MP-202

Vardhaman Mahaveer Open University, Kota

Research Methodology

Course Development Committee

Chairman
Prof. (Dr.) Naresh Dadhich
Vice-Chancellor
Vardhaman Mahaveer Open University, Kota

Convener and Members

Subject Convener and Coordinator
Prof. P.K. Sharma
Department of Management,
Vardhaman Mahaveer Open University, Kota

Members:

1. **Prof.P.N. Mishra**
   Institute of Management Studies,
   Devi Ahilya University, Indore
2. **Prof. Kalpana Mathur**
   Deptt. of Management Studies
   J.N.V. University, Jodhpur
3. **Prof. Harsh Dwivedi**
   R.A. Podar Institute of Management Studies,
   University of Rajasthan, Jaipur
4. **Prof. Naval Kishore**
   School of Management Studies,
   IGNOU, New Delhi
5. **Prof. Geetika Kapoor**
   R.A. Podar Institute of Management Studies,
   University of Rajasthan, Jaipur
6. **Prof. R.K. Jain**
   Professor (Retd.)
   JLN Institute of Business Management
   Vikram University, Ujjain
7. **Prof. P.K. Jain**
   Faculty of Management Studies,
   Mohanlal Sukhadia University, Udaipur
8. **Prof. Manoj Kumar**
   Deptt. of Management Studies
   MDS University, Ajmer
9. **Dr. R.K. Jain**
   Deptt. of Management
   Vardhaman Mahaveer Open University, Kota

Editing and Course Writing

Editor:
Prof. Atul Dhingra
Deptt. of Management Studies,
Haryana Agriculture University, Hisar

Writers:
Prof. B.P. Rao (Unit 1, 2)
Former Pro-Vice Chancellor
Andhra University, Vishakapatnam
Prof. Atul Dhingra (Unit 3, 4, 14)
Deptt. of Management Studies
Haryana Agriculture University, Hisar
Ms. Upasana Tyagi (Unit 5)
Modi Institute of Management & Technology, Kota
Ms. Anamika Sharma (Unit 6)
Modi Institute of Management & Technology, Kota
Dr. Charu Shri (Unit 7, 8, 12)
Institute of Technology & Management,
Navi Mumbai
Prof. Shipra Arora (Unit 9, 10, 11)
KIIT, Gurgogn
Dr. Surendra Bhardwaj (Unit 13)
Govt. Gollege, Kota
Dr. Vishnu Norayan Mishra (Unit 15)
Vaishnaw Institute of Management, Indore

Academic and Administrative Management

Prof. (Dr.) Naresh Dadhich
Vice-Chancellor
Vardhaman Mahaveer Open University, Kota

Prof. B.K. Sharma
Director (Academic)
Vardhaman Mahaveer Open University, Kota

Mr. Yogendra Goyal
Incharge
Material Production and Distribution Department

Course Material Production

Mr. Yogendra Goyal
Assistant Production Officer
Vardhaman Mahaveer Open University, Kota

MP-202

Vardhaman Mahaveer Open University, Kota

CONTENTS

Research Methodology

Unit No.	Name of Unit	Page No.
Block -I:	**Business Research**	
1	Research Methodology	1-12
2	Research Design	13-28
3	Sampling	29-38
4	Scaling Techniques	39-49
5	Data Collection	50-69
Block -II:	**Data Presentation**	
6	Processing and Presentation of Data	70-90
7	Measure of Central Tendency	91-104
8	Measures of Variation and Skewness	105-120
Block III:	**Hypothesis Testing**	
9	Tests of Hypothesis I	121-139
10	Tests of Hypothesis II	140-156
11	Chi-Square Test	157-178
12	Qualitative Research	179-200
Block IV:	**Statistical Analysis**	
13	Correlation and Simple Regression	201-215
14	Computer Applcations in Research Methodology	216-224
Block IV:	**Presentation of Research Findings**	
15	Report Writing	225-235

MP-202

Vardhaman Mahaveer Open University, Kota

Research Methodology

Unit - 1 : Research Methodology

Unit Structure:

1.0 Objectives
1.1 Introduction
1.2 Definition and Objectives
1.3 Characteristics of Research
1.4 Types of Research
1.5 Research Methodology
1.6 Research Process
1.7 Qualities of a Good Researcher
1.8 Summary
1.9 Key Words
1.10 Self Assessment Test
1.11 References

1.0 Objectives

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

- What is Research Methodology?
- What are the objectives of Research Methodology?
- What are the types of Research Methodology?
- What are the qualities of a good research?
- What is the need for Research Methodology?

1.1 Introduction

Research is a scientific process to invent and discover. Invention is the creation of something new which is totally unknown and non-existing hitherto, on the other hand discovery is related to finding of something which already exists but is unknown and hidden. Ever since human beings were conscious about their civilisation, they started inventing and discovering. With the advancement of civilisation they started inventing and discovering to control and conquer the environment in which they were living.

The problems faced by man are numerous. Some problems are created by the nature and others are created by himself in the process of civilisation. Therefore, he started searching and researching the ways and means to tackle and solve the problems or issues in a scientific manner. Therefore, the systematic, scientific and non-scientific research processes have evolved over the year.

Research is often, therefore, described as an active, diligent, meticulous and systematic process of enquiry aimed at discovering, interpreting and revising facts. This intellectual investigation produces a greater understanding of events, behaviours and theories, and makes practical applications through systems, procedures and laws.

Research is undertaken within most professions and fields of studies. It is more than just a study or set of steps followed to achieve a goal. Research is a way of thinking, critical examination of various aspects in the environment. The exploration of hidden facts, unearthing new ones, discovering new methods and remedies is achieved through research.

1.2 Definition and Objectives

Research has been defined in a number of different ways; a broad definition of research is given by Martin Shuttleworth - "In the broadest sense of the word, the definition of research includes any gathering of data, information and facts for the advancement of knowledge."

The Merriam-Webster Online Dictionary defines research in more detail as "a studious inquiry or examination; especially, investigation or experimentation aimed at the discovery and interpretation of facts, revision of accepted theories or laws in the light of new facts, or practical application of such new or revised theories or laws". The Advanced Learner's Dictionary of Current English lays down the meaning of research as "a careful investigation of inquiry specially through search for new facts in any branch of knowledge."

According to Clifford Woody, "Research comprises defining and redefining problems, formulating hypothesis or suggested solutions; collecting, organizing and evaluating data, making deductions and reaching conclusions; and at last carefully testing the conclusions to determine whether they fit the formulating hypothesis". Research is a careful and exhaustive investigation of a phenomenon with an objective of advancing knowledge. Theodorson and Theodorson define it as "…a systematic and objective attempt to study a problem for the purpose of deriving general principles".

With the above definitions it can be said that research is a process which begins with a question or a problem and tries to find answers to such questions through the application of systematic and scientific methods. Thus, research is the systematic approach towards a purposeful investigation. This process includes formulating a hypothesis, collection of data on relevant variables, classifying and tabulating the data, analyzing and interpreting the results and reaching conclusions.

For any research the main aim and purpose is to discover new solutions to problems through the proper and systematic application of various procedures. The research objectives can be grouped into the following:

1. **To gain knowledge**: Research helps one to gain insight and better understanding of a phenomena or a problem etc. This helps him to take proper decisions or other measures for improving the environment. The knowledge thus gained facilitates new dimensions of the existing and emerging problems and provide ability to discern the phenomena.

2. **To generate new concepts and theories:** One of the objectives of research is to try and find new concepts and theories that explain the existing ones or even add to the existing knowledge. In this process, the human endeavour is able to establish new insights that lead to new theories. This can be viewed as Exploratory or Formulative Research.

3. **To test a hypothesis:** When research is conducted with certain premises or suppositions, attempts are made to establish relationship between variables involved to prove the hypothesis. These research studies are known as hypothesis-testing research studies.

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, re-jected 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.
- **Free from Personal Biases:** Bias is a particular tendency or inclination that prevents unprejudiced consideration of a question. A good research should be free from the researcher's personal biases and must be based on objectivity and not subjectivity.
- **Systematic:** A research study must have well planned steps which are interrelated to each other. Each step leads on to the next step which ensures a proper study in a comprehensive manner.
- **Reproducible:** A researcher should be able to get approximately the same results by using an identical methodology by conducting investigation on a population having characteristics identical to the one in the earlier study.
- **Valid and Verifiable Evidence**: A research's observations should be verifiable by others. It can be tested and falsified by others.
- **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 Types of Research

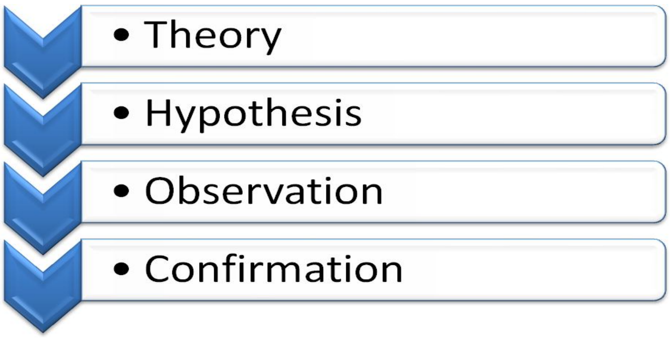
Based upon the reasoning and the arguments that assert the conclusions, Research is classified in two approaches.

- **Deductive**:

Deductive approach begins with a general idea (such as theory, laws, principles) and based on them a specific hypothesis is formed. Deductive reasoning works from the more general to the more specific problem. It is also called a "top-down" approach. In this one starts with a theory about a topic of his interest and then narrows that down and makes it more specific. From this a hypotheses can be tested. It is further narrowed down when observations that address the hypotheses is collected. This ultimately enables one to test the hypotheses with specific data confirmation (or not) of the original theory.

The figure below shows the process of a Deduction.

Deductive Process

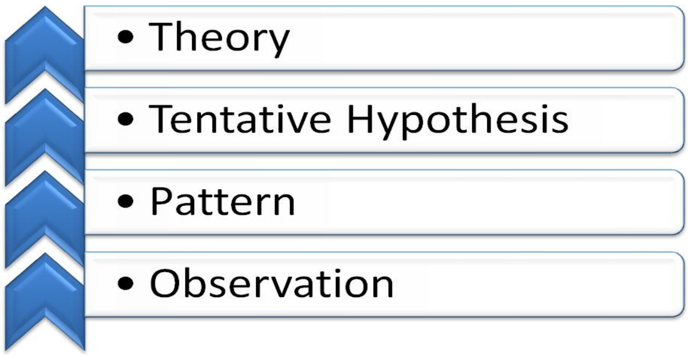


As mentioned it starts with a theory which is more general in nature, and subsequently being narrowed down to smaller testable hypotheses.

• **Inductive**:

Inductive approach begins with specific issues. It starts with observations of individual cases. Based on the accumulation of such observation, one would build a general idea on that observation.

Inductive Process



Inductive reasoning works the other way round as that of deductive approach. It moves from a specific observation to a broader generalization and theory. It is also called as "bottom up" approach. In Inductive reasoning, one begins with a specific observation and measures, then starts to detect the patterns and regularities from which some tentative hypotheses are formulated that can be explored. Finally some general conclusions or theories are developed.

Research further can also be classified into two major categories: (i) **Basic research** and (ii) A**pplied research**.

Basic Research is also known as fundamental research, theoretical research or pure research. Fundamental research is mainly concerned with generalisations and the formulation of a theory. Pauline V. Young quotes "Gathering knowledge for knowledge's sake is termed 'pure' or 'basic' research". It aims at expanding the frontiers of knowledge and does not directly involve pragmatic problems. The essence of basic research is that it addresses itself to more fundamental questions and not to the problems with immediate commercial potential.

Applied Research is known as decisional research. It proceeds with a certain problem and it specifies alternate solutions and the possible outcomes of each alternative. Unlike basic research, it is prompted by commercial considerations. Applied research can be divided into two categories: (i) problem-solving research and (ii) problem- oriented research. Problem-solving research, as the name implies, is concerned with a particular issue or a problem, and is usually proprietary in character, i.e., such a research is undertaken within a firm or by an outside consultant on its behalf. Problem-oriented research, on the other, is concerned with a class of issues or problems in which several firms may be interested. Research of this type is usually concerned with conceptual aspects but is oriented towards applied problems.

Research can also be classified into following categories.

a) **Ad-hoc Research:** Ad hoc research is always conducted for a specific purpose. Generally applied to single surveys which are designed as a 'one-off' and relate to a specific need. For example, 'As and when' the manager faces a problem, he has to find the solution to overcome the problem. So, he immediately appoints a research cell or a committee to study the problem and offer a specific solution. Once the committee finds a solution it loses its existence. This means, once the purpose of the research is over automatically the research team ceases its existence.

b) **Continuous Research:** A survey conducted both regularly or frequently among parallel samples within the same population; or a survey in which the interviews are spaced over a long period of time can be termed as continuous research. In some organizations, particularly, large corporations there will be a separate research department which carries out research activities on behalf of the organization continuously. Many organizations may not wait for the problems to surface, so by the time the problem emerges, the research organization is able to anticipate the problems and offer solutions. In the globalised environment, continuous search is an essential prerequisite for the success of the organization. Continuous studies are made to decide about the market place and product performance.

c) **Exploratory Research:** Exploratory research is often conducted because a problem has not been clearly defined as yet, or its real scope is as yet unclear. It allows the researcher to familiarize with the problem or concept to be studied and perhaps generate hypothesis to be tested. Exploratory research is used when the topic or issue is new and when data is difficult to collect. One of the main characteristics is that it is flexible and can address research questions of all types (such as What, Why, How). It is often used to generate formal hypotheses.

   It is the initial research before more conclusive research is undertaken. Exploratory research helps determine the best research design, data collection method and selection of subjects, and sometimes it even concludes that the problem does not exist. Another common reason for conducting exploratory research is to test concepts before they are put in the marketplace, always a very costly endeavour. Exploratory research can be quite informal, relying on secondary research such as reviewing available literature and/or data, or qualitative approaches such as informal discussions with consumers, employees, management or competitors, and more formal approaches through in-depth interviews, focus groups, projective methods, case studies or pilot studies. The results of exploratory research are not usually useful for decision-making by themselves, but they can provide significant insight into a given situation.

d) **Conclusive Research**: Conclusive research is meant to provide information that is useful in reaching conclusions or decision-making. It tends to be quantitative in nature i.e., in the form of numbers that can be quantified and summarized. The purpose of conclusive research is to provide a reliable or representative picture of the population through the use of a valid research instrument. In the case of formal research, it will also test hypothesis. Both primary and secondary data are used to draw conclusions.

e) **Observation Research:** Observation is a primary method of collecting data by human, mechanical, electrical or electronic means. The researcher may or may not have direct contact or communication with the people whose behaviour is being recorded. Observation techniques can be part of qualitative research as well as quantitative research techniques. There are six different ways of classifying observation methods:

- Participant and Non-Participant Observation
- Obtrusive and Unobtrusive (or Physical Trace) Observation
- Observation in Natural or Contrived Settings
- Disguised and Non-Disguised Observation
- Structured and Unstructured Observation
- Direct and Indirect Observation.

1.5 Research Methodology

Research methods may be understood as all those methods/techniques that are used for conducting research. So, research methods or techniques, thus, refer to the methods the researchers use in performing research operations. In other words, all those methods which are used by the researcher during the course of studying his research problem are termed as research methods. Research methods refer to the behaviour and instruments used in selecting and constructing research technique. They are conventionally divided into quantitative, qualitative and participatory methods each with differing underlying approaches, tools and techniques.

Methodology refers to the rationale and philosophy that underlie a particular study of the available methods. So, it is a way to systematically solve the research problem. This is why scholarly literature often includes a section on the methodology of the researches. Methodology is defined as: The analysis of the principles of methods, rules, and postulates employed. It is the systematic study of methods that can be or have been applied within a discipline or a particular procedure or set of procedures.

Methodology includes: A collection of theories, concepts or ideas, Comparative study of different approaches and Critique of the individual methods.

Methodology is understood as a science of studying how research is done scientifically. Here the researchers study various steps that are generally adopted by a researcher in studying his research problem along with the logic behind them. It is necessary for the researcher to know not only the research methods/techniques but also the methodology. The scope of research methodology is wider than that of research methods. It can be said that research methodology has many dimensions and research methods are part of the research methodology.

1.6 Research Process

Research process is the main content of any research. It defines the research and describes the skills required to identify the problem, the decision alternatives, and the research needs which are critical components of a research activities. It involves a dedicated system of scientific methodology that can be used by researchers to arrive at the right conclusions.

Research process involves a number of inter-related activities which overlap and do not rigidly follow a particular sequence. A researcher is often required to think a few steps ahead, because various steps in research process are inter-woven into each other and each step will have some influence over the other steps. In research, even though our focus is on one particular step, other inter-related steps of operations are also being looked into simultaneously. As we complete one activity or operation, our focus naturally shifts from it to the subsequent one, i.e. the focus is not concentrated exclusively on one single activity or operation at any particular point of time. The research process provides systematic, planned approach to the research project and ensures that all aspects of the research project are consistent with each other.

Earl Babble proposed the following six elements of a research process:

- **Objective or Problem Statement ,** i.e., stating what is to be studied, its worth and practical significance, and its contribution to the construction of social theories.
- **Literature Review,** i.e., what others have said about this topic, what theories have been addressed to it and what are the flaws in the existing research that can be remedied.
- **Subjects for Study,** i.e., from whom is the data to be collected, how to reach persons who are available for Study, whether selecting sample will be appropriate, and if yes, how to select this sample and how to insure that research that is being conducted will not harm the respondents.
- **Measurement,** i.e., determining key variables for the study, how will these variables be defined and measured, how will these definitions and measurements differ from previous researches on the topic.
- **Data Collection Methods,** i.e., determining methods to be used for collecting data-survey or experiment, etc., statistics to be used or not.
- **Analysis,** i.e., spell out the logic of analysis whether variations in some quality are to be accounted or not, and the possible explanatory variables to be analyzed.

Horton and Hunt have elaborately pointed out eight steps in a research:

1. Define the problem, which is worth studying through the methods of science. .
2. Review literature, so that errors of other research scholars may not be repeated.
3. Formulate the hypotheses, i.e., propositions which can be tested.
4. Plan the research design, i.e., outlining the process as to how, what and where the data is to be collected, processed and analyzed.
5. Collect the data, i.e., actual collection of facts and information in accordance with the research design. Sometimes it may become necessary to change the design to meet someu n f o r e s e e n difficulty.
6. Analyze the data, i.e., classify, tabulate and compare the data, making whatever tests are necessary to get the results.
7. Draw conclusions, i.e., whether the original hypothesis is found true or false and is confirmed or rejected, or are the results inconclusive? What has the research added to our knowledge? What implication has it for sociological theory? What new questions have been posed for further research?
8. Replicate the study. Though the above-mentioned seven steps complete a single research study but research findings are confirmed by replication. Only after several researches can the research conclusions be accepted as generally true.

From the above research process can be summed up as:

1. Need for research
2. Define the research objective
3. Identify data needs

4. Identify data sources
5. Choose an appropriate research design and data collection method
6. Design the research instrument or form
7. Identify the sample
8. Collect data, including any relevant secondary data
9. Analyze and interpret the data
10. Present the research findings to decision makers

Need for the Research:

The probability of success of a research is greatly enhanced when the beginning is correctly defined as a precise statement of goals and justification. The research process begins with the recognition of a research problem. Often, considerable analysis of historical data or secondary information basically provides the background of the problem. While deciding about the need for the research one has to consider the use of the research findings, availability of resources, and expected benefits against the cost involved in conducting the research. The formulation of a general topic into a specific research problem, thus, constitutes the first step in a scientific enquiry. Essentially two steps are involved in formulating the research problem, viz., understanding the problem thoroughly, and rephrasing the same into meaningful terms from an analytical point of view.

Albert Einstein went so far as to say that "The formulation of a problem is often more essential than its solution". In an academic institution the researcher can seek the help from a guide who is usually an experienced man and has several research problems in mind. Often, the guide puts forth the problem in general terms and it is up to the researcher to narrow it down and phrase the problem in operational terms. This stage will facilitate the researchers to come to the conclusion not only to recognize the problem situation but also need to proceed further to conduct the research. In abstract, this background leads to correctly define the problem.

Research Objectives:

While deciding to conduct the research one has to have clarity about the objective of the research. The clarity of objectives facilitates the researcher to move in right direction. In other words, accurate definition of purpose of research requires (i) Identifying a number of specific issues to be covered and (ii) Deciding which of those issues is worth to pursue further.

The culmination of the problem definition process is a statement of the research objectives. These objectives are stated in terms of the precise information necessary to address the research problem. Well formulated objectives serve as a road map in pursuing the research project. Research objectives must be as specific and unambiguous as possible. After an introduction which describes the broader context within which the research should be situated, it is important to state the objectives or purpose pursued by the research itself. Often, this is a fairly broad or general statement as well. The purpose of the research is given as: "The following study is an attempt to provide a more meaningful and defensible measure of public opinion". This is a fairly vague statement and should have been followed up with a much more precise research question (in question format) or problem statement (a re-wording of the research question into a statement format). At this point it is important to inform the reader about the breadth or scope of the study, limitations that may have to be imposed on the study due to cost or time constraints or accessibility to respondents.

All research is based on a set of assumptions or factors that are presumed to be true and valid. For instance, it is generally assumed that respondents will reply honestly and accurately as far as they are able to do so. By stating these assumptions up front, the researcher reduces potential criticism of the research, but without them, the research itself would not be possible.

In formal research, the researcher will provide an educated guess regarding the outcome of the study, called **hypothesis**. That means researcher states a research objective in the form of a hypothesis. A hypothesis is a conjectural statement about a relationship between two or more variables that can be tested with empirical data; it is considered to be plausible, given the available information. It does not matter whether researcher correctly predict the outcome of the research or not, since rejecting a hypothesis does not mean that the research itself is poor, but rather that the research has results that are different from what the related literature led to believe should have been expected.

Objectives can be listed under two headings: a) main objectives (aims) and b) sub-objectives. The main objective is an overall statement of the thrust of the study. It is also a statement of the main associations and relationships that one seeks to discover or establish. The sub-objectives are the specific aspects of the topic that one wants to investigate within the main framework of his study.

Identify Data Needs:

After deciding about the purpose of the research, next step is to decide about the nature of data that has to be collected. Generally objectives of research indicate what type of data is required. Researcher should have clarity about the kinds of data that is needed. If there is a mismatch between the purpose of research and data needs, the very purpose of research will be a futile exercise.

Identification of Data Sources:

Once the data needs are identified it becomes important to identify sources of data. Sometimes secondary data may be sufficient to accomplish the research. In some other situations, primary data may be required. However, in most of the cases there is a need to collect both secondary as well as primary data. It is to be noted that, the purpose of the research and urgency of the outcome decides which source of data is needed.

Research Design:

The next step is one of the most important steps in research process, research design. The research design is the plan to be followed to answer the research objectives. The research design is a plan for addressing the research objectives or hypotheses. In essence, the researcher will have to state the conceptual structure within which research would be conducted and also develop a structure or framework to answer a specific research problem. The researcher has to decide which type of research design will be suitable to conduct research.

The research design selection largely depends upon the type of research, purpose of research which has a great influence on the whole process. There is a no single best research design. Instead, different designs offer an array of choices, each with certain advantages and disadvantages. Ultimately, trade-offs are typically involved. A common trade off is between research costs and the quality of the decision-making information needed. The design of the research project is a creative plan to obtain the necessary information in the best way possible. The correct design will save time and money and will result in valid and reliable information.

Select a Basic Method for Research: There is a need to decide what type of, probability or non-probability, method to select for research. Again this depends upon the type of research which will influence.

There are three basic research methods:

a. **Survey Research:** In survey research investigator interacts with the respondents to gather data required for the research. This is often descriptive in nature but can be causal.

b. **Observation Research:** In this the research investigator observes the actions of the respondent without any direct interaction. Observation research is typically descriptive.

c. **Experiment Research:** In this research, investigator changes one or more variables and observes the effect of these changes on another variable. Experiment research is almost always causal.

Selecting the Sampling Procedure: A sample is a subset from a larger population. Although the basic nature of the sample is specified in the research design, selecting the sampling procedure is a separate step in the research process. Several questions must be answered before a sampling procedure is selected. First, the population or universe of interest must be defined. This is the group from which the sample will be drawn. It should include all the elements. Once the population has been defined, the next question arises as to whether to use a probability sample or a non probability sample.

Collection of Data: There are several ways of collecting the appropriate data which differ considerably in context of money costs, time and other resources at the disposal of the researcher. There are various ways in which a researcher can collect data such as, surveys, observations, postal interviews, telephone interviews etc. The researcher should select one of these methods of collecting the data taking into consideration the nature of investigation, objective and scope of the inquiry, financial resources, available time and the desired degree of accuracy.

Analyzing the Data: After the data is collected the next step in the research process is data analysis. The purpose of the analysis is to interpret and draw conclusions from the collected data. The researcher may use a variety of techniques, beginning with simple frequency analysis to various complex multivariate techniques.

Writing and Presenting the Report: After data is analysed; the researcher must prepare a report with his findings, and the conclusions drawn from such and at the end his recommendations. The researcher usually will be required to present both written and oral reports on a project. The nature of the audience must be kept in mind when these reports are being prepared and presented.

1.8 Qualities of a Good Researcher

To undertake a research the researcher must possess some basic qualities such as:

1. **Required Subject Knowledge:** One must have the required knowledge in the subject he is doing the research. A clear understanding of the topic and the process to undertake comes from this knowledge.

2. **Ability and Capacity:** Research involves an in-depth study on a topic and for the research to be successful the researcher should have the ability to undertake such study. He should have capacity to get in-depth information from his respondents.

3. **Unbiased:** Any study undertaken with a bias leads to failure in research. Will be yhe researcher should be capable of collecting information without any bias or prejudices. He should be prepared to analyse the data collected impartially and without twisting that in his favour, no matter whether that is suiting to his viewpoint or not. He should not have preconceived notions.

4. **Persistence:** One has to understand that research involves lot of factors which have a bearing upon the study. These factors may not be available very easily. He can encounter lot of rejection while collecting the data. But one has to be persistent in his efforts to produce a good and successful research.

5. **Accurate and Precise:** Research involves data. One has to be careful in collection of data. There is so much data available, one must be very careful in collection of data that is relevant to his study. He should have desire for accuracy, keen observations and precision of statement. He should not be content with approximate or accepting what prima facie he can see.

6. **Alert:** He should always have an alert mind and capable enough to be receptive to hints while collecting data. It is just possible that a seemingly innocent hint may lead to identifying far-reaching and significant research problems. In the words of Cohen, "It is a mark of scientific genius to be sensitive to difficulties where less gifted people pass by untroubled by doubt."

7. **Time and Cost Management:** Research study is often time bound. Sometimes if the study takes too long then the purpose of such study loses its importance and becomes useless. A researcher should have proper sense of devoting his time in a balanced way. He must realise that time at his disposal is very limited. He cannot spend much on his research either by way of money or time.

8. **He Should be Capable of Realising the Restraints:** as well as favourable circumstances in which he is placed while on work and act according to the situation.

9. **Communication Skills:** A researcher needs to interact with many people during his course of study. It is from these interactions that he gains various inputs for his study. He should be good in conversation and during the course of his research he should speak the language of the respondents from whom he is going to get information.

1.8 Summary

Research is a careful and ex-haustive investigation of a phenomenon with an objective of advancing knowledge. It is an art of scientific investigation. The purpose of research is to discover answers to questions through the application of scientific procedures. Research inculcates scientific and inductive thinking and it promotes the development of logical habits of thinking and organisation. The increasingly complex nature of business and government has focused attention on the use of research in solving operational problems. Research, as an aid to economic policy, has gained added importance, both for government and business. Research is basically classified into major categories: Basic research and Applied research. Research process is step by step process of conducting research.

1.9 Keywords

- **Basic Research** is sometimes called fundamental research, theoretical research or pure research. Fundamental research is mainly concerned with generalisations and the formulation of a theory.

- **Applied Research**: It is the application of basic principles and other existing knowledge to the solution of operational problems

- **Exploratory Research:** Exploratory research provides insights into and comprehension of an issue or situation.

- **Conclusive Research**: Conclusive research is meant to provide information that is useful in reaching conclusions or decision-making.
- **Descriptive Research:** In this the issues are described rather than offering specific solution to the problem. This is more like exploration than cause for the problem.
- **Research Methods** may be understood as all those methods/techniques that are used for conducting research.
- **Research Methodology** refers to the rationale and philosophical assumptions that underlie a particular study of the available methods.
- **Research Process** is a step-by-step process of developing and conducting a research paper.
- **Hypothesis** is a conjectural statement about a relationship between two or more variables that can be tested with empirical data; it is considered to be plausible, given the available information.
- **Research Design** is the plan to be followed to answer the research objectives.

1.10 Self Assessment Test

1. What do you understand by Research Methodology? Why is it needed? Explain.
2. Discuss the objectives and characteristics of a research study.
3. Throw light on various types of research.
4. Provide a detailed account of research process.
5. Discuss various qualities of a good research.

1.11 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

Unit - 2 : Research Design

Unit Structure:

2.0 Objectives
2.1 Introduction
2.2 Benefits of Research Design
2.3 Types of Research Design
2.4 Exploratory Research
2.5 Conclusive Research
2.6 Relationships among Exploratory, Descriptive, & Causal Research
2.7 Pilot Study
2.8 Case Study
2.9 Synopsis
2.10 Summary
2.11 Key Words
2.12 Self Assessment Test
2.13 References

2.0 Objectives

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

- What are the different types of research design?
- What is exploratory research? What methods are adopted during exploratory research?
- What is descriptive research and what methods are adopted for descriptive research?
- Relationship among Exploratory, Descriptive & Causal Research.

2.1 Introduction

Research design is simply a plan for a study. It is a plan which lays out the way the entire research would be done and carried out, specifying the methods and procedures for collecting and analyzing the needed information. So it can be called as a blue print of the study. A major issue in research is the preparation of the research design. Decisions regarding what, where, when, how much, by what means, concerning an enquiry or a research study constitute a research design.

Research design, according to Caldwell & Herbst, can be explained as the strategy and the plan by which the purpose of the study is to be planned and realised. Let us take an example; when a company feels that it requires an additional capital for expansion of its plant it first needs to look at the various sources of capital and also the future expected returns from the project. There is no point in approaching a financial institution or issuing shares or even debentures before ascertaining the facts. Various factors need to be considered as to the estimated cash inflows from the project, the cost of capital, the mix of the capital, repaying ability etc. Until this is done one cannot plan a proper source of capital. Similarly, research needs a design or a structure before data collection or analysis can commence.

The function of a research design is to ensure that the evidence obtained enables one to answer the initial question as unambiguously as possible. This has been aptly put by Claire Selltiz in Research Methods in Social Sciences, wherein he asserts at 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.

So, a research design would mainly contain, a Sample Design explaining the way sample size is arrived at and how it would be selected etc., a Statistical Design containing the statistical tools and techniques to analyse the collected data.

2.2 Benefits of Research Design

Given below are a few benefits of Research Design

1. A researcher realises the objectives of the study with the help of research design which helps to bridge between what has been established and what is to be done.

2. Research design facilitates the researcher to understand and decide what data are needed to achieve the objectives.

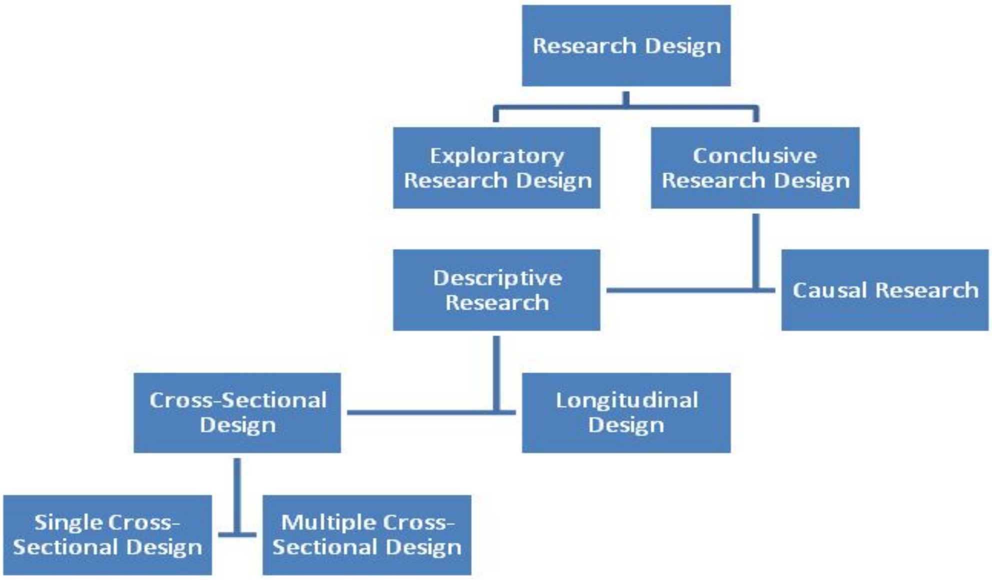
3. A written design helps the researcher exactly in clear terms, what is to be done and why.

4. It helps the researcher to keep the computations and thinking on the path to solutions and recommendations.

5. It guides the researcher how to conduct the research.

2.3 Types of Research Design

There are two types of research designs namely, Exploratory and Conclusive research. Exploratory research is a type of research conducted for a problem that has not been clearly defined. It helps in determining the best research design, data collection method and selection of subjects. It provides the relevant variable that need to be considered. The research methods are flexible, qualitative and unstructured. The researcher in this method does not know "what he will find".



Conclusive research design is more formal and structured compared to exploratory research design. It helps in understanding the relationships between variables. The research helps in decision making.

2.4 Exploratory Research

Exploratory research in general is unstructured, informal research undertaken with an objective to gain background information about the general nature of the research problem. Major emphasis is on converting broad, vague problem statements into small, precise sub-problem statements. It has a great amount of flexibility and does not offer any clear hypothesis about problem. It is generally followed by confirmatory research because it gives a "feel" of various relevant issues. The results of research can give some indication as to the "why", "how" and "when" something occurs, but it cannot explain "how often" or "how many".

It is designed to generate basic knowledge, clarify relevant issues, uncover variables associated with a problem, uncover information needs, and/or define alternatives for addressing research objectives.

In the early stages of research, one usually lacks sufficient understanding of the problem in order to formulate a specific hypothesis. Further, there are often several tentative explanations. Let us take a scenario where the company's top management asks about the dropping sales and product movement, the few feedbacks it received were, "Sales are down because our prices are too high", "Our dealers or sales representatives are not doing a good job", "Our advertisement is weak" and so on. Now, in this scenario, from the statements very little information is available to point out, what is the actual cause of the problem.

The major purpose of exploratory research is to identify the problem more specifically. Therefore, exploratory study is used in the initial stages of research.

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.

The examples given below indicate each of the above type:

Research Purpose	Research Question	Hypothesis
1) What product feature, if stated, will be most effective in the advertisement?	What benefit do people derive from this Ad appeal?	No hypothesis formulation is possible.
2) What new packaging is to be developed by the company (with respect to a soft drink)?	What alternatives exist to provide a container for soft drink?	Paper cup is better than any other forms, such as a bottle.
3) How can our insurance service be improved?	What is the nature of customer dissatisfaction?	Impersonalization is the problem.

In example 1 the research question is posed to determine "What benefit do people seek from the Ad?" Since no previous research is done on consumer benefit for this product, it is not possible to form any hypothesis.

In example 2 some information is currently available about packaging for a soft drink. Here it is possible to formulate a hypothesis which is purely tentative. The hypothesis formulated here may be only one of the several alternatives available.

In example 3 the root cause of customer dissatisfaction is known, i.e. lack of personalised service. In this case, it is possible to verify whether this is a cause or not.

2.4.1 Exploratory Research Methods

The quickest and the cheapest way to formulate a hypothesis in exploratory research are by using any of the four methods:

a) Literature Search

This refers to "referring to a literature to develop a new hypothesis". The literature referred can be trade journals, professional journals, research finding publications, statistical publications etc. Example: Suppose a problem is "Why are sales down?" This can quickly be analysed with the help of published data which should indicate whether the problem is an "industry problem" or a "firm problem".

The following questions help in reviewing the literature:

- What does previous research reveal about the problem?
- What is the theoretical framework for the investigation?
- Are there complementary or competing theoretical frameworks?
- What are the hypotheses and research questions that have emerged from the literature review?

b) Experience Survey

In experience surveys, it is desirable to talk to persons who are well informed in the area being investigated. It includes interviewing persons who are having experienced in the area of study to their memories and experiences. Here, no questionnaire is required. The approach adopted in an experience survey should be highly unstructured so that the respondent can give divergent views. Since the idea of using experience survey is to undertake problem formulation, and not conclusion, probability sample need not be used. Those who cannot speak freely should be excluded from the sample.

Example:

1. A group of housewives may be approached for their choice for a "ready to cook product".
2. A publisher might want to find out the reason for poor circulation of newspaper introduced recently. He might meet (i) Newspaper sellers (ii) Public reading room (iii) General public (iv) Business community, etc.

Given below are some of investigative questions:

1. What is being done?
2. What has been tried in the past without success? With success?
3. How have things changed?
4. What are the change-producing elements of the situation?
5. Who is involved in decisions and what role does each person play?
6. What are problem areas and barriers under study?
7. Whom can we count on to assist and/or participate in the research?
8. What are the priority areas?

c) Focus Group

Another widely used technique in exploratory research is the focus group. In this panels of people are invited to a discussion about a specified topic. The discussion is coordinated by a moderator. He guides and facilitates the group to exchange ideas, feelings, and experiences on the topic. The group usually consists of 8-12 persons. While selecting the group, one must make sure that they have a common background with relevance to the research.

Normally there are 4-5 groups, though there may be even 6-8 groups. The guiding criterion is to see whether the latter groups are generating additional ideas or repeating the same with respect to the subject under study. When this shows a diminishing return from the group, the discussions stopped. A typical focus group session lasts between one and half hour to two hours. The moderator under the focus group has a key role. His job is to guide the group to proceed in the right direction.

A moderator/facilitator should be possessing following characteristics:

- **Listening:** He must have a good listening ability. The moderator must not miss the participant's comment, due to lack of attention.
- **Permissive:** The moderator must be permissive, yet alert to the signs that the group is disintegrating.
- **Memory:** He must have a good memory. The moderator must be able to remember the comments of the participants. Example: A discussion centered around a new advertisement by a telecom company. The participant may make a statement early and make another statement later, which is opposite to what was said earlier. Example: The participant may say that s (he) never subscribed to the views expressed in the advertisement by the competitor, but subsequently may say that the "current advertisement of competitor is excellent".
- **Encouragement:** The moderator must encourage unresponsive members to participate.
- **Learning:** He should be a quick learner.
- **Sensitivity:** The moderator must be sensitive enough to guide the group discussion.
- **Intelligence:** He must be a person whose intelligence is above the average.
- **Kind/firm:** He must combine detachment with empathy.

d) Case Studies

Case refers to the unit of analysis or topic chosen for study (i.e., the individual, organization, or program). Case studies are often used to provide context to other data (such as outcome data), offering a more complete picture of what happened in the programme and why. They excel in making understand a complex issue or even add strength to what is already known through previous research.

Case studies emphasize detailed contextual analysis of a limited number of events or conditions and their relationship. Robert K. Yin defines the case study research method a "an empirical inquiry that investigates a contemporary phenomenon within its real-life context; when multiple sources of evidence are used." Case studies are well suited to carry out exploratory research. However, the results of investigation of case histories are always considered suggestive, rather then conclusive. Case studies are often used to provide context to other data (such as outcome data), offering a more complete picture of what happened in the program and why.

2.5 Conclusive Research

Conclusive research design helps to test specific hypotheses and examine the relation between variables. The primary purpose of conclusive research, also known as confirmatory research, is to help decision makers in determining, evaluating and choosing the best course of action in a situation. It aids the decision makers in selecting a specific course of action and verifying the insights. It is useful especially when the decision maker already in his mind has one or more alternatives and is specifically looking for information pertinent to evaluating them. The major emphasis of conclusive research is o evaluate and establish reliability of the relevant data. It seeks data that can be expressed in numbers, therefore, it can be analysed using a quantitative process. Since data requirements are clearly specified in a situation calling for conclusive research, and since such research is intended as an aid in the final stages of the decision-making process, conclusive research is typically more formal and rigorous than exploratory research. It has clearly defined objectives with specific courses of action taken to solve the problem.

Conclusive research is of two types:

2.5.1 Descriptive Research

The name itself reveals that, it is essentially a research to describe something. Descriptive research studies are those studies which are concerned with describing the characteristics of a particular individual, or of a group. Example: It can describe the characteristics of a group such as-customers, organisations, markets etc. Descriptive research provides "association between two variables" like income and place of shopping, age and preferences. What descriptive research cannot indicate is that it cannot establish a cause and effect relationship between the characteristics of interest. This is the distinct disadvantage of descriptive research.

Descriptive study requires a clear specification of "Who, what, when, where, why and how" of the research. Example: Consider a situation of convenience stores (food world) planning to open a new outlet. The company wants to determine, "How people come to patronize a new outlet?" 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

Hypothesis study at the descriptive research stage (to demonstrate the characteristics of the group)

Management problem	Research problem	Hypothesis
How should a new product be distributed?	Where do customers buy a similar product right now?	Upper class buyers use 'Shopper's Stop' and middle class buyers buy from local departmental stores
What will be the target segment?	What kinds of people buy our product now?	Senior citizens buy our products. Young and married buy our competitors products.

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.

The data collected is in the form of direct and indirect questions and will depend upon: nature of research question, frame of study, available fund and kinds of respondents.

There are two types of descriptive research:

(a) Longitudinal Design

These are the designs in which an event or occurrence is measured again and again over a period of time. This is also known as 'Time Series Study. Longitudinal designs involve panels. Panel once constituted will have certain elements. The panel or sample remains constant throughout the period. There may be some dropouts and additions. The sample members in the panel are being measured repeatedly. The periodicity of the study may be monthly or quarterly etc. Changes that take place during those intervals are registered. This type of design is very useful in monitoring studies, where its performance feedback may signal problems.

Example: Assume that a market research is conducted on 'ready to eat food' at two different points of time 'T_1' and 'T_2' with a gap of 4 months. Each of the above two times, a sample of 2000 household is chosen and interviewed. The brands used most in the household are recorded as follows.

Brands	At T_1	At T_2
Brand X	500(25%)	600(30%)
Brand Y	700(35%)	650(32.5%)
Brand Z	400(20%)	300(15%)
Brand M	200(10%)	250(12.5%)
All others	200(10%)	250(12.5%)
	200	100%

As can be seen between period T_1 and $T_{2,}$ Brand 'X' and Brand 'M' have shown an improvement in market share, while Brand 'Y' and Brand 'Z' have decreased, where as all other categories remains the same. This shows that Brand 'X' and 'M' has gained market share at the cost of Y and Z.

In Longitudinal research there are two types of panels:

- **True Panel:** This involves repeated measurement of the same variables, i.e. each member of the panel is examined at a different time, to arrive at a conclusion on the subject.
- **Omnibus Panel:** In omnibus panel too, a sample of elements is selected and maintained, but the information collected from the member varies. For example, at a certain point of time, the attitude of panel members "towards an advertisement" may be measured. At some other point of time the same panel member may be questioned about the "product performance".

Time Series and Trend Designs: In a time series design, data is collected from the sample or population at successive intervals. The trend data relate to matched samples drawn from the same population at successive intervals. It can be of many types. A simple design can be represented as below:

X- 0

Where X indicates the exposure of a group to an experimental treatment and 0 indicates the observation or measurement taken on the subject or group after an experimental treatment. Another method also involves a control group. This can be represented as below:

X - 01 02 03 04 05 06

Where, 0's represent measurement of the control group. This is termed as multiple time-series design.

(b) Cross-sectional Study

Cross-sectional study is one of the most important types of descriptive research. It studies the effect of different levels of treatments on several groups at the same time. It measure units from a sample of the population at only one point in time. They form a class of research methods that involve observation of all of a population, or a representative subset, at a defined time. It can be done in two ways:

• **Field Study:** This includes a depth study. Field study involves an in-depth study of a problem. **Example:** Reaction of Indian men towards branded ready-to-wear suit. Field study is carried out in real world environment settings. Test marketing is an example of field study.

• **Field survey:** Large samples are a feature of the study. The biggest limitations of this survey are cost and time. Also, if the respondent is cautious, then he might answer the questions in a different manner. Finally, field survey requires good knowledge like constructing a questionnaire, sampling techniques used, etc. **Example:** Suppose the management believes that geographical factor is an important attribute in determining the consumption of a product, like sales of woolen wear in a particular location. Suppose that the proposition to be examined is that, the urban population is more likely to use the product than the semi-urban population. This hypothesis can be examined in a cross-sectional study. Measurement can be taken from a representative sample of the population in both geographical locations with respect to the occupation and use of the products. In case of tabulation, researcher can count the number of cases that fall into each of the following classes:

- Urban population which uses the product - Category I
- Semi-urban population which uses the product - Category II
- Urban population which does not use the product - Category III
- Semi-urban population which does not use the product - Category IV.

Here, the hypothesis has to be supported and tested by the sample data i.e., the proportion of urbanites using the product should exceed the semi-urban population using the product.

i. **Single Cross-Sectional Design:** A research design where subjects are assessed at a single time. In this design only one sample of respondents is drawn from the target population, and information is obtained from this sample only once. These designs are also called sample survey research designs. A cross sectional study is fast and can study a large number of respondents at little cost or effort. Also, one doesn't have to worry about the respondents patients dropping out during the course of the study. This study is efficient at identifying association, but may have trouble deciding cause and effect.

ii. **Multiple Cross-Sectional Designs:** A research design in which there are two or more samples of respondents and information from each sample is obtained only once. Often, information from different samples is obtained at different times over long intervals. These types of designs allow comparison at the aggregate level but not at the individual respondent level. As different samples are taken each time a survey is conducted, there is no way to compare the measures on an individual respondent across surveys. One type of multiple cross-sectional design of special interest is cohort analysis.

(c) Cohort Analysis:

A cohort is a group of persons who share a common experience within a defined time period. A cohort is any group of individuals who are linked in some way or who have experienced the same significant life event within a given period. Any study in which there are measures of some characteristic of one or more cohorts at two or more points in time is cohort analysis.

The study design does not require strict random assignment of subjects, which is, in many cases, unethical or improbable. Cohort analysis is an appealing and useful technique because it is highly flexible. It provides insight into the effects of social, cultural, and political change. In addition, it can be used with either original data or secondary data. In many instances, a cohort analysis can be less expensive than experiments or surveys.

2.5.2 Causal Research

Causal research explores the effect of one variable on another. It used to obtain evidence of cause-and-effect (causal) relationships. The research is used to measure what impact a specific change will have on existing norms and allows market researchers to predict hypothetical scenarios upon which a company can base its business plan. For example, if a clothing company currently sells blue denim jeans, causal research can measure the impact of the company changing the product design to the colour white. Following the research, company will be able to decide whether changing the colour of the jeans to white would be profitable. To summarise, causal research is a way of seeing how actions now will affect a business in the future.

Researchers continually make decisions based on assumed causal relationships. These assumptions may not be justifiable, and the validity of the causal relationships should be examined via formal research. Causal research is appropriate for the following purposes: **1.** To understand which variables is the cause (independent variables) and which variables are the effect (dependent variables) of a phenomenon. and **2.** To determine the nature of the relationship between the causal variables and the effect to be predicted.

Like descriptive research, causal research requires a planned and structured design. Although descriptive research can determine the degree of association between variables, it is not appropriate for examining causal relationships. Such an examination requires a causal design, in which the causal, or independent, variables are manipulated in a relatively controlled environment. A relatively controlled environment is one in which the other variables that may affect the dependent variable are controlled or checked as much as possible. The effect of this manipulation on one or more dependent variables is then measured to infer causality. The main method of causal research is experimentation, designed to provide information on potential cause-and-effect relationships.

Causal Design:

A causal design investigates the cause and effect relationship between two or helps in eliminating factors that are not relevant more variables. It is based on reasoning along well tested lines. The three basic principles based on logic for causal research as stated by John Stuart Mill are:

a) Method of Agreement:

"When two or more cases of given phenomenon have one and only condition in common, then that condition may be regarded as the cause (effect) of the phenomenon." This method is called method of agreement.

If we find observation Z in every case where we find condition C, we may conclude that C and Z are causally related where E is the condition related both C and Z. The effect is shown under:

Situation 1: C-Z

Situation 2: E - Z

Therefore, C - Z are causally related.

b) Method of Negative Agreement:

This states that when condition C is found not to associate with observation with observation without Z, we may conclude that there is negative relationship exists between C and Z. This means, that the method of negative agreement is exactly opposite of method of agreement.

c) Method of Concomitant Variation:

It is also known as method of difference. In the following example the first observation shows that C could cause Z and the second shows that other possible factors could not cause Z.

Situation 1:	A ⟶ B ⟶ C		**Produce**	Z
Situation 2:	A ⟶ B		Without C Produce	No-Z
Therefore		C	Produces	Z

2.6 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) When little is known about the problem, it is always desirable to go with exploratory research. Exploratory research is appropriate when the problem needs to be defined more precisely, alternative courses of action identified, research questions or hypotheses development, and key variables isolated and classified as dependent or independent.

b) Exploratory research is the initial step in the overall research design framework. It should, in most instances, be followed by descriptive or causal research. For example, hypotheses developed via exploratory research should be statistically tested using descriptive or causal research.

c) It is not necessary to begin every research design with exploratory research. It depends upon the precision with which the problem has been defined and the researcher's degree of certainty about the approach to the problem. A research design could well begin with descriptive or causal research.

d) Although exploratory research is generally the initial step, it need not be. Exploratory research may follow descriptive or causal research.

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

2.7 Pilot Study

Pilot study is a small scale replica of the main study. Before conducting a study, the researcher may be interested in knowing in advance what will be the possible repercussions of that study in future. Its main aim is to eliminate unnecessary issues that are to be studied and help to incorporate new dimensions which are beyond the imagination of the researcher in the study. If the researcher fails to do this, it may be impossible to rectify the limitations once the main study is over. It is basically a rehearsal to the main study.

A pilot study may play an important role in providing in any of the following services.

- Pilot study findings may provide better knowledge about the problems under study.
- It helps the researcher to eliminate some unnecessary dimensions of the study.
- It helps in identifying the nature of relationship between variables and in formulating hypothesis.

- It helps the researcher to have a better understanding about the population of the study.
- It helps the researcher to have better clarity about the sampling procedures adopted in the study.
- It helps the researcher to assess the suitability of data collection methods adopted in the study.
- It facilitates the researcher the adequacy of the tools adopted in data collection.
- It helps to identify the actual problems in the field work.
- It facilitates the researcher what better methods are to be adopted in improving the authenticity of the study.
- It helps the researcher to restructure the entire process to complete the same in a more systematic way.
- It also helps the researcher to estimate the entire project completion in terms of time, money and other resources.

The nature, scope, size and other issues in conducting pilot study is purely based on the researcher and his experience and also resources availability like money, time and infrastructure. If the main study is conducted in the light of the pilot study, the researcher will have enough confidence and study can be conducted with better clarity. This in turn results in better outcome.

2.8 Case Study

Case study is one of the very popular forms of qualitative data generation method. Basically it is a study of a few units or limited numbers for in-depth analysis. The study may be related to an individual, a family an institution, a cultural group or a community. According to Odum, the case study method is a technique by which individual factor whether it is an institution or just an episode in the life of an individual, or a group is analysed in its relationship to any other in the group. According to Young a case study is a comprehensive study of a social unit is that unit a person, a group, a social institution, a district or a community. So case study is a method that studies the social unit carefully and completely for the purpose of generalization.

Characteristics of Case Study:

1. The researcher will selected one unit only and study the same comprehensively.
2. The scope of the case study is complete coverage of each and every dimension of the selected unit for the study.
3. Since the case study is qualitative in nature, it is the researcher's experience and perception play an important role in the study.
4. The study will respectively into the future, results in insights.
5. The case study methods results generally facilitate the researcher to formulate hypothesis.
6. This type of study is more appropriate to understand the behavioural issues.

Advantages of Case Study:

1. Case study research helps to thorough understanding of the issues that are to be studies.
2. Case study facilitates the researcher to study qualitative issues which is not possible under quantitative research.

3. It helps the researcher to proceed further with the help of the case study method.

4. The case study method may be an alternative method to the organized data collection methods.

5. These methods facilitate the research to study the social unit retrospectively and develop historical issues and analysis.

6. Case study helps to understand the social change with the help of intensive study of limited units.

7. Case study intensifies the researcher interests out of case study observations. Case study approach is essential in the case of therapeutic and administrative purposes.

Limitations:

1. Case study observations are not comparable as these are stray cases.

2. Case study findings cannot be generalized as these studies are not generally conducted under some scientific set of rules.

3. Case study findings may not be very realistic because it is based on certain set of assumptions.

4. Case study approach is used only in limited areas where the quantitative approaches may not really useful.

2.9 Synopsis

Synopsis is an abstract form of research which underlines the research procedure followed and is presented before the guide for evaluating its potentiality. In one sentence it may be described as a condensation of the final report. The structure of synopsis varies and also depends on the guides' choice. However, for our understanding a common structure may be framed as under:

1. **Defining the Problem:** In defining the problem of the research objective, definition of key terms, general background information, limitations of the study and order of presentation should be mentioned in brief.

2. **Review of Existing Literature:** In this head, researcher should study the summary of different points of view on the subject matter as found in books, periodicals and approach to be followed at the time of writing.

3. **Conceptual Framework and Methodology:** Under this head the researcher should first make a statement of the hypothesis. Discussion on the research methodology used, duly pointing out the relationship between the hypothesis and objective of the study and finally discussions about the sources and means of obtaining data should also be made. In this head the researcher should also point out the limitations of methodology, if any, and the natural crises from which the research is bound to suffer for such obvious limitations.

4. **Analysis of Data:** Analysis of the data involves testing of hypothesis from data collected and key conclusions thus arrived.

5. **General Conclusions:** In general conclusions, the researcher should make a restatement of objectives. Conclusion with respect to the acceptance or rejection of hypothesis, conclusion with respect to the stated objectives, suggested areas of further research and final discussion of possible implications of the study for a model, group, theory and discipline.

Finally the researcher should mention about the bibliographies and appendices. The above format is drawn after a standard framework followed internationally in preparation of a synopsis. However, in our country, keeping in view the object of research, style and structure of synopsis varies and quite often it is found that the research guide exercises his own discretion in synopsis preparation than following some acceptable international norms. A standard format for preparation of synopsis commonly used in management and commerce research in India may be drawn as follows:

1. **Introduction:** This includes definition of the problem and its review from a historical perspective.
2. **Objective of the Study:** It defines the research purpose and its specialty from the existing available research in the related field.
3. **Literature Review:** It includes among other things, different sources from which the required abstract is drawn.
4. **Methodology:** It is intended to draw out the sequences followed in research and ways and manners of carrying out the survey and compilation of data.
5. **Hypothesis:** It is a formal statement relating to the research problem and it need to be tested based on the researchers' findings.
6. **Model:** It underlies the nature and structure of the model that the researcher is going to build in the light of survey findings.

2.10 Summary

Exploratory research helps the researcher to become familiar with the problem. It helps to establish the priorities for further research. It may or may not be possible to formulate Hypothesis during exploratory stage. To get an insight into the problem, literature search, experience surveys, focus groups, and selected case studies assist in gaining insight into the problem. The role of moderator or facilitator is extremely important in focus group. There are several variations in the formation of focus group. Conclusive research is designed to assist the decision maker in determining, evaluating and selected the best course of action to take in a given situation. It is helpful in verifying insights and aid decision makers in selecting a specific course of action. The purpose of conclusive research, also known as confirmatory research, is to help decision makers choose the best course of action in a situation. It is especially useful when decision maker already has in mind one or more alternatives and is specifically looking for information pertinent to evaluating them.

2.11 Key Words

- **Exploratory Study** is usually to develop hypotheses or questions for further research.
- **Ex post Facto Design**: In this the researcher has no control over the variables; they can only report what has happened.
- **Literature Research:** It refers to "referring to a literature to develop a new hypothesis".
- **Conclusive Research:** This is a research having clearly defined objectives. In this type of research, specific courses of action are taken to solve the problem.
- **Descriptive Research:** It is essentially a research to describe something and tries to explain relationships among variables.
- **Longitudinal Study:** These are the studies in which an event or occurrence is measured again and again over a period of time.

- **Field Study:** Field study involves an in-depth study of a problem, such as reaction of young men and women towards a product.
- **Causal Study:** A type of conclusive research where the major objective is to obtain evidence regarding cause and-effect (causal) relationships.
- **Single Cross-sectional Design:** A cross-sectional design in which one sample of respondents is drawn from the target population and information is obtained from this sample only.
- **Multiple Cross-sectional Design:** A cross-sectional design in which there are two or more samples of respondents, and information from each sample is obtained only once.
- **Cohort Analysis:** A multiple cross-sectional design consisting of a series of surveys conducted at appropriate time intervals.

2.12 Self Assessment Test

1. Why is research design necessary to conduct a study?
2. What are the various types of research design? Explain with examples.
3. What is exploratory research? Give Example under what circumstances, exploratory research is ideal.
4. What are the sources available for data collection at exploratory stage?
5. What are the characteristic that a moderator should possess while conducting the focus group?
6. What are the various types of descriptive studies? When are they used?
7. Distinguish exploratory from descriptive research.

2.13 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.
- William G. Zikmund, “Business Research Methods”, Thomson, South-Western Publication, Singapore.
- C.R. Kothari, “Research Methodology – Methods and Techniques”, Wilely Eastern Limited, Delhi.

Unit - 3 : Sampling

Unit Structure:

3.0 Objectives
3.1 Introduction
3.2 Essentials of Sampling
3.3 Merits of Sampling
3.4 Limitations of Sampling
3.5 Sampling Methods
3.6 Non-probability Sampling Methods
3.7 Probability Sampling Methods
3.8 Summary
3.9 Key Words
3.10 Self Assessment Test
3.11 References

3.0 Objectives

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

- The concept of sampling.
- The advantages and limitations 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.

3.1 Introduction

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.

Statistical data may be collected by complete enumeration called census enquiry or by partial enumeration called sample enquiry. In the former case the information is collected about each and every item comprising the whole (called universe or population in statistical parlance), while in the later case information is collected about a small number of items which are representative of the whole so as to form an estimate of the characteristics of the whole. If such a sample is adequate representative of the whole, properly drawn and interpreted, then it is most likely to represent the conditions of the whole and can be fairly relied upon as if the observations had been based on complete enumeration. For example, in the field of business while the accounting results are based on complete compilation of business transactions, test checks or test audits cover only a small number of entries to verify the truth of all the entries.

While taking 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 and the number included in the sample are called 'population size' and 'sample size' respectively.

A statistical population is called 'finite' or 'infinite' according to its size. When the number of members of the population can be expressed as a definite quantity the population is said to be 'finite'. Otherwise the population is 'infinite'. In particular, if a sample is known to have been drawn from a continuous probability distribution, then the population is infinite. The primary objective of sampling is to obtain the maximum information about the population with the minimum effort, and also to set out the limits of accuracy of estimates based on sampling.

3.2 Essentials of Sampling

If the sample results have any worthwhile meaning, it is necessary that a sample possesses the following essentials characteristics:

(i) **Representativeness:** A sample should be so selected that it truly represents the universe otherwise the results obtained may be misleading. To ensure representativeness, the random method of selection should be used.

(ii) **Adequacy:** The size of sample should be adequate; otherwise it may not represent the characteristics of the universe.

(iii) **Independence:** All items of the sample should be selected independently of one another and all items of the universe should have the same chance of being selected in the sample. By independence of selection we mean that the selection of a particular item in one draw has influence on the probabilities of selection in any other draw.

(iv) **Homogeneity:** When we talk of homogeneity we mean that there is no basic difference in the nature of units of the universe and that of the sample. If two samples from the same universe are taken, they should give more or less the same unit.

3.3 Merits of Sampling

Given below are some of the merits of sampling:

1. **Less time-consuming:** Since the sample is a study of a part of the population, considerable time and labour are saved when sample survey is carried out. Time is saved not only in collecting data but also in processing it. For these reasons a sample provides more timely data in practice than a census.

2. **Less cost:** The total financial burden of a sample survey is generally less than that of a complete census. This is because of the fact that in sampling, we study only a part of population and the total expense of collecting data is less than that required when the census method is adopted. This is a great advantage particularly in an underdeveloped economy where much of the information would be difficult to collect by the census method for lack of adequate monetary resources.

3. **Reliable results:** Although the sampling involves certain inaccuracies owing to sampling errors, the results obtained are fairly reliable.

There are several reasons for it. First, it is always possible to determine the extent of sampling errors. Secondly, other types of errors to which a survey is subjected, such as inaccuracy of information, incompleteness of returns, etc., are likely to be more serious in a complete census than in a sample survey. This is because more effective precautions can be taken in a sample survey to ensure that information is accurate and complete. For these reasons not only the total error made is smaller in a sample survey but also sample results can also be used with a greater degree of confidence because of our knowledge of the probable size of error. Thirdly, it is possible to avail of the services of experts to impart thorough training to the investigators in a sample survey which further reduces the possibility of errors. Follow-up work can also be undertaken much more effectively in sampling method.

4. Only method sometimes - 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. For example, if one is interested in testing the breaking strength of chalks manufactured in a factory; under the census method all the chalks would be broken in the process of testing. Hence, census method is impracticable and the sample method has to be used. Similarly, if the producer wants to find out whether the tensile strength of a. lot of steel wires meets the specified standard, he must resort to sample method because census would mean complete destruction of all the wires. Also if the population under investigation is infinite, sampling is the only possible solution.

3.4 Limitations of Sampling

Despite various advantages of sampling, it is not completely free from limitations. Some of the difficulties involved in sampling are stated below:

1. 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.
2. 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.
3. At times the sampling plan may be so complicated that it requires more time, labour and money than a complete count. This is so if the size of the sample is a large proportion of the total population and if complicated weighted procedures are used. With each additional complication in the survey, the chances of error multiply and greater care has to be taken which, in turn, means more time and labour.
4. If the information is required for each and every unit in the domain of study, a complete enumeration survey is necessary.

3.5 Sampling Methods

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

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

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

3.6 Non-probability Sampling Methods

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. The selection process is at least partially subjective. The following methods are generally categorised as non-probability sampling methods:

1. **Judgment sampling:** It is a method under which items for the samples are selected generally on certain predetermined criteria. The fixation of criteria and deliberate choice of sampling units may bring in personal element and introduce bias. The selection of items would differ from person to person, at times by personal fancy and judgment of the individual determining the sample. This method is also known as purposive or deliberate sampling method. This is generally used and considered appropriate in small inquiries and researches by individuals, especially when they are familiar with almost all terms of universe. Inferences drawn under this method are not amenable to statistical treatment.

   Though the principles of sampling theory are not applicable to judgment sampling, the method is sometimes used in solving many types of economic and business problems. The use of judgment sampling is justified under a variety of circumstances:

   (i) When only a small number of sampling units are in the universe, simple random selection may miss the more important elements, whereas judgment selection would certainly include them in the sample.

   (ii) When we want to study some unknown traits of a population, some of whose characteristics are known; we may then stratify the population according to these known properties and select sampling units from each stratum on the basis of judgment. This method is used to obtain a more representative sample.

   (iii) In solving everyday business problems and making public policy decisions, executives and public officials are often pressed for time and cannot wait for probability sample designs. Judgment sampling is then the only practical method to arrive at solutions to their urgent problems.

   There are some limitations of this method:

   (i) This method is not scientific because the population units to be sampled may be affected by the personal prejudice or bias of the investigator. Thus judgment sampling involves the risk that the investigator may establish foregone conclusions by including those items in the sample which conform to his preconceived notions. For example, if an investigator holds the view that the wages of workers in a certain establishment are very low, and if he adopts the judgment sampling method, he may include only those workers in the sample whose wages are low and thereby establish his point of view which may be far from the truth. Since an element of subjectiveness is possible, this method cannot be recommended for general use.

   (ii) There is no technique of evaluating the reliability of sample results.

The success of this method depends upon the excellence in judgment. If the individual making decisions is knowledgeable about the population and has good judgment, then the resulting sample may be representative, otherwise the inferences based on the sample may be erroneous. It may be noted that even if a judgment, sample is reasonably representative, there is no objective method for determining the size or likelihood of sampling error. This is a big defect of the method.

2. **Quota Sampling:** It is a method in which population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion. The actual selection of items for the sample is left to the investigator's discretion. This method is convenient and is relatively inexpensive but this allows

some bias to enter into the inquiry; Inferences drawn using this method are not amenable to statistical treatment in a formal way.

Quota sampling and stratified random sampling are similar in as much as in both methods the universe is divided into parts and the total sample is allocated among the parts. However, the two procedures diverge radically. In stratified random sampling the sample within each stratum is chosen at random. In quota sampling, the sampling within each cell is not done at random and the field representatives are given wide latitude in the selection of respondents to meet their quotas.

Quota sampling is often used in public opinion studies. It occasionally provides satisfactory results if the interviewers are carefully trained and if they follow their instructions closely. It is often found that since the choice of respondents within a cell is left to the field representatives, the more accessible and articulate people within a cell will usually be the ones who are interviewed. Slight negligence on the part of interviewers may lead to interviewing ineligible respondents. Even with alert and conscientious field representatives it is often difficult to determine such control category as age, income, educational qualifications, etc.

3. **Block or Cluster Sampling:** It is used when there is unequal concentration of individual units in the universe. Under this method certain blocks or clusters of higher concentration are selected for complete inquiry. Clusters and not 'individual elements' are selected. However, all elements in a selected cluster are studied. This method makes the sampling procedure relatively easier, and may increase the efficiency of field work.

4. **Area Sampling:** It is closer to cluster sampling. Under area sampling the total geographical area (when it happens to be a big one) is divided into a number of smaller non-overlapping areas and then some of the smaller areas are selected and all units of the selected areas constitute the sample. This method generally, makes field interviewing efficient and is suited in inquiries to be conducted over a large area, when the list of population concerned is not available.

5. **Convenience Sampling:** A convenience sample is obtained by selecting 'convenient' population units. The method of convenience sampling is also called the 'chunk'. A chunk refers to that fraction of the population being investigated which is selected neither by probability nor by judgment but by convenience. A sample obtained from readily available lists such as automobile registrations; telephone directories, etc., is a convenience sample and not a random sample even if the sample is drawn at random from the lists. If a person is to submit a project report on, labour management relations in textile industry and he takes a textile mill close to his office and interviews some people over there, he is following the convenience sampling method. Convenience samples are prone to bias by their very nature. Selecting population elements which are convenient to choose almost always make them special or different from the best of the elements in the population in some way.

Hence the results obtained by following convenience sampling method can hardly be representative of the population. They are generally biased and unsatisfactory. However, convenience sampling is often used for making pilot studies. Questions may be tested and preliminary information may be obtained by the chunk before the final sampling design is decided upon.

3.7 Probability Sampling Methods

Probability sampling methods are those in which every item in the universe has a known chance, or probability, of being chosen for the sample. This implies that the selection of sample items is independent of the person making the study, that is, the sampling operation is controlled so objectively that the items will be chosen strictly at random.

1. Simple Random Sampling:

This is a method of sampling in which each unit of the population has exactly the same chance of being included in the sample. Under it, it is the chance alone that determines whether one item or the other is selected. There can be several ways, of selecting items for the sample under random sampling designs but most frequently, items are selected with the help of random number tables. The results obtained under this method can be assessed in terms of probability, i.e., errors of estimation can be measured. This method ensures the law of statistical regularity and generally gives representative sample without selector's bias. In short, this method provides an objective basis for selection.

Random sampling may be of two types: Random sampling with replacement, and random sampling without replacement. In first type, the same unit of population may occur more than once; and in the second, the same population unit cannot come more than once in the sample.

Given below are some of the advantages of simple random sampling:

(a) Since the selection of items in the sample depends entirely on chance, there is no possibility of personal bias affecting the results.

(b) As compared to judgment sampling a random sample represents the universe in a better way. As the size of the sample increases, it becomes increasingly representative of the population.

(c) The analyst can easily assess the accuracy of his estimate because sampling errors follow the principles of chance.

However simple random sampling has following limitations also:

(a) The use of simple random sampling necessitates a completely catalogued universe from which to draw the sample. But it is often difficult for the investigator to have up-to-date lists of all the items of the population to be sampled. This restricts the use of this method in economic and business data where very often we have to employ restricted random sampling designs.

(b) The size of the sample required to ensure statistical reliability is usually larger under random sampling than stratified sampling.

(c) From the point of view of field survey it has been claimed that cases selected by random sampling tend to be too widely dispersed geographically and that the time and cost of collecting data become too large.

(d) Random sampling may produce the most non-random-looking results. For example, thirteen cards from a well-shuffled pack of playing cards may consist of one suit. But the probability of this type of occurrence is very low.

2. Stratified Sampling:

When the population is not homogeneous with respect to a character under study, we divide the population into a number of homogeneous groups (called Strata), which differ in that character from one another but each group is homogeneous within itself. Then units are sampled at random from each of these stratum. The sample size in each stratum, varies according to the relative importance of the stratum in the population. The sample which is the aggregate of the sampled units of each of the stratum is called a Stratified Sample and this technique is called Stratified Sampling. Such a sample can be considered as the representative of the population from which it has been drawn.

The main purposes of stratification are:

(i) to ensure that all sections of the population are adequately represented;

(ii) to avoid a large size of the population;

(iii) to avoid the heterogeneity of the population;

(iv) to bring gain in precision of the estimates obtained.

Given below are some of the advantages of stratified sampling:

(a) Since the population is first divided into various strata and then a sample is drawn from each stratum there is a little possibility of any essential group of the population being completely excluded. A more representative sample is thus secured. It has been rightly pointed out that this type of sampling balances the uncertainty of random sampling against the bias of deliberate selection.

(b) Stratified sampling ensures greater accuracy. The accuracy is maximum if each stratum is so formed that it consists of uniform or homogeneous items.

(c) As compared to random sample, stratified samples can be more concentrated geographically, i.e., the units from the different strata may be selected in such a way that all of them are localised in one geographical area. This would greatly reduce the time and expenses of interviewing.

However, stratified sampling has following limitations also:

(a) Utmost care must be exercised in dividing the population into various strata. Each stratum must contain, as far as possible, homogeneous items otherwise the results may not be reliable. If proper stratification of the population is not done, the sample may have the effect of bias.

(b) The items from each stratum should be selected at random. But this may be difficult to achieve in the absence of skilled sampling supervisors and a random selection within each stratum may not be ensured.

(c) Because of the likelihood that a stratified sample will be more widely distributed geographically than a simple random sample, cost per observation may be quite high.

3. Systematic Sampling:

It involves the selection of sample units at equal intervals, after all the units in the population have been arranged in some order. If the population size is finite, the units may be serially numbered and arranged in order. From the list of these, a single unit is chosen at random. This unit and every k-th unit thereafter constitute a Systematic Sample. In order to obtain a systematic sample of 100 villages out of 10,000, i.e., one out of 100 on an average, all the villages have to be numbered serially. From the first 100 of these, a village is selected at random, suppose with the serial number 6. Then the villages with serial numbers 6,106,206,306, ……… constitute the systematic sample.

The systematic sampling design is simple and convenient to adopt. The time and work involved in sampling by this method are relatively less. The results obtained are also found to be generally satisfactory provided care is taken to see that there are no periodic features associated with the sampling interval. If populations are sufficiently large, systematic sampling can often be expected to yield results similar to those obtained by proportional stratified sampling.

The main limitation of the method is that it becomes less representative if we are dealing with populations having "hidden periodicities". Also if the population is ordered in a systematic way with respect to the characteristics the investigator is interested in, then it is possible that only certain types of items will be included in the population, or at least more of certain types than others. For instance, in a study of workers wages, list may be such that every tenth worker on the list gets wages above Rs. 750 per month.

4. Multi-stage Sampling:

It refers to a sampling procedure which is carried out in several stages. The population is first divided into large groups, called first-stage units. These first-stage units are again divided into smaller units, called second-stage units the second-stage units into third-stage units, and so on, until we reach the ultimate units, e.g., in a survey of rural debt in India, one may like to apply a sampling technique for selection of districts; then villages; then household, and so on.

Multi-stage sampling introduces flexibility in the sampling method which is lacking in the other methods. It enables existing divisions and sub-divisions of the population to be used as units at various stages, and permits the field work to be concentrated and yet large area can be covered. Another advantage of the method is that subdivision into second stage units (i.e., the construction of the second stage frame) need be carried out for only those first stage units which are included in the sample. It is, therefore, particularly variable in surveys of under-developed areas where no frame is generally sufficiently detailed and accurate for subdivision of the material into reasonably small sampling units.

However, a multi-stage sample is in general less accurate than a sample containing the same number of final stage units which have been selected by some suitable single stage process.

3.7.1 Merits of Probability Sampling Methods

Given below are some of the merits of probability sampling methods:

(i) Since the sample is objective and unbiased, it is defensible before the superiors or even before the courts of law.

(ii) The size of the sample does not depend upon the expediency or mere tradition but on demonstrable statistical method and therefore, has a justification for the expenditure involved.

(iii) The degree of deviation from the parameter i.e., the statistical measures based on the population can be estimated and evaluated in terms of certain degree of precision required.

(iv) It provides a more accurate method of drawing conclusion about the characteristics of the population expressed as parameters. Even when qualified persons are not available, it provides an objective basis for selection.

(v) The sample may be combined and evaluated, even though accomplished by different individuals.

(vi) The results obtained can be assessed in terms of probability, and the sample accepted or rejected on a consideration of the extent to which it can be considered representative.

3.7.2 Limitations of Probability Sampling Methods

Despite the great advantages of probability sampling techniques mentioned above it has certain limitations because of which non-probability sampling is quite often used in practice. These limitations are:

(i) Probability sampling requires a very high level of skill and experience for its use.

(ii) It requires a lot of time to plan and execute a probability sample.

(iii) The costs involved in probability sampling are generally large as compared to non-probability sampling.

3.8 Summary

A true representative sample not only gives accurate results but also saves on time, effort, and money. Actually, if primary data are collected using census method, time and cost increases considerably. 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. Probability sampling include simple random, stratified, systematic, and multistage sampling etc. Each of the methods has its own advantages and limitations.

3.9 Key Words

- **Sample:** Small number of items which are representative of the whole, so as to form an estimate of the characteristics of the whole, selected for the purpose of study.
- **Population:** Each and every item comprising the whole.
- **Non-probability Sampling Methods:** 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.
- **Judgment Sampling:** It is a method under which items for the samples are selected generally on certain predetermined criteria.
- **Quota Sampling:** It is a method in which population is first segmented into mutually exclusive sub-groups, just as in stratified sampling. Then judgment is used to select the subjects or units from each segment based on a specified proportion.
- **Cluster Sampling:** Under this method certain blocks or clusters of higher concentration are selected.
- **Convenience Sampling:** A convenience sample is obtained by selecting 'convenient' population units.
- **Probability Sampling Methods:** Probability sampling methods are those in which every item in the universe has a known chance, or probability, of being chosen for the sample.

- **Simple Random Sampling:** This is a method of sampling in which each unit of the population has exactly the same chance of being included in the sample.

- **Systematic sampling:** involves the selection of sample units at equal intervals, after all the units in the population have been arranged in some order.

- **Multi-stage sampling:** refers to a sampling procedure which is carried out in several stages.

3.10 Self Assessment Test

1. "Sampling is necessary under certain conditions". Explain this with suitable examples.

2. Critically examine the various probability sampling methods.

3. What are the methods of sampling? How do you select a sample to study consumer behaviour?

4. Distinguish between random sampling, purposive sampling and stratified sampling. How is a random sample obtained?

5. What is sampling? Point out the merits and demerits of sampling techniques.

6. Discuss the role and limitations of sampling.

3.11 References

- Donald Cooper, Donald R. Cooper, and Pamela S. Schindler, "Business Research Methods", (a(aMcGraw 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.

- 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

Unit - 4 : Scaling Techniques

Unit Structure:

4.0 Objectives
4.1 Introduction
4.2 Nominal Scales
4.3 Ordinal Scales
4.4 Interval Scales
4.5 Ratio Scales
4.6 Examples of Self Rating Scales
4.7 Measurement Scales and Statistical Tests
4.8 Measurement Errors
4.9 Summary
4.10 Key Words
4.11 Self Assessment Test
4.12 References

4.0 Objectives

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.

4.1 Introduction

Measurement is the most fundamental aspect of any research. It 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. 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.

Scaling is 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. As the rules become more restrictive, the kinds of arithmetic operations for which the numbers can be used are increased.

4.2 Nominal Scales

The nominal scale of measurement only satisfies the identity property of measurement. Nominal scale is one in which numbers are used as labels and have no quantitative or numerical significance. Variables such as sex, geographic location, occupation, religion, brand awareness are studied under nominal scales. The numbers serve only as labels or tags for identifying and classifying objects. When used for identification, there is a strict one-to-one correspondence between the numbers and the objects. The numbers do not reflect the amount of the characteristic possessed by the objects. The only permissible operation on the numbers in a nominal scale is counting. Only a limited number of statistics, all of which are based on frequency counts, are permissible, e.g., percentages, and mode. A mean or a median can not be calculated for nominal data. A chi-square statistical test, however, can be conducted to determine if differences between the numbers falling in the various categories is likely to be the result of chance or randomness.

A Nominal measurement scale is used for variables in which each participant or observation in the study must be placed into one mutually exclusive and exhaustive category. For example, categorizing study participants into "male" and "female" categories (i.e., Male = 0 and Female = 1). demonstrates that 'gender' is measured on a nominal scale.

Nominal scales are very simple scales. They consist of assignment of facts/choices to various alternative categories which are usually exhaustive as well mutually exclusive. These scales are just numerical and are the least restrictive of all the scales. In a Nominal Scale numbers are no more than labels and are used specifically to identify different categories of responses. Following example illustrates –

What is your gender? (Please Tick)
[ ] Male
[ ] Female
Another example is – What is your education level? (Please Tick)
[ ] Matriculation
[ ] Graduation
[ ] Post Graduation
Still another example – Indicate your income per year.
[ ] Less than 2,00,000
[ ] 2,00,000 – upto 5,00,000
[ ] 5,00,000 and more

Every observation in the study falls into one, and only one, nominal category. With a nominal measurement scale, there is no relative ordering of the categories — the assignment of numeric scores to each category (Male, Female) is purely arbitrary. Religion and political affiliation are other examples of variables that are normally measured on a nominal scale.

4.3 Ordinal Scales

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. Ordinal scales are used to measure attitudes, preferences, occupation, social classes etc. Ordinal scales help to place different entities like objects, individuals or responses in relative position with respect to a particular aspect. It is ranking scale in which numbers are assigned to objects to indicate the relative extent to which the objects possess some characteristic. It can determine whether an object has more or less of a characteristic than some other object, but not how much more or less. For example, if you think of some type of competition or race (swimming, running), it is possible to rank order the finishers from first place to last place. If someone tells you they finished 2nd, you know that one person finished ahead of them, and all other participants finished behind them. Ordinal variables, however, do not tell us anything about the absolute magnitude of the difference between 1st and 2nd or between 2nd and 3rd. That is, we know 1st was before 2nd, and 2nd was before 3rd, but we do not know how close 3rd was to 2nd or how close 2nd was to 1st. The 1st place finisher could have been a great deal ahead of the 2nd place finisher, who finished a great deal ahead of the 3rd place finisher; or, the 1st, 2nd, and 3rd place finishers may have all finished very close together. Let us consider a few examples:

Rank the following attributes (1 - 5), on their importance in the purchase of a Television.

1. Brand Name
2. Functions
3. Price
4. After sale services
5. Design

The most important attribute is ranked 1 by the respondents and the least important is ranked 5. Instead of numbers, letters or symbols too can be used to rate in a ordinal scale. Such scale makes no attempt to measure the degree of favourability of different rankings.

If there are 4 different types of fertilizers and if they are ordered on the basis of quality as Grade A, Grade B, Grade C, Grade D is again the use of an Ordinal Scale. If there are 5 different brands of Talcom Powder and if a respondent ranks them based on say, “Freshness” into Rank 1 having maximum Freshness Rank 2 the second maximum Freshness, and so on, this is the use of an ordinal scale. In addition to the counting operation allowable for nominal scale data, ordinal scales permit the use of statistics like percentile, quartile etc. A mode or median may be used but not a mean.

4.4 Interval Scales

The interval scale of measurement has the properties of identity, magnitude, and equal intervals. When a variable is measured on an interval scale, the distance between numbers or units on the scale is equal over all levels of the scale. An example of an Interval scale is the Farenheit scale of temperature. In the Farenheit temperature scale, the distance between 20 degrees and 40 degrees is the same as the distance between 75 degrees and 95 degrees. With Interval scales, there is no absolute zero point. For this reason, it is inappropriate to express Interval level measurements as ratios; it would not be appropriate to say that 60 degrees is twice as hot as 30 degrees.

For an example - How do you rate your present Microwave for the following qualities?

Brand Name	Less Known	1	2	3	4	5	Well known
Features	Few	1	2	3	4	5	Many
Price	Low	1	2	3	4	5	High
Quality	Poor	1	2	3	4	5	Good

Such a scale permits the researcher to say that position 5 on the scale is above position 4. Such a scale, however, does not permit conclusion that position 4 is twice as strong as position 2 because no zero position has been established. The data obtained from the Interval Scale can be used to calculate the Mean scores of each attributes over all respondents. The Standard Deviation (a measure of dispersion) can also be calculated.

This scale is used to measure common statistical measures such as range, standard deviation and correlation. A researcher gauges preference, liking or importance of a particular aspect of brand on a continuous basis and distance between one point of scale to another is the same. The location of the zero point is not fixed. Both the zero point and the units of measurement are arbitrary. Any positive linear transformation of the form y = a + bx will preserve the properties of the scale. It is not meaningful to take ratios of scale values.

Virtually the entire range of statistical analyses can be applied to interval scales. Such descriptive measures as the mean, median, mode, range, and standard deviation are applicable. Bivariate correlation analyses, t-tests, analysis of variance tests, and most multivariate techniques applied for purposes of drawing inferences can be used on interval scaled data.

4.5 Ratio Scales

The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and an absolute zero. These are the most powerful of the above scales. Unlike interval scales they have a zero point. Not only are numbers or units on the scale equal over all levels of the scale, but there is also a meaningful zero point which allows for the interpretation of ratio comparisons. Time is an example of a ratio measurement scale. Not only can we say that difference between three hours and five hours is the same as the difference between eight hours and ten hours (equal intervals), but we can also say that ten hours is twice as long as five hours (a ratio comparison).

Now consider the following questions :

Q1) What is your monthly take home salary? _____

Answer to this question has a natural, unambiguous starting point, namely zero. Since starting point is not chosen arbitrarily, computing and interpreting ratio makes sense. For example we can say that a respondent with a monthly take home salary of Rs. 20,000 earns twice as much as one with take home salary of Rs. 10,000.

Ratio scales can be used for all kind of statistical calculation. They are used for variables such as age, number of customers, sales, costs, height, length etc. Only proportionate transformations of the form y = bx, where b is a positive constant, are allowed. All statistical techniques can be applied to ratio data.

Primary Scales of Measurement

Primary Scale	Basic Characteristics	Common Examples	Marketing Examples	Permissible Statistics
Nominal	Numbers identify and classify objects	Social Security numbers; numbering of football players	Brand numbers; store types; sex classification	Percentages; mode
Ordinal	Numbers indicate the relative positions of the objects but not the magnitude of differences between them	Quality rankings; rankings of teams in a tournament	Preference rankings; market position; social class	Percentile; median
Interval	Differences between objects can be compared; zero point is arbitrary	Temperature (Fahrenheit, Celsius)	Attitudes; opinions; index numbers	Range; mean; standard deviation
Ratio	Zero point is fixed; ratios of scale values can be computed	Length; weight	Age; income; costs; sales; market shares	Geometric mean (All)

4.6 Examples of Self Rating Scales

Described below are a few measurement scales that are frequently used in business researches:

a) Graphic Rating Scale -

The respondents rate the objects by placing a mark at the appropriate position on a line that runs from one extreme of the criterion variable to another.

Poor Quality ____ Good Quality

(0) (10)

This line can be vertical or horizontal and scale points may or may not be provided. No other indication is there on the continuous scale. A range is provided. To quantify the responses to question that "indicate your overall opinion about our pop-corns by placing a tick mark at appropriate position on the line", we measure the physical distance between the right extreme position and the response position on the line; the lesser the distance, the more favourable is the response or attitude towards pop-corns. Sometimes options are also indicated on the line.

very easy easy quite easy can't say quite hard hard very hard

||——————|——————|——————|——————|——————|

Its limitation is that coding and analysis will require substantial amount of time, since we first have to measure the physical distances on the scale for each respondent.

b) Likert Scale – The Likert Scale is the most widely used method of scaling in the social sciences today. Perhaps this is because they are much easier to construct and because they tend to be more reliable than other scales with the same number of items. It was developed Rensis Likert. Here the respondents are asked to indicate a degree of agreement and disagreement with each of a series of statement. Each scale item has 5 response categories ranging from strongly agree and strongly disagree.

5	4	3	2	1
Strongly agree	Agree	Indifferent	Disagree	Strongly disagree

Each statement is assigned a numerical score ranging from 1 to 5. It can also be scaled as -2 to +2.

-2 -1 0 +1 +2

Each degree of agreement is given a numerical score and the respondents total score is computed by summing these scores. This total score of respondents reveal the particular opinion regarding a statement.

Two examples of Likert Scales are as follows:

Likert Scales

Please circle the number that represents how you feel about the computer software you have been using

I am satisfied with it
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It is simple to use
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It is fun to use
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It does everything I would expect it to do
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

I don't notice any inconsistencies as I use it
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

It is very user friendly
Strongly Disagree ---1---2---3---4---5---6---7--- Strongly Agree

Likert Scales

Please fill in the number that represents how you feel about the computer software you have been using

I am satisfied with it

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

It is simple to use

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

It is fun to use

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

It does everything I would expect it to do

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

I don't notice any inconsistencies as I use it

1	2	3	4	5
Strongly Agree	Agree	Neither	Disagree	Strongly Disagree

Likert Scales are of ordinal type, they enable one to rank attitudes, but not to measure the difference between attitudes. A typical Likert scale has 20 - 30 statements. While designing a good Likert Scale, first a large pool of statements relevant to the measurement of attitude has to be generated and then from the pool statements, the statements which are vague and non-discriminating have to be eliminated. Thus, likert scale is a five point scale ranging from 'strongly agreement' to 'strongly disagreement'. No judging gap is involved in this method.

When treated as ordinal data, Likert responses can be collated into bar charts, central tendency summarised by the median or the mode (but not the mean), dispersion summarised by the range across quartiles (but not the standard deviation), or analyzed using non-parametric tests, e.g. Chi-square test, Mann-Whitney Test, Wilcoxon signed rank test or Kruskal Wallis Test.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 (Acquiesence Bias); or try to portray themselves or their organization in a more favorable light (Social Desirability Bias).

c) Stapel's Scale -

It was developed by Jan Stapel. This scale has some distinctive features:-

- Each item has only one word/phrase indicating the dimension it represents.
- Each item has ten response categories.
- Each item has an even number of categories.
- The response categories have numerical labels but no verbal labels.

For example, in the following items, suppose for quality of ice cream, we ask respondents to rank from +5 to -5. Select a plus number for words which best describe the ice cream accurately. Select a minus number for words you think do not describe the ice cream quality accurately. Thus, we can select any number from +5,for words we think are very accurate, to -5, for words we think are very inaccurate. This scale is usually presented vertically.

+5

+4

+3

+2

+1

High Quality

-1

-2

-3

-4

-5

This is a unipolar rating scale.

d) Thurston Scales -

In psychology , the **Thurstone scale** was the first formal technique for measuring an attitude. It was developed by Louis Leon Thurston in 1928, as a means of measuring attitudes towards religion. It is made up of statements about a particular issue, and each statement has a numerical value indicating how favorable or unfavorable it is judged to be. People check each of the statements to which they agree, and a mean score is computed, indicating their attitude.

The Thurstone attitude measurement procedure is generally more consistent with empirical characteristics of disagree-agree responses, It is constructed by th method of equal appearing intervals, in which a large pool of candidate statements about an attitude object, ranging from strongly negative (Abortion is never justified) through neutral (There are arguments both for and against abortion) to strongly positive (Abortion is every woman's right), are sorted by a group of judges into eleven categories. They are assumed to appear equally spaced on the attitude continuum, according to how favorable the statements are towards the attitude object. Items that yield the highest level of agreement among the judges as to their scale position, and that collectively represent an adequate range of contents and scale positions, are then selected for the final scale. Respondents to the scale endorse just those items with which they agree, and an individual respondent's score is calculated as the mean (or occasionally median) of the items endorsed, such scores being assumed to lie on an interval scale of measurement. It is also called an equal-appearing interval scale.

Thurstone was one of the first and most productive scaling theorists. He actually invented three different methods for developing a unidimensional scale: the method of equal-appearing intervals; the method of successive intervals; and, the method of paired comparisons. The three methods differed in how the scale values for items were constructed, but in all three cases, the resulting scale was rated the same way by respondents.

e) Guttman Scales/Scalogram Analysis –

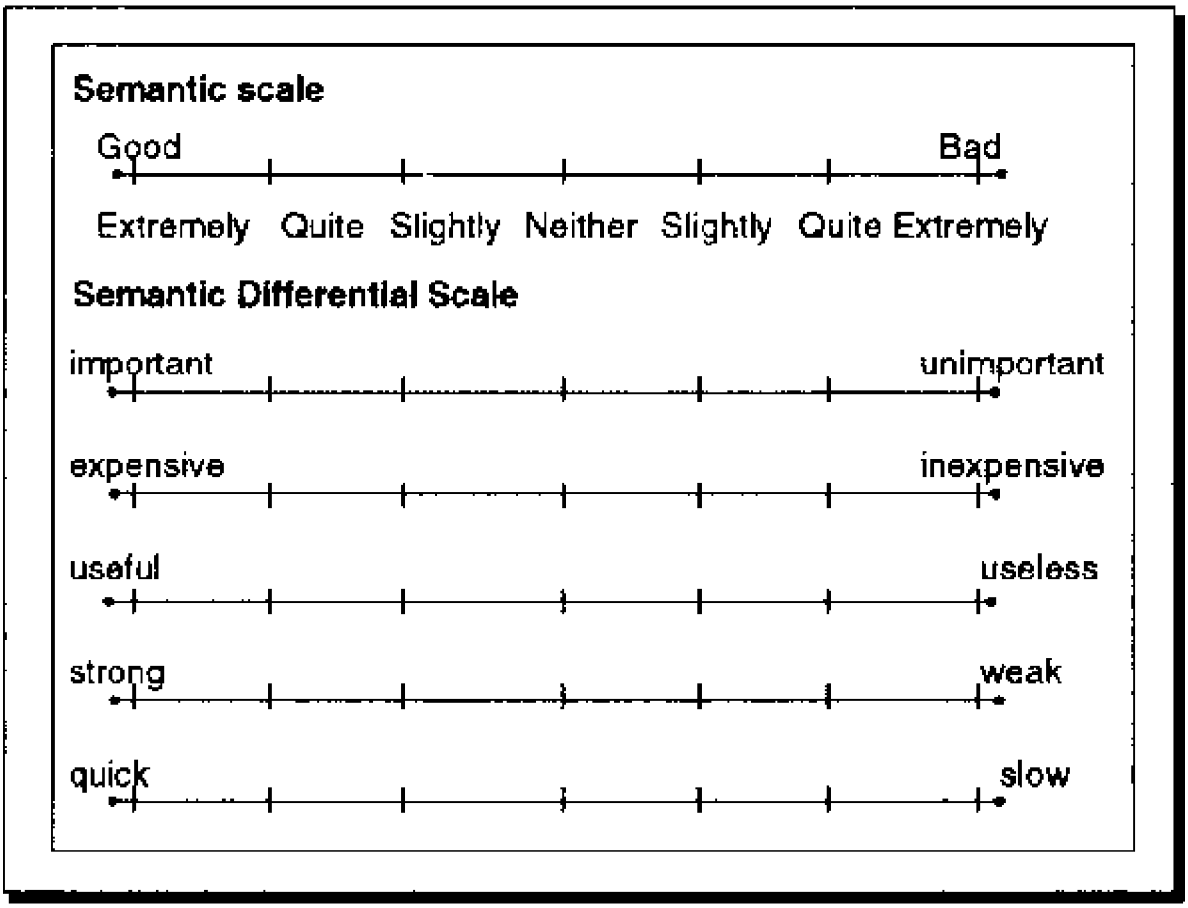
This is a scale which is named after the US (later Israeli) psychologist Louis H. Guttman (1916–87). It is based on the idea that items can be arranged along a continuem in such a way that a person who agrees with an item or finds an item acceptable will also agree with or find acceptable all other items expressing a less extreme position. For example - Children should not be allowed to watch indecent programmes or government should ban these programmes or they are not allowed to air on the television. They all are related to one aspect.In this scale each score represents a unique set of responses and therefore the total score of every individual is obtained. This scale takes a lot of time and effort in development.Theyare very commonly used in political science, anthropology, public opinion, research and psychology.

f) The Q Sort Scales -

It is used to discriminate among large number of objects quickly. It uses a rank order procedure and the objects are sorted into piles based on similarity with respect to some criteria. For example, on the basis of taste we classify the brands into tasty, moderate and non tasty. We can classify on the basis of price also- Low, medium, high. Then we can attain the perception of people that whether they prefer low priced brand, high or moderate. We can classify brands or pile it into three piles. So the number of objects is to be placed in three piles-low, medium or high. Thus, the Q-sort technique is an attempt to classify subjects in terms of their similarity to attribute under study.

g) Semantic Differential Scales -

This type of scale makes extensive use of words rather than numbers. Respondents describe their feelings about the products or brands on scales with semantic labels. When bipolar adjectives are used at the end points of the scales, these are termed semantic differential scales.



4.7 Measurement Scales and Statistical Tests

One of the primary purposes of classifying variables according to their level or scale of measurement is to facilitate the choice of a statistical test used to analyze the data. There are certain statistical analyses which are only meaningful for data which are measured at certain measurement scales. For example, it is generally inappropriate to compute the mean for Nominal variables. Suppose you had 20 subjects, 12 of which were male, and 8 of which were female. If you assigned males a value of '1' and females a value of '2', could you compute the mean sex of subjects in your sample? It is possible to compute a mean value, but how meaningful would that be? How would you interpret a mean sex of 1.4? When you are examining a Nominal variable such as sex, it is more appropriate to compute a statistic such as a percentage (60% of the sample was male). When a researcher wishes to examine the relationship or association between two variables, there are also guidelines concerning which statistical tests are appropriate. For example, let's say a University administrator was interested in the relationship between student gender (a Nominal variable) and major field of study (another Nominal variable). In this case, the most appropriate measure of association between gender and major would be a Chi-square test. Let's say our University administrator was interested in the relationship between undergraduate major and starting salary of students' first job after graduation. In this case, salary is not a Nominal variable; it is a ratio level variable. The appropriate test of association between undergraduate major and salary would be a one-way Analysis of Variance (ANOVA), to see if the mean starting salary is related to undergraduate major.

Finally, suppose we were interested in the relationship between undergraduate grade point average and starting salary. In this case, both grade point average and starting salary are ratio level variables. Now, neither Chi-square nor ANOVA would be appropriate; instead, we would look at the relationship between these two variables using th e Pearson Correlation Coefficient..

In the social and behavioral sciences, however, much of what we study is measured on what would be classified as an ordinal level. We often ask if people "Strongly Disagree", "Slightly Disagree", or are "Neutral" to a series of statements. We then assign a value of '1' if they Strongly Disagree with a statement, up to a '5' if they Strongly Agree with a statement. To be sure, this type of measurement is ordinal, in the sense that "Strongly Agree" reflects more agreement than "Slightly Agree". This type of measurement is **not** an interval or a ratio level of measurement, because we can not state for certain that the interval between "Strongly Disagree" and "Slightly Disagree" is equivalent to the interval between "Slightly Disagree" and "Neutral". Nor can we say that there is an absolute zero point for level of agreement. However, if we were to rigidly follow the rules of "permissible" analyses for ordinal variables, many of the analyses we conduct in social sciences research would be deemed impermissible. On the other hand, some scientists have conducted computer simulations to try and find out what would
happen if we violated certain "rules" of data analysis. They have found that for the most part, it is alright to treat ordinal data (such as variables which have been measured using Strongly Disagree to Strongly Agree response alternatives) as though it were interval level data, and conduct statistical tests that are appropriate for interval level data.

Before finalizing on the scale usage, it is, however, advisable to consider the following aspects:

- Type/details of information required.
- Time and cost constraints.
- Literary level of target respondents (for easy communication)
- Stimulus characteristics
- Analysis methodology to be used
- Past experience to be used

4.8 Measurement Errors

There are various measurement errors which are as follows:

a) **Reflection of additional stable characteristics** – Perhaps the most troublesome measurement error occurs when the measurement reflects a stable characteristic of the object or event in addition to the one of interest to the researcher. The extraneous variables as gender, education, age etc. have been found to be sources of bias in the measurement of attitudinal reactions. Stable characteristics of respondent are particularly troublesome when they differ across respondent groups of interest such as gender, culture, sub-culture, nationality etc. For example, the Japanese are generally reluctant to say 'no' directly. Similarly Hispanics are more likely than other groups to provide positive responses as per a research study. This issue, therefore, must be given careful consideration when conducting multi-national or sub-cultural research studies.

b) **Temporary characteristics of the respondent** – An equally common source of error is the influence of short term characteristics of the object. Such factors as fatigue, health, hunger and emotional state may influence the measure of the other characteristics. For example, responses of a person may differ if the person is in a bad mood because of cold or fatigue. Researchers, however, frequently assume that such temporary fluctuations are randomly distributed in their effect on the measurement and will cancel each other out.

c) **Situational characteristics** – Many measurements that involve human subjects reflect both the true characteristic under consideration and the characteristics under which the measurement is taken. For example, husbands and wives tend to report one level of influence in a purchase decision if their spouses are present and another level if their spouses are absent.

d) **Characteristics of measurement process** – The measurement also can include influences from the method of gathering the data. Gender, age, ethnic background, and style of dress of the interviewer influence an individual's response patterns on certain questions. In addition, various methods of contact – telephone, mail, personal interview, and the like – sometimes alter response patterns.

e) **Characteristics of measuring instrument** – Aspects of the measuring instrument itself can cause constant or random errors. Unclear instructions, ambiguous questions, confusing terms, irrelevant questions, and omitted questions can all introduce errors.

f) **Characteristics of the response process** – Response errors are another reason why responses may not reflect the 'true' characteristic accurately. For example, a respondent may have inadvertently checked a positive response when the intention was to check a negative one.

g) **Characteristics of the analysis** – Finally, mistakes can occur in interpreting, coding, tabulating, and analyzing an individual's or a group's response. For example, researcher might enter 8 instead of 3 for a response.

4.9 Summary

Measurement 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. Measurement 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. As the rules become more restrictive, the kinds of arithmetic operations for which the numbers can be used are increased.

4.10 Key Words

- **Scales of Measurement**: Instruments and/or mechanism for measuring qualitative and quantitative responses.
- **Nominal Scale**: Type of scale where numbers are used as labels or tags for identifying and classifying objects.
- **Ordinal Scale**: It is ranking scale in which numbers are assigned to objects to indicate the relative extent to which the objects possess some characteristic.
- **Interval Scale**: It is a scale of measurement that has the properties of identity, magnitude, and equal intervals.
- **Ratio Scale**: The ratio scale of measurement satisfies all four of the properties of measurement: identity, magnitude, equal intervals, and an absolute zero.

4.11 Self Assessment Test

1. Discuss various types of scales of measurement highlighting their respective usage.
2. Differentiate between ordinal and interval scales citing suitable examples.
3. Write short notes on the following:
   (a) Likert scale
   (b) Thurston scale
   (c) Semantic Differential Scale
4. Discuss various scales that may be used to measure attitudes.
5. Discuss some of the measurement errors.

4.12 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.
- 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

Unit - 5 : Data Collection

Unit Structure:

5.0 Objectives
5.1 Introduction
5.2 Importance of Data
5.3 Types of Data
5.4 Primary Data Collection Methods
5.5 Secondary Data Collection Methods
5.6 Summary
5.7 Key Words
5.8 Self Assessment Test
5.9 References

5.0 Objectives

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.

5.1 Introduction

Whenever the research work is carried out there is need for information. Information means the communication of knowledge about an event of a given condition or the extension and multiplication of knowledge resulting from observation, study, experience or education. Knowledge, facts, data, news, message etc. are the terms used as synonym or near synonyms to information. All type of information collected without proper aim or objective is of no use. For example: In 6^{th} standard school student's height ranges from 4 feet to 5 feet and on average 2 to 5 students are absent daily. So these are not data. We acquire the information about people and things through collecting data. Data collection is the vehicle through which researchers collect information to answer their research questions and define their conclusions and recommendations based on the finding from research. Data collection means gathering information to address those critical evaluation questions which have been identified earlier in the evaluation process. These are many methods available to gather information, and a wide variety of information sources. The important aspect of data collection is selecting the most appropriate information or evidence that can satisfy the problem question. To plan data collection, first question reads to be formed and the information source available

5.2 Importance of Data

The data serve as the bases or raw materials for analysis. Without an analysis of factual data, no specific inferences can be drawn on the questions under 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.

Data form the basis for testing the hypotheses formulated in a study. Data also provide the facts and figures required for constructing measure-ment 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.

Data collection in marketing research is a detailed process in which a planned search for all relevant data is made by researcher and some inferences are drawn an specific issues or problems. While deciding about the methods of data collection to be used for the study, the researchers should consider two type of sources for collecting data - Secondary and Primary. Example of secondary sources include the use of census data to obtain information on the age- sex structure of a population; the use of hospital records to find out the morbidity and mortality pattern of a community; and the collection of data from sources such as articles of generals, magazines, books and periodicals to obtain historical and other types of information . on other side, finding out number of people in a country, determining the job satisfaction of the employees of an organization, and knowing the preferences of the customers are examples of information collected from primary sources.

5.3 Types of Data

The data can be classified into primary data and secondary data.

5.3.1 Primary Data

Primary data is the data which is collected first hand specially for the purpose of study 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 collected through various methods such as observation, interviewing, mailing etc. It is collected for addressing the problem at hand. 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.”

Given below are some of the advantages of primary data:

- As it is collected by researcher itself so they are original in nature.
- 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.

Given below are some of the disadvantages of primary data

- It is very time consuming process as researchers has to collect all data personally.
- Also an expensive affair. As to visit individual respondents or to collect data from individual respondent requires heavy expenses.
- Chances of biasness are at great extent. If the researchers work is to be completed in a given time period then possibility are there that on part of researcher biasness can be seen either to approach the respondent; or in receiving the response.

- 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.

5.3.2 Secondary Data

Secondary Data is the data that have been already collected by readily available from other sources. Such data are less expensive and obtain more quickly than the primary data and also may be available when primary data cannot be obtained at all. These data might have been collected and compiled for another purpose. The secondary sources consist of readily available compiled statistical statements and reports whose data may be used by researches for their studies, e.g., census reports, annual reports and financial statements of companies, reports of government departments, annual reports on currency and finance statistical Statements relating to cooperatives, federal cooperative commission, commercial Banks and Micro Finance Credit Institutions reports of the national sample survey organisation, 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 firms and organisations, e.g., accounting and financial records, personnel records, register of members, minutes of meetings, inventory records, etc.

Given below are some of the advantages of secondary data:

- Secondary data is economical. It saves efforts and expenses.
- It is time saving as wider geographical area and longer reference period may be covered without much cost.
- It helps to make primary data collection more specific since with the help of secondary data, we are able to identify the gaps or deficiencies and what additional information is required to be collected.
- It helps to understand the problem.
- It act as a bench mark or a basis of comparison for the data that is collected by the researcher.
- Helps in exploring the hidden area of research.
- The use of secondary data broadens the database from which scientific generalizations can be made.

Given below are some of the disadvantages of secondary Data:

- Accuracy of secondary data is not known. It is not easy to find out under what circumstances the secondary data is collected or there can be doubt if the organization to which the data belongs is not authenticate or not having credibility.
- Data may be outdated. It may not match with the current research because of time lag in producing them.
- Units or variables of secondary data may differ from the current research, e.g. the size of firm can be expressed as (i) member of employees, (ii) paid up capital employed, (iii) gross sales, (iv)gross or net profit etc. So it is not necessary that it may suit to the current research project.

5.4 Primary Data Collection Methods

The researcher directly collects primary data from their original sources. In this case, the researcher can collect the required data precisely according to his research needs, he can collect them when he wants them and in the form he needs them. But the collection of Primary data is costly and time consuming. Yet, for several types of social science research such as socio-economic surveys, social anthropological studies of rural communities and tribal communities, sociological studies of social problems and social institutions, marketing research, leadership studies, opinion polls, attitudinal surveys, readership, radio listening and T.V. viewing surveys, knowledge-awareness practice (KAP) studies, farm management studies, business management studies, etc., required data are not available from secondary sources and they have to be directly gathered from the primary sources. In all cases where the available data are inappropriate, inadequate or obsolete, primary data have to be gathered

There are various methods of primary data collection. A 'Method' is different from a 'Tool'. While a method refers to the way or mode of gathering data, a tool is an instrument used for the method. For example, a schedule is used for interviewing. Each method has its advantages and disadvantages. No method is universal. Each method's unique features should be compared with the needs and conditions of the study and thus the choice of the methods should be decided.

The important methods of collecting primary data are:

5.4.1 Observation Method:

Observation means viewing or seeing. We go on observing something or other while we are awake. Most of such observations are just casual and have no specific purpose. But observation as a method of data collection is different from such casual viewing. Observation may be defined as a systematic viewing of a specific phenomenon in its proper setting or the specific purpose of gathering data for a particular study. Observation as a method includes both 'seeing' and 'hearing.' It is accompanied by perceiving as well. In this method of data collection, the researcher or investigator collect information by the way of own direct observation of relevant people, actions and situations without interviewing the respondents. The researcher collects the information of the current events and it has no relevance with the past behavior or future intentions or attitudes of respondents. As the researcher has to individually examine the respondent so this method is not only expensive but also time consuming. Also feelings, beliefs and attitudes that motivate buying behaviour and infrequent behaviour cannot be observed.It usually happens what is seen may or may not be true. This type of method is used in the researches related to behavioral sciences.

Example:

A food service operator sends researchers into competing restaurants to learn menu items prices, check portion sizes and consistency and observe point-of purchase merchandising.

A restaurant evaluates possible new locations by checking out locations of competing restaurants, traffic patterns and neighborhood conditions.

Through observation the information can be yield which people are normally unwilling or unable to provide. **Example 6:** Observing numerous plates containing uneaten portions of the same menu items indicates that food is not satisfactory.

Types of Observation:

Observation can be following types:

a) **Structured Observation:** A systematic method of data collection, where there is considerable pre-coding and the observation takes the form of recording when, how often, or for how long the pre-coded behaviors occur. In this method people are observed in their natural environments using specified methods and measurement procedures. Structured observation permits stronger generalizations and has checks on reliability and validity. This is more like a survey, where every respondent is asked the same set of questions. But in this case, questions are not asked. Instead, particular types of behavior are looked for and counted. Structured observation places more emphasis on identifying and recording the frequency of people's actions. The aim of this quantitative approach is to plan every aspect of the research before data collection takes place. It is important that there are set instructions for the observer to follow on precisely when and what to observe and how the data should be recorded. For example, standard collection sheets should be used to record, tick or mark the type of behaviour being observed. The idea is to be as objective as possible to avoid any contamination of the results. Under this type of method data is reliable and easy to collect, cost effective and simultaneously several observations can be run, and ethical considerations are also taken into account. The disadvantages of structured observation are- the data collected may be shallow and there is no way of noting any unusual happenings as recording methods have already been pre-determined.

b) **Unstructured Observation:** It can be considered as early phase of the research. In unstructured or non-systematic observations, one or more observers provide narrative or qualitative accounts of their observations, without using structured, numerical categories to observe and record target problems. The purpose of such unstructured observation is to provide a richer and more direct account of the behavioral phenomenon under study. In this approach, an attempt is made to understand and analyze the complexities of a particular situation without imposing any rigid structure over it. Flexibility and the absence of imposed structure allow an observer to gain authentic information. In unstructured observation, the observer monitors all aspects of the phenomenon that seem relevant to the problem at hand, e.g., observing children playing with new toys.

c) **Participant Observation:** Participant observation refers where the researcher takes on a role in the social situation under observation. The social researcher immerses themselves in the setting under study, getting to know key actors in that location in a role which is either covert or overt, although in practice, the researcher will often move between these two roles. This enables them to develop a deeper understanding of the motives and actions of the people they are studying. The advantages of participant observation are-observations of real life in natural settings produce sound data, provides a rounded picture, and researchers can make "on the spot" evaluations of behaviour as it occurs. On the other hand there are some disadvantages also. Ethical issues can arise, researchers may lose their objectivity or even influence behaviour and outcomes, and reliability of data collected can be low as observations are often personal and non-repeatable.

Covert observation involves:

• the social researcher participating fully without informing members of the social group of the reasons for her presence, thus the research is carried out secretly or covertly.

• contact with a 'gatekeeper', a member of the group under study who will introduce the researcher into the group.

Overt observation involves:

• the researcher being open about the reason for her presence in the field of study since the researcher is given permission by the group to conduct her research.

- the use of a 'sponsor', who is an individual likely to occupy a high status within the group, therefore lessening any potential hostility towards the research

d) **Non- participant observation:** When the observer does not actually participate in the activities of the group, but simply observes them from a distance, it is known as a non-participant observation. In this observation the researcher remains passive observers, who watches and listens to the activities and drawing conclusions from the group activities.

For example: The observer or researcher want to study the sales person response to the customer. An observer could watch, follow and record the activities as they performed. After making a number of observations, conclusions could be drawn about the behavior salesperson carried out in the store.

Given below are some of the the advantages of observation method:

1. Data can be collected from the place when an event or activity is occurring.
2. Does not rely on people's willingness or ability to provide information.
3. Allows direct observation of what people do rather than relying on what people say they did.

Given below are some of the the disadvantages of observation method:

1. Susceptible to observer bias.
2. Susceptible to the "hawthorne effect," that is, people usually perform better when they know they are being observed, although indirect observation may decrease this problem.
3. Can be expensive and time-consuming compared to other data collection methods.
4. Does not increase your understanding of why people behave as they do.

5.4.2 Interview:

Interviewing is one of the major methods of data collection. It is a oral questioning of respondents either with an individual or to a group of persons. It may be defined as two-way systematic conversation between an investigator and an informant, initiated for obtaining information relevant to as a specific study. It is a face to face interaction with the respondent. During interview a response are recorded either by writing or recording or by combination of both.It involves not only conversation, but also learning from the respondents' gestures, facial expressions and pauses, and his environment. It is done by using a structured schedule or an unstructured guide. The purpose of interview is to know or gain certain information from the respondent which cannot be obtained or gathered from any source. The interviews are generally based on the past incidences, feelings, experiences or the reaction of the individual towards the subject which are to be listed by the researcher.

Interviewing may be used either as a main method or as a supplementary one in studies of persons. Interviewing is the only suitable method for gathering information from illiterate or less educated respondents. It is useful for collecting a wide range of data from factual demographic data to highly personal and intimate information relating to a person's opinions, attitudes, values, beliefs, past experience and future intentions. When qualitative information is required or probing is necessary to draw out fully, then interviewing is required. Where the area covered for the survey is a compact, or when a sufficient number of qualified interviewers are available, personal interview is feasible.

Interview is often superior to other data-gathering methods. People are usually more willing to talk than to write. Once rapport is established, even confidential information may be obtained. It permits probing into the context and reasons for answers to questions. Interview can add flesh to statistical information. It enables the investigator to grasp the behavioural context of the data furnished by the respondents. It permits the investigator to seek clarifications and brings to the forefront those questions, that, for one reason or another, respondents do not want to answer.

Types of Interviews

The interview may involve presentation of oral-verbal stimuli and reply in terms of oral-verbal responses. This method can be are through personal interviews i.e. face to face or through telephone.

1) **Personal Interview**

It is very flexible and can be used to collect large amounts of information. Trained interviewers can hold the respondent's attention and are available to clarify difficult questions.They can guide interviews, explore issues, and probe as the situation requires. Personal interview can be used in any type of questionnaire and can be conducted fairly quickly. Interviewers provide a guard against confusing items. If a respondent has misunderstood a question, the interviewer can clarify, thereby obtaining relevant responses. Personal interviews are a good way to gather information from community leaders, particularly those who might be unwilling or too busy to complete a written survey.

Some of the advantages of the personal interview are:

- **Flexibility-** Allows flexibility in the questioning process and allows the interviewer to clarify terms that are unclear.
- **Control of the interview situation -** Can ensure that the interview is conducted in private, and respondents do not have the opportunity to consult one another before giving their answers.
- **High response rate -** Respondents who would not normally respond to a mail questionnaire will often respond to a request for a personal interview.
- **Characteristics of respondent assessed** – During interview the interest of the respondent can be easily assessed through the tone of voice, facial expression, hesitation, etc.

Some of the disadvantages are:

- **Higher cost -** Costs are involved in selecting, training, and supervising interviewers; perhaps in paying them; and in the travel and time required to conduct interviews.
- **Interviewer bias -** The advantage of flexibility leaves room for the interviewer's personal influence and bias, making an interview subject to interviewer bias.
- **Lack of anonymity -** Often the interviewer knows all or many of the respondents. Respondents may feel threatened or intimidated by the interviewer, especially if a respondent is sensitive to the topic or to some of the questions.
- **Time Consuming** – It took time if person is interviewed individually.

Personal interview are usually carried out in structured way. Interviews may be classified into: (a) structured or directive interview, (b) unstructured or non-directive interview, (c) focused interview, and (d) in- depth interview .

a) Structured Interview: In a structured interview the interviewer has a pre-determined list of questions with answers. The candidates interviewed are asked questions only from the prepared list. Marks may be awarded to the candidates based on the answer they give and thereby their suitability for the job may be assessed. The person asking the questions ("the interviewer") is allowed to explain things the interviewee (or "respondent" - the person responding to the questions) does not understand or finds confusing. Structured interview method is used when data is to be collected from the large number of persons. It helps in systematic collection, comparison, organization and analysis of data within a limited period of time. Thus, in this method the interviewer follows a rigid procedure laid down, asking questions in a form and order prescribed.

This method has the following strengths:

1. It enables the researcher to examine the level of understanding a respondent has about a particular topic - usually in slightly more depth than with a postal questionnaire.
2. It can be used as a powerful form of formative assessment. That is, it can be used to explore how a respondent feels about a particular topic before using a second method (such as observation or indepth interviewing) to gather a greater depth of information. Structured interviews can also be used to identify respondents whose views you may want to explore in more detail (for example : through the use of focused interviews).
3. All respondents are asked the same questions in the same way. This makes it easy to repeat ("replicate") the interview. In other words, this type of research method is easy to standardise.
4. Provides a reliable source of quantitative data.
5. The researcher is able to contact large numbers of people quickly, easily and efficiently
6. It is relatively quick and easy to create, code and interpret (especially if closed questions are used).
7. There is a formal relationship between the researcher and the respondent with the latterknowing exactly what is required from them in the interview If, for example, a respondent is unable or unwilling to answer a question the researcher (because they are present at the interview) is aware of the reasons for a failure to answer all questions.
8. The researcher does not have to worry about response rates, biased (self-selected) samples, incomplete questionnaires and the like.

This method has the following weaknesses:

1. Can be time consuming if sample group is very large (this is because the researcher or their representative needs to be present during the delivery of the structured interview).
2. The quality and usefulness of the information is highly dependent upon the quality of the questions asked. The interviewer cannot add or subtract questions.
3. A substantial amount of pre-planning is required.
4. The format of questionnaire design makes it difficult for the researcher to examine complex issues and opinions.
5. There is limited scope for the respondent to answer questions in any detail or depth.
6. There is the possibility that the presence of the researcher may influence the way a respondent answers various questions, thereby biasing the responses. For example, an aggressive interviewer may intimidate a respondent into giving answers that don't really reflect the respondent's beliefs.

7. A problem common to both postal questionnaires and structured interviews is the fact that by designing a "list of questions", a researcher has effectively decided - in advance of collecting any data - the things they consider to be important and unimportant.

b) Unstructured Interview: In this case the interviewer does not keep any list of questions. He can ask any question that he thinks is relevant and see how the interviewee responds. Sometimes the interviewer may ask the candidate to express his views on the job he has applied for, the work he did earlier, the organizations he served earlier, the present organization and so on. The interviewer, in this kind of an interview, is a patient listener. He does not interrupt the interviewee nor does he give his opinions on what is stated.

Advantages of unstructured interviews:

Some of the advantages of unstructured interviews are:

- Unstructured interviews are good in the initial stages of the project as they provide a general understanding of the problem.
- Permit full exploration of ideals and beliefs.
- It is more like a conversation and the interviewee is relaxed.

Disadvantages of unstructured interviews:

Some of the disadvantages of unstructured interviews are:

- The information from the expert may be vast and too unrelated for the engineer to unravel.
- There is a temptation on the part of the expert to discuss more unusual and different areas that he/ she deals with, whereas the primary purpose of the interview is to concentrate on the central issues.
- They also tend to be time consuming and difficult to analyze, so it is usual to move to a structured interview or special method once the knowledge engineer is reasonably sure of his/her ground.
- Attention not focused on a given issue.
- Very little factual information is provided.
- Less details provided on general concepts and objects.

c) Focused Interview: The method of interviewing participants in focus groups is widely adapted to collect primary data. The groups are generally composed of 7 to 10 people (although groups range from as small as 4 to as large as 12) who are unfamiliar with one another and have been selected because they share certain characteristics relevant to the study's questions. The interviewer creates a supportive environment, asking focused questions to encourage discussion and the expression of differing opinions and points of view. These interviews may be conducted several times with different individuals so that the researcher can identify trends in the perceptions and opinions expressed, which are revealed through careful, systematic analysis.This method assumes that an individual's attitudes and beliefs do not form in a vacuum: People often need to listen to others' opinions and

understandings to form their own. the questions in a focus-group setting are deceptively simple. The trick is to promote the participants' expression of their views through the creation of a supportive environment.

The advantages of focus-group interviews are that this method is socially oriented, studying participants in an atmosphere more natural than artificial experimental circumstances and more relaxed than a one-to-one interview. When combined with participant observation, focus groups are especially useful for gaining access, focusing site selection and sampling, and even for checking tentative conclusions.

There are, however, certain disadvantages to this method as well: First and foremost is the issue of power dynamics in the focus-group setting, time can be lost while dead-end or irrelevant issues are discussed; the data are difficult to analyze because context is essential to understanding the participants' comments; the method requires the use of special room arrangements and highly trained observer moderators; the groups can vary a great deal and can be hard to assemble; and logistical problems may arise from the need to manage a conversation while getting good quality data.

d) In-depth Interview: In-depth interviews are one-to-one encounters in which the interviewer makes use of an unstructured or semi-structured set of issues/topics to guide the discussion. The object of the exercises is to explore and uncover deep-seated emotions, motivations and attitudes. They are most often employed when dealing with sensitive matters and respondents are likely to give evasive or even misleading answers when directly questioned. Most of the techniques used in the conduct of depth interviews have been borrowed from the field of psychoanalysis. Depth interview are usually only successful when conducted by a well trained and highly skilled interviewer.

Other instances when depth interviewers can be particularly effective are: where the study involves an investigation of complex behaviour or decision-making processes; when the target respondents are difficult to gather together for group interviewers (e.g. farmers, veterinary surgeons, haulage contractors, government officials); and where the interviewee is prepared to become an informant only if he/she is able to preserve his/her anonymity.

2) Telephone Interview

Telephone interviewing is a non-personal method of data collection. Telephone surveys are a quantitative marketing research methodology that involve calling and interviewing a representative sample of people within a geographic area or a targeted market served by a business or organization

Advantages of Interview Method:

- Quick method of interview
- More flexible as interviewer can explain questions not understood by the respondent
- Depending on respondent's answer interviewer can skip some questions and probe more on others
- Allows greater sample control
- Response rate tends to be higher than mail
- Easily administered

Disadvantages of Interview Method:

- Cost per respondent is higher
- Some people may not want to discuss personal questions with interviewer
- Interviewer's manner of speaking may affect the respondent's answers
- Different interviewers may interpret and record response in a variety of ways

- Under time pressure ,data may be entered without actually interviewing
- Visual aids cannot be used

5.4.3 Survey Method:

This approach is most suited for gathering descriptive information. Surveys are based on short interviews either face-to-face or on telephone. These are based on simple questions. Survey research is the appropriate mode of inquiry for making inferences about a large group of people based on data drawn from a relatively small number of individuals in that group. Its basic aim is to describe and explain statistically the variability of certain features in a population. The general logic of survey research gives a distinctive style to the research process; the type of survey instrument is determined by the information needed.

Surveys have definite advantages when the goals of research require obtaining quantitative data on a certain problem or population. Large surveys often focus on sensitive or controversial topics within the public domain. The strengths of surveys include their accuracy, generalization, and convenience. Surveys are conducted in three ways: by mail, telephone, and personal interview. Survey can be performed by sending a questionnaire and quantifying the response when questionnaires are returned.

Advantages:

- Can be used to collect many different kinds of information.
- Quick and low cost as compared to observation and experimental method.

Limitations:

- Respondent's reluctance to answer questions asked by unknown interviewers about things they consider private.
- Busy people may not want to take the time
- May try to help by giving pleasant answers
- Unable to answer because the respondent cannot remember or never gave a thought to what they do and why
- May answer in order to look smart or well informed.

Questionnaire:

The mail survey is one of the method of collecting primary data. This method involves sending questionnaires to the respondents with a request to complete them and return them by post. This can be used in the case of educated respondents only. The mail questionnaire should be simple so that the respondents can easily understand the questions and answer them. It should preferably contain mostly closed-end and multiple-choice questions so that it could be completed within a few minutes.

The distinctive feature of the mail survey is that the questionnaire is self-administered by the respondents themselves and the responses are recorded by them, and not by the investigator as in the case of personal interview method. It does not involve face-to-face conversation between the investigator and the respondent. Communication is carried out only in writing and this requires more cooperation from the respondents than does verbal communication.

In contrast with interviews, where an enumerator poses questions directly, questionnaires refer to forms filled in by respondents alone. Questionnaires can be handed out or sent by mail and later collected or returned by stamped addressed envelope. This method can be adopted for the entire population or sampled sectors.

In order to maximise return rates, questionnaires should be designed to be as simple and clear as possible, with targeted sections and questions. Most importantly, questionnaires should also be as short as possible. If the questionnaire is being given to a sample population, then it may be preferable to prepare several smaller, more targeted questionnaires, each provided to a sub-sample. If the questionnaire is used for a complete enumeration, then special care needs to be taken to avoid overburdening the respondent.

Questionnaires can contains three types of questions:

Dichotomous is a question when there are two possible responses. The choice between two alternatives reduces the issues to its simplest terms. Surveys often use dichotomous questions that ask for a Yes/No, True/False or Agree/Disagree response. There are a variety of ways to lay these questions out on a questionnaire. For example:

- Please enter your gender:

a) Male []

b) Female []

- Do you believe that the death penalty is ever justified:

a) Yes []

b) No []

Open-ended (or "open question") is a question where the researcher doesn't provide the respondent with a set answer from which to choose. Rather, the respondent is asked to answer "in their own words". This produces mainly **qualitative** data. Such questions are useful for obtaining in-depth information on the facts with which researcher is not very familiar, opinion, suggestions and attitude of informants, or sensitive issues. For example: What are your suggestions to improve our services?

Multiple Choice (categorised) is a question which provides several set alternative for its answers. Thus, it is a middle ground between free answers and dichotomous questions. Multiple choice questions can be used when an issue has more than two aspects, or when gradation, opinion or degree of intensity is involved. For example: In a study, a researcher want to know from respondent–What is your opinion on the following statement (tick any one):

'Nuclear families are better than joint families.'

1) Strongly Agree

2) Agree

3) No opinion

4) Disagree

5) Strongly Disagree

5.4.4 Case-Study Method:

The term case-study usually refers to a fairly intensive examination of a single unit such as a person, a small group of people, or a single company. Case-studies involve measuring what is there and how it got there. In this sense, it is historical. It can enable the researcher to explore, unravel and understand problems, issues and relationships. It cannot, however, allow the researcher to generalise, that is, to argue that from one case-study the results, findings or theory developed apply to other similar case-studies. The case looked at may be unique and, therefore not representative of other instances. It is, of course, possible to look at several case-studies to represent certain features of management that we are interested in studying. The case-study approach is often done to make practical improvements. Contributions to general knowledge are incidental.

The case-study method has four steps:

- Determine the present situation.
- Gather background information about the past and key variables.
- Test hypotheses. The background information collected will have been analysed for possible hypotheses. In this step, specific evidence about each hypothesis can be gathered. This step aims to eliminate possibilities which conflict with the evidence collected and to gain confidence for the important hypotheses. The culmination of this step might be the development of an experimental design to test out more rigorously the hypotheses developed, or it might be to take action to remedy the problem.
- Take remedial action. The aim is to check that the hypotheses tested actually work out in practice. Some action, correction or improvement is made and a re-check carried out on the situation to see what effect the change has brought about.

The case-study enables rich information to be gathered from which potentially useful hypotheses can be generated. It can be a time-consuming process. It is also inefficient in researching situations which are already well structured and where the important variables have been identified. They lack utility when attempting to reach rigorous conclusions or determining precise relationships between variables.

5.4.5 Projective Techniques:

The direct methods of data collection, viz., personal interview, telephone interview and mail surveyrely on respondents' own report of their behaviour, beliefs, attitudes, etc. But respondents may be unwilling to discuss controversial issues or to reveal intimate information about themselves or may be reluctant to express their true views fearing that they are generally disapproved. In order to overcome these limitations, indirect methods have been developed. Projective Techniques are such indirect methods. They become popular during 1950s as a part of motivation research. It is a psychological test in which a subject's responses to ambiguous or unstructured standard stimuli, such as a series of cartoons, abstract patterns, or incomplete sentences, are analyzed in order to determine underlying personality traits and feelings. This entails indirect question which enables the respondent to "project beliefs and feelings onto a third party". The respondents are expected to interpret the situation through their own experience, attitude and personality and express hidden opinion and emotions. The basic assumption of projective techniques is that a person projects his own thoughts, ideas and attributes when he perceives and responds to ambiguous or unstructured stimulus materials. Thus a person's unconscious operations of the mind are brought to a conscious level in a disguised and projected form, and the person projects his inner characteristics.

Of many techniques, word association, sentence completing and ink-blot tests are very common. In these techniques, both verbal and non-verbal (hesitation, time-lag and facial expression) are noted and interpreted. Such tests are useful for finding out consumer preference, buying attitude and behavior. Eventually,these are used for product development or finding out reason for failure of an apparently efficient product.

Types of Projective Techniques:

1) **Word Association**

In word association, respondents are presented with a list of words, one at a time and asked to respond to each with the first word that comes to mind. The words of interest, called test words, are interspersed throughout the list which also contains some neutral, or filler words to disguise the purpose of the study. Responses are analyzed by calculating:

(1) the frequency with which any word is given as a response;

(2) the amount of time that elapses before a response is given; and

(3) the number of respondents who do not respond at all to a test word within a reasonable period of time.

Example:

STIMULUS	MRS. X	MRS. Y
washday	everyday	ironing
fresh	and sweet	clean
pure	air	soiled
scrub	don't; husband does	clean
filth	this neighborhood	dirt
bubbles	bath	soap and water
family	squabbles	children
towels	dirty	wash

2) **Sentence Completion:**

In **Sentence completion**, respondents are given incomplete sentences and asked to complete them. Generally, they are asked to use the first word or phrase that comes to mind. A variation of sentence completion is paragraph completion, in which the respondent completes a paragraph beginning with the stimulus phrase.

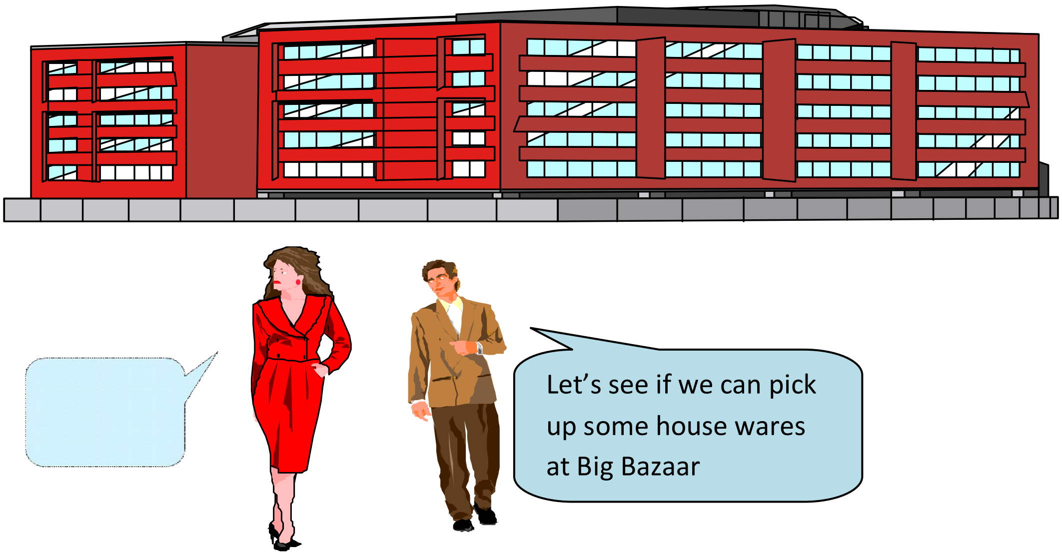
Example:

- A person who shops at Big Bazaar is ______________________
- Maggi is most liked because _________________________
- When I think of shopping in a department store, I ________

In **story completion**, respondents are given part of a story that is enough to direct attention to a particular topic but not to hint at the ending. They are required to give the conclusion in their own words.

With a **picture response,** the respondents are asked to describe a series of pictures of ordinary as well as unusual events. The respondent's interpretation of the pictures gives indications of that individual's personality.

In **cartoon tests**, cartoon characters are shown in a specific situation related to the problem. The respondents are asked to indicate what one cartoon character might say in response to the comments of another character. Cartoon tests are simpler to administer and analyze than picture response techniques.



Let’s see if we can pick up some house wares at Big Bazaar

3) **Expressive Techniques**

In expressive techniques, respondents are presented with a verbal or visual situation and asked to relate the feelings and attitudes of other people to the situation.

Role Playing - Respondents are asked to play the role or assume the behavior of someone else.

Third-Person Technique - The respondent is presented with a verbal or visual situation and the respondent is asked to relate the beliefs and attitudes of a third person rather than directly expressing personal beliefs and attitudes. This third person may be a friend, neighbor, colleague, or a “typical” person.

Given below are some of the advantages of projective techniques are:

- They may elicit responses that subjects would be unwilling or unable to give if they knew the purpose of the study.
- Helpful when the issues to be addressed are personal, sensitive, or subject to strong social norms.
- Helpful when underlying motivations, beliefs, and attitudes are operating at a subconscious level.

Given below are some of the disadvantages of projective techniques are:

- Suffer from many of the disadvantages of unstructured direct techniques, but to a greater extent.
- Require highly trained interviewers.
- Skilled interpreters are also required to analyze the responses.
- There is a serious risk of interpretation bias.
- They tend to be expensive.
- May require respondents to engage in unusual behavior.

Given below are some of the Guidelines for Using Projective Techniques:

- Projective techniques should be used because the required information cannot be accurately obtained by direct methods.

- Projective techniques should be used for exploratory research to gain initial insights and understanding.
- Given their complexity, projective techniques should not be used naively.

5.5 Secondary Data Collection Method

All methods of data collection can supply quantitative data (numbers, statistics or financial) or qualitative data (usually words or text). Quantitative data may often be presented in tabular or graphical form. Secondary data is data that has already been collected by someone else for a different purpose. For example, this could mean using:

- Data collected by a hotel on its customers through its guest history system
- Data supplied by a marketing organisation
- Annual company reports
- government statistics.

Secondary data are indispensable for most organizational research. Secondary data refer to information gathered by someone other than the researcher conducting the current study. Such data can be internal or external to the organization and accessed through the internet or perusal of recorded or published information. Secondary data can be used among other things for forecasting sales by constructing models based on past sales figures and through extrapolation. There are several sources of secondary data, including books and periodicals government publications of economic indicators, census data, statistical abstracts, data bases, the media, annual reports of companies, case studies and other archival records. Secondary sources of data provide a lot of information for research and problem solving. Such data are qualitative in nature. Also included in secondary sources are schedules maintained for or by key personnel in organizations, the desk calendar of executives and speeches delivered by them. Much of such internal data though could be proprietary and not accessible to all. Financial databases readily available for research are also secondary data sources. The Compustat Database contains information on thousands of companies organized by industry and information on global companies is also available through Compustat. The advantage of seeking secondary data sources is savings in time and cost of acquiring information.

Secondary data is data which has been collected by individuals or agencies for purposes other than those of our particular research study. For example, if a government department has conducted a survey of, say, family food expenditures, then a food manufacturer might use this data in the organisation's evaluations of the total potential market for a new product.

No marketing research study should be undertaken without a prior search of secondary sources (also termed desk research). There are several grounds for making such a bold statement. Secondary data may be available which is entirely appropriate and wholly adequate to draw conclusions and answer the question or solve the problem. Sometimes primary data collection simply is not necessary.

It is far cheaper to collect secondary data than to obtain primary data. For the same level of research budget a thorough examination of secondary sources can yield a more information than can be had through a primary data collection exercise.

The time involved in searching secondary sources is much less than that needed to complete primary data collection. Secondary sources of information can yield more accurate data than that obtained through primary research. This is not always true but where a government or international agency has undertaken a large scale survey, or even a census, this is likely to yield far more accurate results than custom designed and executed surveys when these are based on relatively small sample sizes.

It should not be forgotten that secondary data can play a substantial role in the exploratory phase of the research when the task at hand is to define the research problem and to generate hypotheses. The assembly and analysis of secondary data almost invariably improves the researcher's understanding of the marketing problem, the various lines of inquiry that could or should be followed and the alternative courses of action which might be pursued.

Secondary sources help define the population. Secondary data can be extremely useful both in defining the population and in structuring the sample to be taken. For instance, government statistics on a country's agriculture will help decide how to stratify a sample and, once sample estimates have been calculated, these can be used to project those estimates to the population.

5.5.1 Problems of Secondary Data:

Whilst the benefits of secondary sources are considerable, their shortcomings have to be acknowledged. There is a need to evaluate the quality of both the source of the data and the data itself. The main problems may be categorised as follows:

1. **Definitions:** The researcher has to be careful, when making use of secondary data, of the definitions used by those responsible for its preparation. Suppose a researcher is interested in rural communities and their average family size. If published statistics are consulted then a check must be done on how terms such as "family size" have been defined. They may refer only to the nucleus family or include the extended family. Even apparently simple terms such as 'farm size' need careful handling. Such figures may refer to any one of the following: the land an individual owns, the land an individual owns plus any additional land he/she rents, the land an individual owns minus any land he/she rents out, all of his land or only that part of it which he actually cultivates. It should be noted that definitions may change over time and where this is not recognised erroneous conclusions may be drawn. Geographical areas may have their boundaries redefined, units of measurement and grades may change and imported goods can be reclassified from time to time for purposes of levying customs and excise duties.

2. **Measurement Error:** When a researcher conducts fieldwork she/he is possibly able to estimate inaccuracies in measurement through the standard deviation and standard error, but these are sometimes not published in secondary sources. The only solution is to try to speak to the individuals involved in the collection of the data to obtain some guidance on the level of accuracy of the data. The problem is sometimes not so much 'error' but differences in levels of accuracy required by decision makers. When the research has to do with large investments in, say, food manufacturing, then management will set very tight margins of error in making market demand estimates. In other cases, having a high level of accuracy is not so critical. For instance, if a food manufacturer is merely assessing the prospects for one more flavour for a snack food already produced by the company then there is no need for highly accurate estimates in order to make the investment decision.

3. **Source Bias:** Researchers have to be aware of vested interests when they consult secondary sources. Those responsible for their compilation may have reasons for wishing to present a more optimistic or pessimistic set of results for their organisation. It is not unknown, for example, for officials responsible for estimating food shortages to exaggerate figures before sending aid requests to potential donors. Similarly, and with equal frequency, commercial organisations have been known to inflate estimates of their market shares.

4. **Reliability:** The reliability of published statistics may vary over time. It is not uncommon, for example, for the systems of collecting data to have changed over time but without any indication of

this to the reader of published statistics. Geographical or administrative boundaries may be changed by government, or the basis for stratifying a sample may have altered. Other aspects of research methodology that affect the reliability of secondary data is the sample size, response rate, questionnaire design and modes of analysis.

5. **Time scale:** Most censuses take place at 10 year intervals, so data from this and other published sources may be out-of-date at the time the researcher wants to make use of the statistics.

The time period during which secondary data was first compiled may have a substantial effect upon the nature of the data.

5.5.2 Sources of Secondary Data:

Secondary sources of information may be divided into two categories: internal sources and external sources.

(A) Internal Sources of Secondary Information

1) **Sales Data** : All organizations collect information in the course of their everyday operations. Orders are received and delivered, costs are recorded, sales personnel submit visit reports, invoices are sent out, returned goods are recorded and so on. Much of this information is of potential use in marketing research but a surprising amount of it is actually used. Organisations frequently overlook this valuable resource by not beginning their search of secondary sources with an internal audit of sales invoices, orders, inquiries about products not stocked, returns from customers and sales force customer calling sheets. For example, consider the information that can be obtained from sales orders and invoices:
   - Sales by territory
   - Sales by customer type
   - Prices and discounts
   - Average size of order by customer, customer type, geographical area
   - Average sales by sales person and
   - Sales by pack size and pack type, etc.

   This type of data is useful for identifying an organisation's most profitable product and customers. It can also serve to track trends within the enterprise's existing customer group.

2) **Financial Data:** An organization has a great deal of data within its files on the cost of producing, storing, transporting and marketing each of its products and product lines. Such data has many uses in marketing research including allowing measurement of the efficiency of marketing operations. It can also be used to estimate the costs attached to new products under consideration.

3) **Transport Data:** Companies that keep good records relating to their transport operations are well placed to establish which are the most profitable routes, and loads, as well as the most cost effective routing patterns. Good data on transport operations enables the enterprise to perform trade-off analysis and thereby establish whether it makes economic sense to own or hire vehicles, or the point at which a balance of the two gives the best financial outcome.

4) **Storage Data:** The rate of stockturn, stockhandling costs, assessing the efficiency of certain marketing operations and the efficiency of the marketing system as a whole. More sophisticated accounting systems assign costs to the cubic space occupied by individual products and the time period over which the product occupies the space. These systems can be further refined so that the profitability per unit, and rate of sale, are added. In this way, the direct product profitability can be calculated.

(B) External Sources of Secondary Information

There are large numbers of organisations that provide information including national and local government agencies, quasi-government agencies, trade associations, universities, research institutes, financial institutions, specialist suppliers of secondary marketing data and professional marketing research enterprises. Printed sources of secondary data begin with referral texts such as directories, indexes, handbooks and guides. These sorts of publications rarely provide the data in which the researcher is interested but serve in helping him/her locate potentially useful data sources.

The main sources of external secondary sources are:

Government Statistics - These may include all or some of the following:

- Population censuses
- Social surveys, family expenditure surveys
- Import/export statistics
- Production statistics
- Agricultural statistics

1. Trade Associations- Trade associations differ widely in the extent of their data collection and information dissemination activities. However, it is worth checking with them to determine what they do publish. At the very least one would normally expect that they would produce a trade directory and, perhaps, a yearbook

2. Commercial Services - Published market research reports and other publications are available from a wide range of organisations which charge for their information. Typically, marketing people are interested in media statistics and consumer information which has been obtained from large scale consumer or farmer panels. The commercial organisation funds the collection of the data, which is wide ranging in its content, and hopes to make its money from selling this data to interested parties.

3. National and International institutions- Bank economic reviews, university research reports, journals and articles are all useful sources to contact. International agencies such as World Bank, IMF, IFAD, UNDP, ITC, FAO and ILO produce a plethora of secondary data which can prove extremely useful to the marketing researcher.

5.6 Summary

The choice of a particular method of data collection depends upon the purpose of collecting information, the type of information being collected, the available resources and the skills required to collect the data. Mostly, data is collected through use of secondary source. If no such data is available from libraries or on the internet, one has to collect primary data for which a number of methods are available such as observations, interview, survey and indepth techniques. Construction of research instrument is the most important part of research endeavor as it determines the nature and quality of information. Research instruments must be developed in the light of objectives of the research.

5.7 Key Words

- **Data**- Data are facts, figures and other relevant materials, past and present, serving as bases for study and analysis.
- **Primary Data**- Data that has not been previously published, i.e. the data is derived from a new or original research study and collected at the source by researcher.

- **Secondary Data**- It refers to the statistical material which is not originated by the investigator himself but obtained from someone else's records.
- **Questionnaire-** A form containing a set of questions, especially one addressed to a statistically significant number of subjects as a way of gathering information for a survey.
- **Observation**- It is the systematic process of recording the behavioral patterns of people, objects and occurrences without questioning or communicating with them.

5.8 Self Assessment Test

1 What are the various methods of data collection? Explain.

2 What are primary data? List some of its advantages.

3 What are secondary data? Explain its advantages and disadvantages.

4 Explain observation as a source of primary data collection.

5 Explain some of the major projective techniques as a tool of data collection.

6 Write short notes on:
   a. Interview Method
   b. Case Study
   c. Survey

7 Differentiate between the primary and secondary data with some examples.

5.9 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.

Unit - 6 : Processing and Presentation of Data

Unit Structure:

6.0 Objectives
6.1 Introduction
6.2 Editing
6.3 Coding
6.4 Classification
6.5 Tabulation
6.6 Graphic and Diagrammatic Presentation
6.7 Summary
6.8 Key Words
6.9 Self Assessment Test
6.10 References

6.0 Objectives

After reading this unit, you should be able to:

- Convert raw data to useful information.
- Understand the difference between types of data i.e. categorical and numerical.
- Learn how to analyze each type of data.
- Eplore the categorical levels of data.
- Explore the numerical levels of data.
- Put data with frequency distribution.
- Classify data and present it in tabular form.
- Draw graphic presentation of data.
- Present of data through histogram, ogives and polygons.

6.1 Introduction

The data collected by primary and secondary sources is in raw form. It is called a data set and a single observation is a data point. To understand this data and to reach at logical and required conclusion, it is essential to organize the data set. Observation and statistics acquired from the collection can help the decision makers in drawing the suitable and accurate conclusions only when the acquired data is processed and presented aptly.

Information before it is arranged and analyzed is called raw data. It is raw because it is unprocessed by statistical methods.For example, the following data shows rainfall figures (in inches) for a certain location in specified months over a 10 year period.

Figure - 1 : **Raw Data**

18.6	13.8	11.6	17.7	25.2	17.6	19.2	19.4	18.4	14.8

Here is another example of concrete material. Engineers need to check whether concrete material to be used in building a bridge can withstand the stress. They test a sample of the concrete and on the basis of average conclude how much stress it can withstand. Figure-2 present the raw data gathered from a sample of 20 batches of concrete in constructing a bridge.

Figure -2: Data gathered of pounds of pressure/square inch that concrete can withstand

2500.2	2500.0	2502.8	2496.8
2500.8	2492.8	2499.9	2498.8
2498.4	2499.4	2500.4	2500.6
2499.2	2500.6	2500.6	2500.2
2500.9	2494.6	2498.8	2500.4

So it is to be understood that in case of research the data is drawn from a sample and it is assumed that it is representation of the whole population, but this large amount of collected data cannot support prompt and effective decision, unless it is processed and summarized in a useful way. The data set can be summarized and presented in the ascending or descending order. This simplest way of arranging data in the ascending or descending order is known as Data Array. This type of simple arrangement of data is in fact, a type of data processing.

Thus, data processing involves certain operations i.e. editing, coding, classification and tabulation of collected data so that it can be analyzed as per the need. According to C.Rajendra, "Some of the experts are of the view that the processing of data means concentrating and dealing with the data set". But it is an important and essential step of research because the acquired data is logically arranged and systematically organised at this stage.

6.2 Editing

Editing involves a careful scrutiny of the compiled data. It is an operation for assuring accuracy, consistency and uniformity of the data, so that the data set can be arranged to facilitate coding and tabulation. According to C.R. Kothari, "Editing of data is a process of examining the collected raw data(specially in surveys) to detect errors and omissions to correct these when possible".

Editing deals with errors and omissions in the data set and checking it for some inconsistencies in the response or responses given in the primary data. If any two answers are given by the respondent is not in consistency then while editing, the editor chooses the answer which he feels is correct or can suitably change the answer. Similarly in case of two answers given for a particular question, the editor can carefully decide and choose the suitable one or may code the 'no information' for the same. These corrections by the editor should be considered differently, and should be made by using different colour pencil so that they can be easily distinguishable.

Thus, it can be understood clearly that editing is done to fulfil following objectives –

a) **For Completeness** – Here the editor makes sure that all the questions are answered by a respondent. If any question of vital importance has not been answered then the respondent should be contacted again. If such a question is not related to some significant information, the editor may mark 'No Answer' there.

b) **For Consistency** – The editor checks that the answers should not be contradictory. For example, if a respondent in following questions-

   i) **Do you have a vehicle**? Answer – No

ii) **Which type of vehicle do you have?**

Four Wheeler ☐ **Two Wheeler** ☐

Answer – Two Wheeler.

These answers are inconsistent. The editor gets the correct answer and makes the required changes.

c) **For Accuracy** – To get reliable results from any data set, it is required that the information given by the respondents is accurate. The editor tries to make sure that the given information is accurate.

d) **For Homogeneity** – To get right information from the data and acquire sensible interpretations, the editor needs to check that the information supplied by various respondents is uniform. For example- in response to a question related to income, if one respondent has given monthly income and other has given yearly income, then the information can't yield sensible results.

Editing is of two types i.e. Field Editing and Central Editing. Field Editing, as its name suggests, is performed at the time of collection of data. At the time of recording the respondents' responses which the investigator translates or rewrites, if any responses are having errors, omissions or ambiguity due to respondents' writing style, they should be edited. In case of Central Editing after complete collection of data, the responses from the respondents are reviewed. It can be done after a week or a month of data collection. The central editing is undertaken after the questionnaires have been received at the headquarters and it is usually done by single editor. In case of Central editing following points should be taken care of –

- Possibly by single editor to maintain consistency.
- If more than one editor, uniform instructions should be given.
- Done prior to coding, classification and tabulation.
- In case of large data, entire data should be divided in two parts for consistency.
- Editor having specially knowledge of editing numeric data should be assigned this task.

The editors can correct the obvious error or strike out the answer. Editing tasks are performed by knowledgeable and well-informed editors. They must also be given instructions for editing along with general instructions given to the respondents. The editor's initials and the date of editing should be placed on each completed form or schedule.

6.3 Coding

Coding is the next major stage of qualitative data analysis. It is here that one carefully reads the transcribed data, line by line, and divides the data into meaningful analytical units (i.e., segmenting the data). When the meaningful segments are located, they are coded. Coding is defined as marking the segments of data with symbols, descriptive words, or category names, whenever meaningful segment of text is found in a transcript; a code or category name is assigned to signify that particular segment. This process is to be continued until all the data is segmented and coded.

During coding, one must keep a master list (i.e., a list of all the codes that are developed and used in the research study). It is required for classifying and tabulating the data set appropriately. These activities can't be performed effectively without accurate and apt Coding. Coding involves two steps i.e. identifying different categories and putting different responses in different categories.

According to C.R.Kothari "Coding refers to the process of assigning numerals or other symbols to the answers so that responses can be put into a limited number of categories or classes. It is necessary for efficient analysis and through it several replies may be reduced to a small number of classes which can contain the critical information required for analysis."

Coding is the technical procedure for converting verbal information into numbers or other symbols which can be more easily counted and tabulated. While identifying different categories, it should be taken care that all responses can fit in one of the categories. These categories should also not overlap each other, it means that one response should not be suitable for two different categories. These categories should be comprehensive as well as exclusive.

Coding depends on the information. When the responses of the respondents to the questions are observed, three categories of responses can be found i.e. Quantitative Responses, Categorical Responses (can be both quantitative & qualitative), and Descriptive Responses (always qualitative). Coding process is different for different types of responses.

6.3.1 Coding Quantitative and Categorical Responses

To do coding of quantitative and categorical responses, a code is developed. Here is an example of two questions from a questionnaire, to understand this coding process –

a) Question 1: Your current age ——————.

b) Question 2: Your marital status-

    i) Married   ii) Never Married   iii) Divorced/ Separated.

For coding responses of these two questions responses can be coded in the following manner-

Figure-3 Coding of Quantitative and Categorized Questions

Question No.	Variable	Response Pattern	Code
1	Age	21-25	1
		26-30	2
		31-35	3
		36-40	4
		41-45	5
		46-50	6
		50 and above	7
		No Response	9
2	MS	Married	1
		Never Married	2
		Divorced/Separated	3
		No Response	9

In the given example, it can be seen that that Question no. 1 is an open-ended quantitative question. For such questions, it is required that categories should be developed beforehand. Here in Question no. 2, categories are given within the question so codes can be easily assigned.

6.3.2 Coding Descriptive/ Qualitative Responses

When descriptive or qualitative data is to be coded, coding is done on the basis of 'Content Analysis.' Here is an example of an open-ended descriptive question-

Question: In your opinion, what are the main differences between businessmen at present and past times? (Note- Terms past and present are already specified)

For this question, responses given by the respondents could be several. To develop appropriate coding for this question, one needs to identify most probable themes of the responses. Like some common themes of responses could be – Lack of Satisfaction, Lack of Money, Low Morale, More Responsibilities, Greater Risk, More Restrictions & Laws, Higher Volatility, Better Opportunity and Greater Future Growth. After identifying these themes of the responses, one assigns codes to these most probable themes and categorizes them. Coding for this question can be as given in Figure -4.

Figure - 4 : Coding of an Open-ended Descriptive Question

Question No.	Variable	Response Pattern	Code
1	BSPNP	Lack of Satisfaction	
		• Present	1
		•Past	2
		Lack of Money	
		• Present	3
		• Past	4
		Low Morale	
		• Present	5
		• Past	6
		More Responsibilities	
		• Present	7
		• Past	8
		Greater Risk	
		• Present	9
		• Past	10
		More Restrictions & Laws	
		• Present	11
		• Past	12
		Higher Volatility	
		• Present	13
		• Past	14
		Better Opportunity	
		• Present	15
		• Past	16
		Greater Future Growth	
		• Present	17
		• Past	18
		No Response	19

After assigning codes to all the questions, it is required to pre-test the coding by taking certain questionnaires and coding the responses. After coding all the questions of those selected actual questionnaires, if any responses remain unencoded then coding should be done again to include those responses, if no responses remain without a code, the process of coding is considered complete.

6.4 Classification

Classification is a sorting operation where collected data set is put into different groups. It is an operation of dividing the large data into homogeneous groups to get meaningful relationships. It is the process of arranging the acquired data on the basis of common characteristics and placing it into different classes. It refers to the policies and procedures by which stored data is categorized so the information can be accessed, updated, protected, recovered and managed more efficiently in accordance with specific application requirements.

Classification makes the data easily understandable and thus, fulfil following objectives-

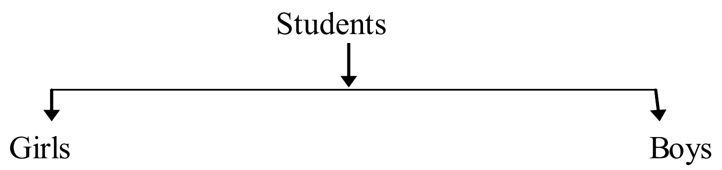
- To easily understand similarities and dissimilarities of data.
- To identify significant features of the data.
- To facilitate comparison.
- To sort significant information by dropping insignificant one.
- To facilitate statistical treatment
- Data can be classified on the basis of geographical area and time.

Classification becomes necessary when there is diversity in the data collected, as it facilitates meaningful presentation and analysis. A good classification should have the characteristics of clarity, homogeneity, & equality of scale, purposefulness, accuracy, stability, flexibility, and unambiguity. Classification is of two types, viz., **quantitative classification**, which is on the basis of variables of quantity; and **qualitative classification**, which is according to attributes.

6.4.1 Qualitative Classification -

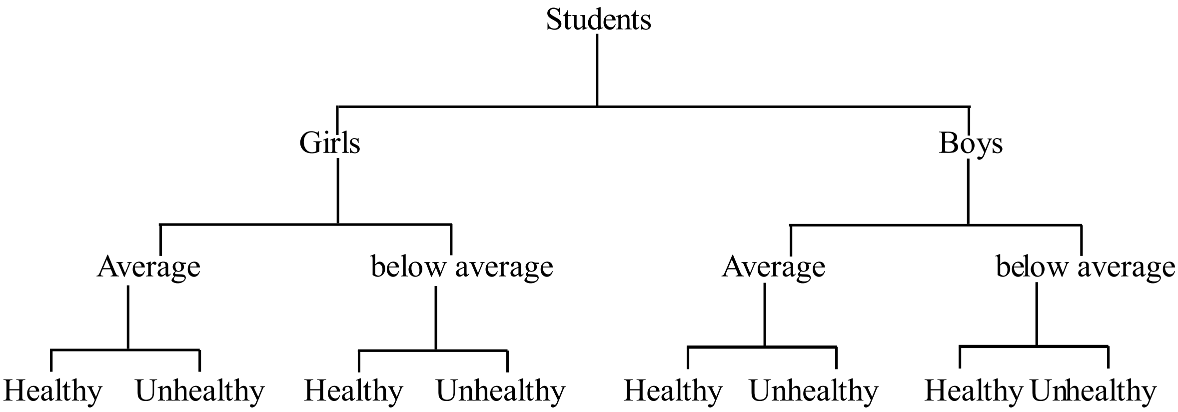
Classification when it is done on the basis of qualities again it is of two type's i.e. **two-fold and manifold classification**. When the universe is divided into two classes based on one attribute i.e. one class having that attribute and other class not having that attribute and only two classes are formed then this is known as simple or two-fold classification. For example in the data collected in a research related to the strength of girls and boys in an educational institution, the data can be classified in two classes.

Figure –5: Two-fold Classification



In manifold classification the universe is divided into more than two classes. Here classification is done by further dividing the data on the basis of other attributes. Thus when two or more attributes of the universe are to be considered and analyzed simultaneously, manifold classification is required. If in case of the given data of the students of an educational institution, other attributes of the data like 'average performance' and 'physical health' are also to be considered, then the classification will be known as manifold classification. Here total number of the classes can be found by the formula 2^n , where n denotes the number of attributes. In the given case n is equal to 3, so total number of classes will be 2^3, which is 8.

Figure-6: Manifold Classification



6.4.2 Quantitative Classification

Classification on the basis of some attribute which can be measured in numbers is called Quantitative classification. For example students of an educational institution can be classified on the basis of their age, weight, height etc. which are expressed in numbers. These measurable characteristics age, weight, height etc. are variables and these could be discrete or continuous. A discrete data variable is independent and doesn't have any conceivable fractional value. It is derived by counting whereas continuous data variable is derived through measurement. Quantitative data can be both discrete and continuous. When the data is classified on the basis of measurable attributes, it can also be classified according to class intervals.

While doing classification, it is significant to understand class interval. When the data set has numeric characteristics they are divided into groups which are called class intervals. There are certain things which we need to understand about class intervals –

- Each group has upper and lower class limits; it means values between these limits are included in that particular group. For example- in a class group of 20-30, the lower limit is 20 and the upper limit is 30. The value below 20 cannot be included in this group and the value above 30 cannot be included in this group.
- The difference between the lower and the upper limits of a class is known as class magnitude or interval of the class. In the class 20-30, class interval is 10. Generally multiple of 2, 5 or 10 are preferred for class interval but to identify appropriate class interval i, there is a simple formula i.e.

$$i = \frac{L - S}{k}$$

where, L is the largest item of the data.

S is the smallest item of the data.

k is the number of classes.

- The value lying half-way between the lower and the upper class limits of a class interval is known as Class Mid-point. It can be obtained by this formula –

$$\text{Class Mid point} = \frac{\text{Upper limit of the class} + \text{Lower limit of the class}}{2}$$

- Class limit can be of two types i.e. inclusive and exclusive type of class limits. In Inclusive type of class limit, the upper limit of one class is included in that class itself. For example in a class 20-29, the value 29 will be included in this class. It is an inclusive type of class limit. On the other side in a class 20-30 when the upper limit 30 becomes the lower limit of other class 30-40, the upper limit of earlier class and the lower limit of the later class are same. In this exclusive method of classification, it is always assumed that the upper limit of the class is excluded from the class. The inclusive method is generally used in case of the discrete variables whereas the exclusive method of classification is preferred in case of the continuous variables.

- Number of the classes (k), in which the data set is to be divided depends on the data set itself and on the nature of the problem. The purpose of dividing the data set is to make the data meaningful. It can be determined arbitrarily or by using following Sturges' formula –

$$k = 1 + 3.322 \log N$$

where N is the total number of observation.

- In case of discrete variables, the number of times a particular value is repeated is known as the frequency, whereas in case of continuous variables, the number of items in a class is called the frequency of that class.
- The frequency of a class is the number of times a particular value is repeated in a data set, or in each class of a data set.

Thus Classification is the way of arranging the data in different classes in order to give a definite form and a coherent structure to the data collected, facilitating their use in the most systematic and effective manner.

6.5 Tabulation

Table is the simplest device of summarizing data. Tabular method is used to condense the information from a mass of figures or raw data to a form, which provides an informative summary and readily conveys the essential features of the data. It is the most systematic way of presenting numerical data in an easily understandable form. It facilitates a clear and simple presentation of the data, a clear expression of the implication, and an easier and more convenient comparison. Here is san example of Ice-cream sales of Annand Dairy in each of the 40 Ice-cream parlours.

Figure - 7: Table of Ice-cream Sale in Number of Packs of Annand Dairy in Each of the 40 Ice-cream Parlours

44	49	56	65	64	43	44	55	65	49
66	45	58	34	66	43	54	48	45	55
55	65	49	54	48	45	55	43	49	56
48	45	43	49	44	49	56	35	34	66

6.5.1 Frequency Distribution Table

A simple way to sort this data is to draw up a tabular summary in the form of frequency distribution. It is a slightly more convenient method of tabulating a collection of data i.e. by arranging it in rank order, so making it easier to see how many times each number appears. This is known as "ranked data". The next table shows the previous sales figures in this form. It is possible to save a little space by making a table in which each individual item of the ranked data is written down once only, but paired with the number of times it occurs. The data is then presented as a "frequency distribution table". There may also be a column of "tallied frequencies" if the table is being constructed from the raw data.

Figure - 8: Frequency Distribution of Ice-cream of Anand Dairy Sale in Number of Packs in Each of the 40 Ice-cream Parlours

Ice-cream Sale in number of packs	Tally Mark	Ice-cream Parlours
34	II	2
35	I	1
43	IIII	4
45	III	3
44	IIII	4
48	III	3
49	~~IIII~~ I	6
54	II	2
55	IIII	4
56	III	3
58	I	1
64	I	1
65	III	3
66	III	3
	40	40

6.5.2 Grouped Frequency Distribution Table

For about forty or more items in a set of numerical data, it usually most convenient to group them together into 10 and 15 "classes" of values, each covering a specified range or "class interval". Each item is counted every time it appears in order to obtain the "class frequency" and each class interval has the same "class width". Too few classes mean that the data is over-summarised while too many classes means that there is little advantage in summarising at all. Here, we use the convention that the lower boundary of the class is included while the upper boundary is excluded. A "grouped frequency distribution table" normally has columns which show the class intervals, class frequencies, and cumulative frequencies, the last of these being a running total of the frequencies themselves. The cumulative frequency shows, at a glance, how many items in the data are less than a specified value. It is sometimes more useful to use the ratio of the cumulative frequency to the total number of observations (relative cumulative frequency).

Figure - 9: Cumulative & Relative Frequency Distribution Table of Ice-cream Sale

Class Intervals	Frequency	Cumulative Frequency	Percentage/Relative Frequency
30-40	3	3	7.5%
40-50	20	23	50%
50-60	10	33	2.5%
60-70	7	40	17.5%
Total	40	40	100

6.6 Graphic and Diagrammatic Presentation

Diagrams and Graphs both are powerful and effective ways for presenting, understanding and comparing statistical data. They need to be constructed and presented carefully. They can present only approximate values and limited amount of information. They don't help in analysing the data. They can be easily manipulated because they cannot be accurately appraised visually.

After summarizing the collected data by classification and tabulation, it is presented in a systematic manner. To make an interesting and effective presentation of the summarized data diagrams and graphs are used. Diagrammatic and Graphic presentations make the data not only convincing but also appealing. In case of Graphic presentation, Data recorded in experiments or surveys is displayed by a statistical graph. There are different types of statistical graphs. Selection of the graph is determined by the type and breadth of the data, the audience it is directed to, and the questions being asked.

A graph refers to the plotting of different values of the variables on a graph paper, which gives the movement or a change in the variable over a period of time. Diagrams can present the data in an attractive style but still there is a method more reliable than this. Diagrams are often used for publicity purposes but are not of much use in statistical analysis. Hence graphic presentation is more effective and result oriented.

6.6.1 Advantages of Graphs

The presentation of statistics in the form of graphs facilitates many processes in economics. The main uses of graphs are as under:

- **Attractive and Effective presentation of Data**: The statistics can be presented in attractive and effective way by graphs. It is said that a picture is worth of a thousand words, therefore a fact that an ordinary man cannot understand easily, could understand that in a better way by graphs.
- **Simple and Understandable Presentation of Data:** Graphs help to present complex data in a simple and understandable way. Therefore, graphs help to remove the complex nature of statistics.
- **Useful in Comparison:** Graphs also help to compare the statistics. If the profits made by two different competitors are presented through graphs, then it becomes easy to understand the difference between the two.
- **Useful for Interpretation:** Graphs also help to interpret the conclusion. It saves time as well as labour.
- **Remembrance for long period:** Graphs help to remember the facts for a long time and they cannot be forgotten.
- **Helpful in Predictions:** Through graphs, tendencies that could occur in near future can be predicted in a better way.
- **Universal utility:** In modern era, graphs can be used in all spheres such as trade, economics, government departments, advertisement, etc.
- **Helpful in Transmission of Information:** Graphs help in the process of transmission as well as information of facts.
- **No Need of Training :**There isn't any need for special training for the interpretation, when facts are presented through graphs.

6.6.2 Construction of Graphs

There are different types of graphs and they are useful for presenting different types of data. Following are the main rules to construct a graph:

- Every graph must have a suitable title which should clearly convey the main idea, the graph intends to portray.
- The graph must suit to the size of the paper.
- The scale of the graph should be in even numbers or in multiples.
- Footnotes should be given at the bottom to illustrate the main points about the graph.
- Graph should be as simple as possible.
- In order to show many items in a graph, index for identification should be given.
- Every graph should be given with a table to ensure whether the data has been presented accurately or not.
- The test of a good graph depends on the case with which the observer can interpret it. Thus economy in cost and energy should be exercised in drawing the graph.

6.6.3 Limitation of Graphs

Although Graphs are very useful yet there are various drawbacks about them. Following are the main limitations of graphs:

- **Limited Application:** Graphic representation is useful for a common man but for an expert, its utility is limited.
- **Lack of Accuracy:** Graphs do not measure the magnitude of the data. They only depict the fluctuations in them.
- **Subjective:** Graphs are subjective in character. Their interpretation varies from person to person.
- **Misleading Conclusions:** The person who has no knowledge can draw misleading conclusions from graphs.
- **Index:** In order to show many items in a graph, index for identification should be given.

The scale indicates the unit of a variable that a fixed length of axis would represent. Scale may be different for both the axes. It should be taken in such a way so as to accommodate whole of the data on a given graph paper in a lucid and attractive style. Sometimes data to be presented does not have low values but with large terms. One has to use the graph so as it may present the given data for comparison even.

6.6.4 Types of Graphs

There are different types of graphs. They have their uses with several advantages and disadvantages. Some of the common graphical methods of data presentation are as follows –

- Line Graphs.
- Bar Charts.
- Subdivided Bar Charts
- Histograms &Frequency Polygons.

- Cumulative Frequency Curves or Ogives
- Pie Charts
- Pictograph

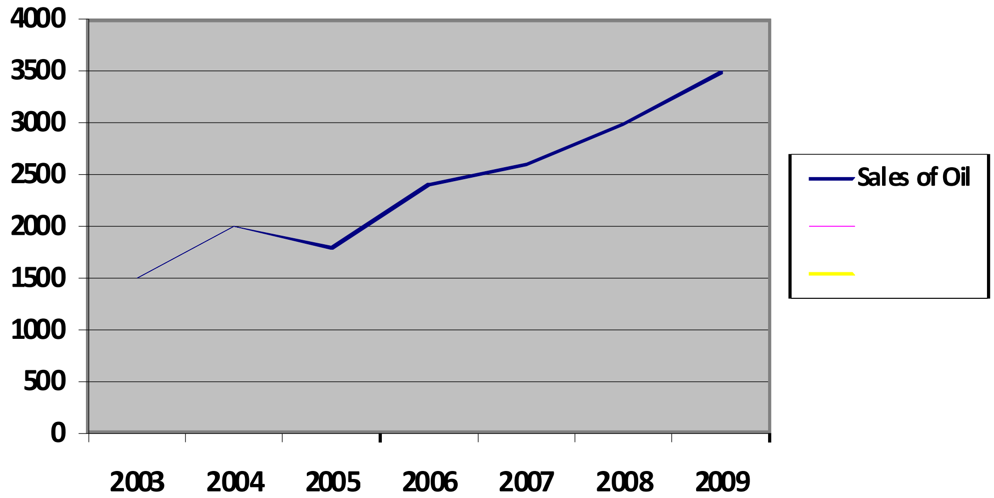
6.6.5 Line Graphs

Line Graph is a visual presentation of a set of data values, joined by straight lines. A line graph plots continuous data as points and then joins them with a line. Multiple data sets can be graphed together, but a key must be used. It can compare multiple continuous data sets easily and interim data can be inferred from graph line. It can be effectively used to describe committed relationships between two variables by plotting their respective values on the x- and y-axes. It has limitation as it can be used only with continuous data. Let us take an example of Sales of edible oil over a period of time. If sales of the oil are as shown in Figure 10, the line graph for the same shown in Figure 11.

Figure 10: Sales of Edible oil Over a Period of Time

Year	2003	2004	2005	2006	2007	2008	2009
Sales of Edible Oil (in 1000 litres)	1500	2000	1800	2400	2600	3000	3500

Figure 11: Line Graph of Sales of Edible oil Over a Period of time



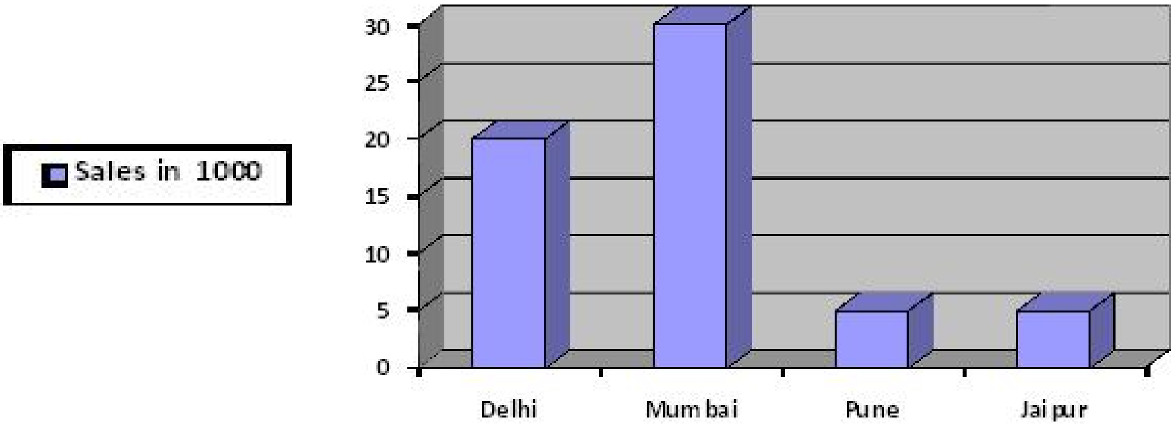
6.6.6 Bar Chart

Bar Chart comprises a series of bars of equal width, where the base of the bars being equal to the class interval of a data. The bars stand on a common base line and the heights of the bars being proportional to the frequency of the interval. The bars can be drawn vertically or horizontally. The data set given in the table presents the distribution of sales of a cosmetic brand in a particular month in five different cities.

Figure 12: Distribution of Sales of a Cosmetic Brand over a period of time

Cities	No. of Items Sold (in 1000)
Delhi	20
Mumbai	30
Pune	5
Jaipur	5

Figure - 13: Bar Chart of the Distribution of Sales of a Cosmetic Brand



6.6.7 Sub-divided Bar Chart

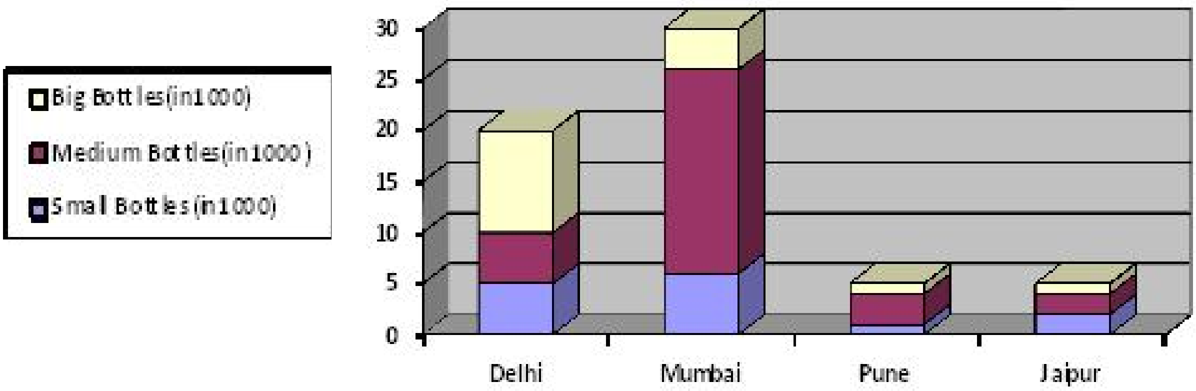
It is a bar chart wherein each bar is divided into further components. It is a chart which shows further division of any data, if required. If in the above mentioned data of sales given in Figure 10, the sales include three different packing sizes i.e. small, medium and big bottles. The data of their sales is given in the Figure 14.

Figure - 14: Sales as per Different Packing Sizes

Cities	Small Bottles Sold (in 1000)	Medium Bottles Sold (in 1000)	Big Bottles Sold (in 1000)	No. of Items Sold (in 1000)
Delhi	5	5	10	20
Mumbai	6	20	4	30
Pune	1	3	1	5
Jaipur	2	2	1	5

The Sub-divided Bar Chart of the above given data will be presented as given under:

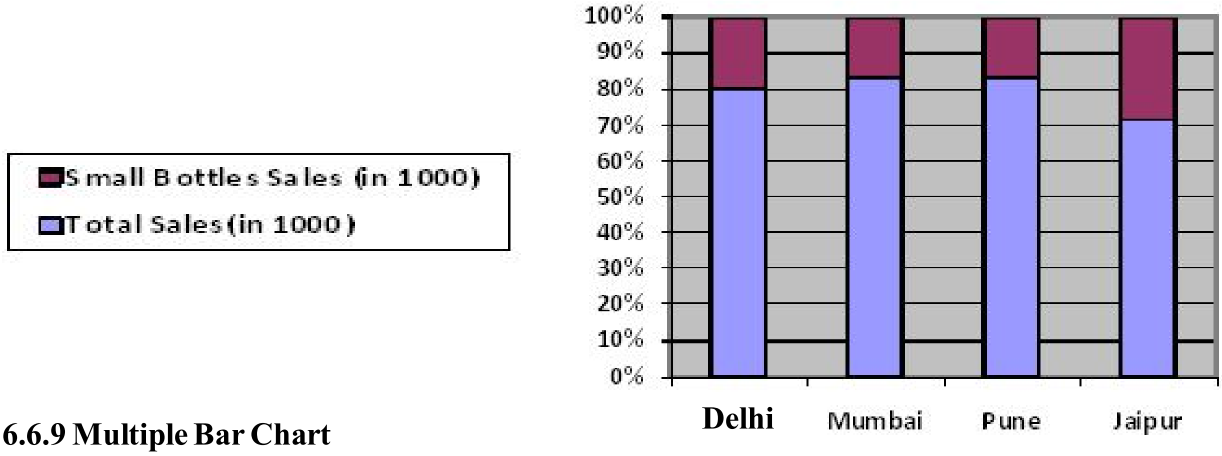
Figure - 15: Sub-divided Bar Chart of Sales as per Different Packing Sizes



6.6.8 Percentage Bar Chart

Percentage Bar Chart is one in which each bar is divided into components which are expressed as percentage of the total bar. In the above example if the company studies sales of its small bottles in comparison to other types of bottles, the percentage bar may show it. In Figure 14 given percentage bar chart shows that in Jaipur and Pune small bottles are preferred.

Figure 16: Percentage Bar Chart of Sales of Small Packing Sizes



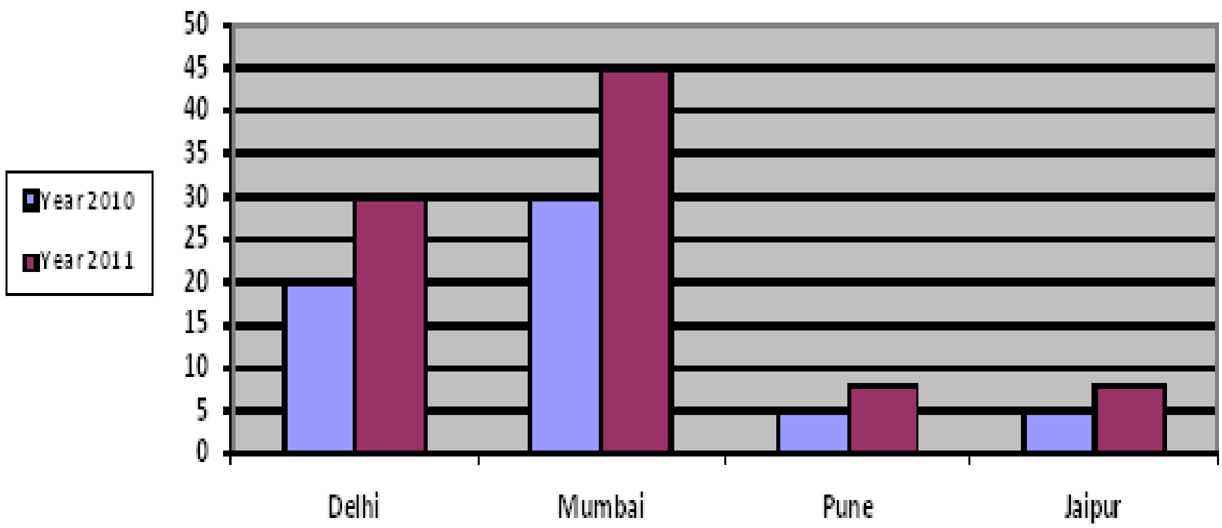
6.6.9 Multiple Bar Chart

Multiple Bar Chart is a bar chart in which one or more bars are placed together for each entity. In this type of chart, two or more bars are placed together to provide comparative assessment of values of same parameters over two different periods of time. If we take the same example of sales of a cosmetic brand over two different years, the data set becomes as under-

Figure 17: Distribution of Sales of a Cosmetic Brand in Years 2010 & 2011

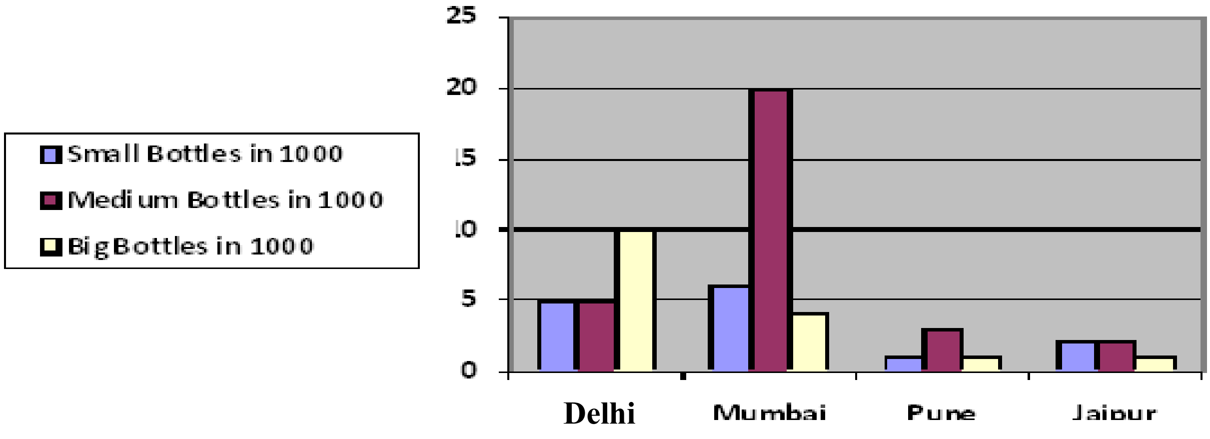
Cities	of Items Sold (in 1000) 2010	No. of Items Sold (in 1000)2011
Delhi	20	30
Mumbai	30	45
Pune	5	8
Jaipur	5	8

Figure 18: Multiple Bar Chart of Distribution of Sales in Years 2010 & 2011



To understand it better we can again take the example of sales and size as well, the given data of different size of packing in different cities (Figure14) can also be compared by presenting multiple bar chart as shown in Figure19.

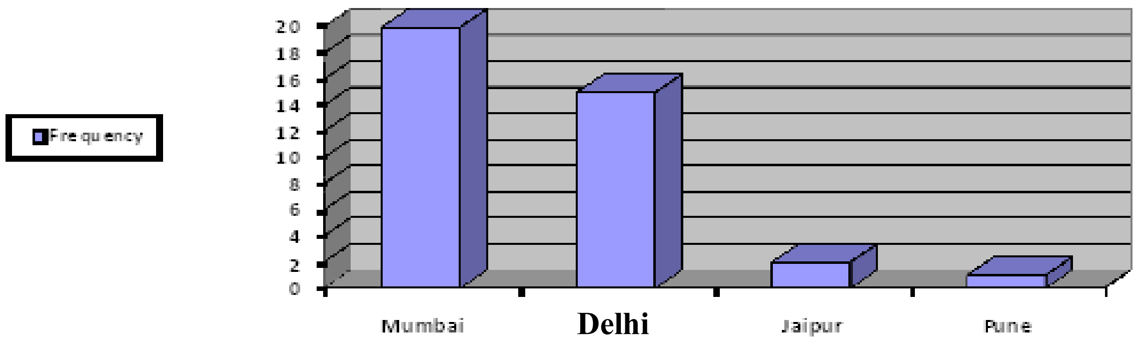
Figure 19: Multiple Bar Chart of Sales as per Different Packing Sizes



6.6.10 Pareto Chart

Pareto Chart is specialized bar chart, which is used to classify a variable into groups or intervals from the largest to smallest frequency. It is named after famous Italian economist. It helps in identifying the most frequent occurrence of an event or phenomenon. It can be used to sort the data according to any of the criteria like geographical regions, cities etc. For the given example of the data set given in the Figure 12 presents the distribution of sales of a cosmetic brand in a particular month in five different cities the chart of the same data will be presented as given in Figure 20.

Figure 20: Pareto Chart of Distribution of Sales of a Cosmetic Brand



6.6.11 Histogram and Frequency Polygon

Histogram has vertical rectangles whose bases are proportional to the class interval and height are is proportional to frequencies. A "histogram" is a diagram which is directly related to a grouped frequency distribution table and consists of a collection of rectangles whose height represents the class frequency (to some suitable scale) and whose breadth represents the class width. The histogram shows, at a glance, not just the class intervals with the highest and lowest frequencies, but also how the frequencies are distributed.

Frequency Polygon is formed by joining the middle points of these rectangles. Using the fact that each class interval may be represented, on average, by its class mid-point, one may plot the class mid-points against the class frequencies to obtain a display of single points. By joining up these points with straight line segments and including two extra class midpoints, it gives a "Frequency Polygon". A histogram displays continuous data in ordered columns. Categories are of continuous measure such as time, inches, temperature, etc. A histogram is visually strong and it can be compared to normal curve. But it has many disadvantages like it cannot read exact values because data is grouped into categories and it makes comparing two data sets more difficult. It can be used only with continuous data.

Figure 21: Frequency Distribution Table of Ice-cream Sale

Class Intervals	Class Mid-point	Frequency
30-40	35	3
40-50	45	20
50-60	55	10
60-70	65	7
Total	**Class Mid-point**	**40**

Figure 22: Histogram of Frequency Distribution Table of Ice-cream Sale

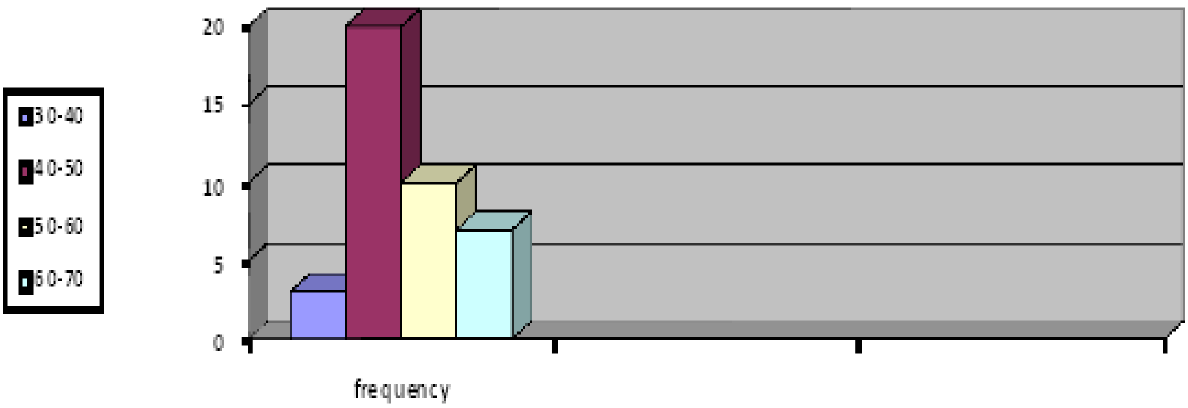
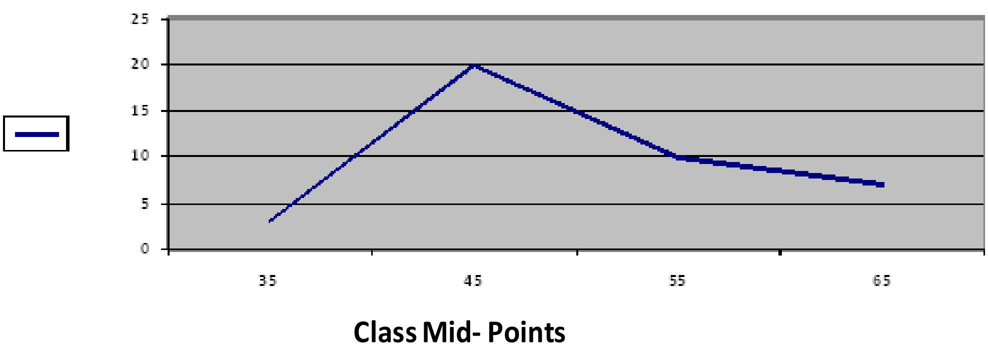


Figure 23: Frequency Polygon of Ice-cream Sale



6.6.12 Ogive or Cumulative Frequency Curve

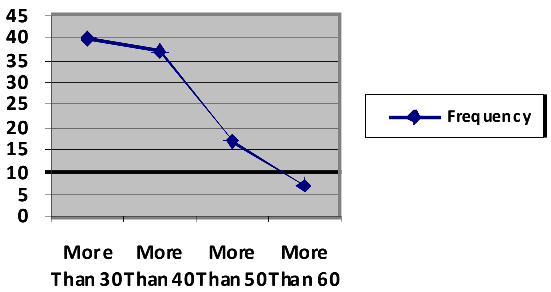
The Ogive or Cumulative Frequency Curve is a graph which gives an idea about the number of observations less than or greater than the values in the range of the variable. There are two types of Ogives i.e. 'Less Than' and 'Greater Than'. The earlier use of the cumulative frequency to estimate the number (or proportion) of values less than a certain amount may be applied graphically by plotting the upper class-boundary against cumulative frequency; then joining up the points plotted with straight line segments. The graph obtained is called the "cumulative frequency polygon" or "ogive".

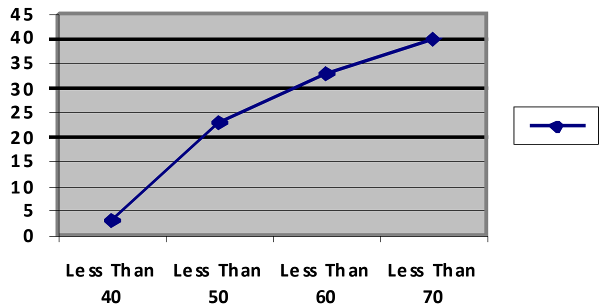
For Example -

Figure 24: Frequency Distribution Table of Ice-cream Sale

Class Intervals	Frequency	For 'Less Than' Curve		For 'More Than' Curve	
		Class Intervals	Cumulative Frequency	Class Intervals	Cumulative Frequency
30-40	3	<40	3	>30	40
40-50	20	<50	23	>40	37
50-60	10	<60	33	>50	17
60-70	7	<70	40	>60	7

Figure 25: 'Less Than' Ogive & 'More Than' Ogive of Ice-cream Sales





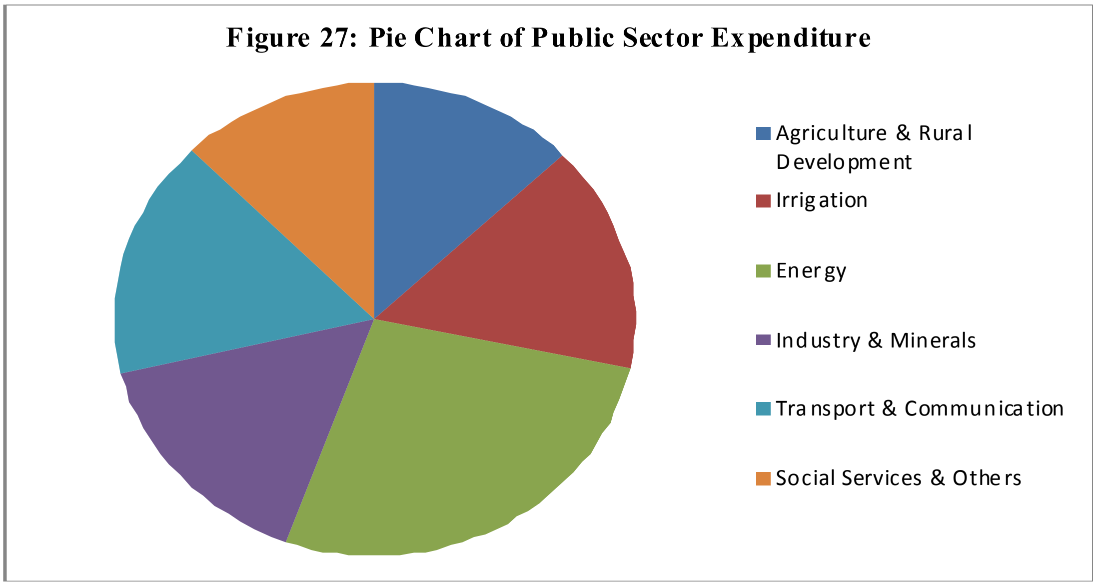
6.6.13 Pie diagrams

Pie diagrams are very popularly used to present data according to percentage breakdowns. If the data of VI five year plans, public sector expenditure is given in Figure 26.

Figure 26: Public Sector Expenditure Data Computation for Pie Chart

Sector	Percentage	Angle Outlays
Agriculture and Rural Development	12.9	12.9 x 360/100 = 46°
Irrigation	15.4	15.4 x 360/100 = 56°
Energy	27.2	27.2 x 360/100 = 98°
Industry and Minerals	15.9	15.9 x 360/100 = 57°
Transport and Communication	16.1	16.1 x 360/100 = 58°
Social Services and Others	12.5	12.5 x 360/100 = 45°

The Pie Diagram of the given data will be as follows-



Pie Chart is useful to show relative proportions. It is less effective than bar charts for accurate interpretations and readings and when readings are large in number. It is also not suitable for comparison.

6.6.14 Pictographs

Pictographs are not abstract presentations such as lines are bars. They elepict a data using attractive pictures and make the data easy to understand. By pictorial symbols even a bay man can understand it. for constructing a pictograph certain points should be taken care of ie symbol should be relected carefully and must represent the iden. It should be clear interesting. Symbol should be distinguishable and suitable in size. It has greater attraction value and stimulate higher interest, but it does not give details and appropriate symbol selection is not easy.

Example

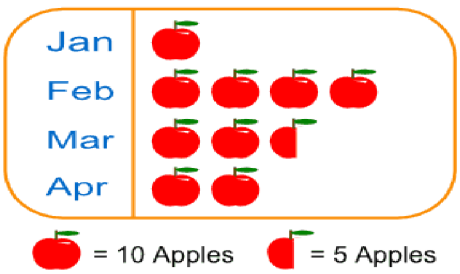
Figure 28: Sale of Apples at a Shop in Few Months.

Months	Sale of Apples in Number of Apples
January	10
February	40
March	25
April	20

Note that each picture of an apple means **10 apples** (and the half-apple picture means 5 apples).

The pictograph is shown in Figure 29.

Figure 29 : Sales of Apples



Graphs and diagrams both are not very different. Yet are should understand the difference between these two. Diagram are generally drawn on paper but graphs are plotted on the graph paper. Diagram are more attractive to look at and better suitable for publicity but graphs are accurately measured and better saitable for staticiation and their statistical analyris. Graphs are more suitable aralyris. Graphs are more suitable for frequency distribution and time series etc. For proper contraction of there two a proper proportion between width and height with suitable selection of scale should be these.

6.7 Summary

In this chapter, 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. 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. A graphic presentation is constructed in relation to two axes : horizontal and vertical. It also has a title and labeled axes, to describe content. There are different type of graphs or charts the common one's are Line Graphs Bar Charts, Subdivided Bar Charts, Percentage Bar Charts, Multiple Bar Charts, Histograms, Frequency Polygons, Ogives, Pie Chats, Scattergrans etc.The data is sometime presented by using pictures or figures which makes if more interesting, then it is called pictogram.

6.8 Key Words

- **Frequency Distribution :** It is a tabular summary of a set of data showing the frequency (or number) of items in each of several non-overlapping classes.
- **Classes -** These are the groups into which a set of data may be classified.
- **Class Intervals-** These are the ranges of values which define the width of each class (group)
- **Class Freqencies -** These are the number of observations contained in each class.
- **Class Limits -** Class limits are the upper and lower limits for the classes of a freqency distribution.
- **Line Graph-** Graphical presentation of data often used for showing movements in a variable over time.
- **Bar Chart -** It is the chart which is used when discrete data are grouped into classes with each class corresponding to a rectangle, the baseof which is the class interval and the height of which is the class frequency.
- **Histogram -** It is similar to a bar chart and may be used in the cases of continuous or discrete grouped data but with no gaps left between the bars.
- **Freqency Polygons -** This is a graph formed by plotting class frequencies against the mid points of the corresponding classes, connecting the points to form a graph and extending it to each end of the distribution to meet the horizontal axis at the mid point at what would have been the next class below or above, respectively.

- **Cumulative Frequency Curve (Ogive) -** It is a graph which presents the cumulative frequency distribution.
- **Pie Chart-** It is a pictorial device for presenting categorical data in which a circle is divited into wedges like slices of a pie each wedge corresponding to the relative frequency of each class.
- **Pictogram -** It is a visual representation of categorical data in which pictures are used to summarize data each picture being scaled in size or repeated a number of times to indicate relative magnitudes.

6.9 Self Assessment Test

1 The Data shows the percentage marks of 40 students in the form of a table-

33	41	17	83	63	54	92	60	06	24	00	09	21	29	43	46	60	70	71	72
81	27	86	25	64	32	68	38	92	11	88	87	51	88	56	89	59	82	44	88

(a) Arrange these data as a frequency distribution of 10 classes of equal width.

(b) Present the frequency distribution as

i) Histogram.
ii) Frequency Polygon.

2 What is a Pie Chart?

3 What do you understand by the processing of the data ? What are the different steps of it?

4 What is 'raw data' ? Why do you edit the data ?

5 What are different types of Editing ? Explain them

6 What is Graphical Presentation ? Give any two methods of graphical presentation with example.

7 Explain Coding with example.

8 What do you understand by Classification ? Define Classes, Class intervals, Class Limits, Mid Points & Class Frequencies.

9 The given data in Figure below shows the the monthly expenditure in rupees of two families A and B-

Figure : Expenditure in Rupees of Two Families A and B.

Items	Family A	Family B
Food	3800	3000
Clothing	2000	1000
Fuel & Power	1000	500
Miscellaneous	1200	1500

Present the data using

(i) Pie Chart
(ii) Bar Charts
(iii) Percentage Bar chart

10 What is the difference between Graphical Presentation& Diagrammatic Presentation ? Give Pictogram with an example.

6.10 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.
- 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.

Unit - 7 : Measure of Central Tendency

Unit Structure:

7.0 Objectives
7.1 Introduction
7.2 Mean
7.3 Median
7.4 Mode
7.5 Comparing Measures of Central Tendency
7.6 Choosing a Measure of Central Tendency
7.7 When to Use the Mean, Median and Mode
7.8 Empirical Relation between Mean, Median And Mode
7.9 Weighted Mean
7.10 Summary
7.11 Key Words
7.12 Self Assessment Test
7.13 References

7.0 Objectives

After going through this unit, you should be able to:

- Explain meaning of measures of central tendency
- Calculate Mean, Median and Mode
- Comparing measures of central tendency
- Choosing a Measure of Central Tendency
- When to Use the Mean, Median, and Mode
- Empirical Relation Between Mean, Median And Mode
- Calculate Weighted Mean

7.1 Introduction

Whenever you collect data, you end up with a group of scores on one or more variables. If you take the scores on one variable and arrange them in order from lowest to highest, what you get is a distribution of scores. Researchers often want to know about the characteristics of these distributions of scores, such as the shape of the distribution, how spread out the scores are, what the most common score is, and so on. One set of distribution characteristics that researchers are usually interested in is central tendency. This set consists of the mean, median, and mode.

The mean is probably the most commonly used statistic in all research. The mean is simply the arithmetic average of a distribution of scores, and researchers like it because it provides a single, simple number that gives a rough summary of the distribution. It is important to remember that although the mean provides a useful piece of information, it does not tell you anything about how spread out the scores are (i.e., variance) or how many scores in the distribution are close to the mean. It is possible for a distribution to have very few scores at or near the mean.

The median is the score in the distribution that marks the 50th percentile. That is, 50% of the scores in the distribution fall above the median and 50% fall below it. Researchers often use the median when they want to divide their distribution scores into two equal groups (called a median split). The median is also a useful statistic to examine when the scores in a distribution are skewed or when there are a few extreme scores at the high end or the low end of the distribution.

The mode is the least used of the measures of central tendency because it provides the least amount of information. The mode simply indicates which score in the distribution occurs most often, or has the highest frequency.

7.2 Mean

The symbol for the mean of a sample is $\bar{X}$ (pronounced "mean" or "Xbar"). The symbol for the mean of a population is μ (a Greek letter, pronounced "mew"). Of course, if you had an entire population of scores you could calculate μ, and it would carry no uncertainty with it. Most of the time, however, the population is not available and you must make do with a sample. Fortunately, mathematical statisticians have shown that the formula for produces a value that is the best estimator of μ. It also turns out that this formula for is the same as the formula for μ. The difference in and μ then is in the interpretation. carries some uncertainty with it; μ does not.

Table 1: Formula for Calculating the Mean of a Distribution

$$\mu = \frac{\Sigma X}{N}$$

$$\bar{X} = \frac{\Sigma X}{n}$$

$\bar{X}$ is the sample mean
μ is the population mean
Σ means "the sum of"
X is an individual score in the distribution
n is the number of scores in the sample
N is the number of scores in the population

Here's a very simple example of a sample mean. Suppose a college freshman arrives at school in the fall with a promise of a monthly allowance for spending money. Sure enough, on the first of each month, there is money to spend. However, 3 months into the school term, our student discovers a recurring problem: too much month left at the end of the money.

In pondering this problem, our student realizes that money escapes from his pocket at the Student Center. So, for a 2-week period, he keeps a careful accounting of every cent spent at the center (soft drinks, snacks, video games, coffee, and so forth). His data are presented in Table 2. You already know how to compute the mean of these numbers, but before you do, eyeball the data and then write down your estimate of the mean in the space provided. The formula for the mean is given in Table 1.

Table 2: Amounts of money spent at the Student Center during a 2-week period

Day	Money spent ($)
1	3.25
2	2.50
3	4.47
4	0.00
5	3.81
6	1.75
7	0.00
8	0.00
9	6.78
10	2.40
11	0.00
12	0.00
13	8.50
14	4.20

Σ=$37.66

Your estimate of the mean _______

For the data in Table 2,

$$\bar{X} = \frac{\Sigma X}{n} = \frac{37.66}{14} = \$2.69$$

These data are for a 2-week period, but our freshman is interested in his expenditures for at least 1 month and, more likely, for many months. Thus, the result is a sample mean and the symbol is appropriate. The amount, $2.69, is an estimate of the amount that our friend spends at the Student Center each day. $2.69 may seem low to you. If so, note that $0.00 was spent several days.

Now we come to an important part of any statistical analysis, which is to answer the question, So what? Calculating numbers or drawing graphs is a part of almost every statistical problem, but unless you can tell the story of what the numbers and pictures mean, you won't find statistics worthwhile.

The first use you can make of Table 2 is to estimate the student's monthly Student Center expenses. This is easy to do. Thirty days times $2.69 is $80.70. Now, let's suppose our student decides that this $80.70 is an important part of the "monthly money problem." The student has three apparent options. The first is to get more money. The second is to spend less at the Student Center. The third is to justify leaving things as they are. For this third option, our student might perform an economic analysis to determine what he gets in return for his $80+ a month. His list might be pretty impressive: lots of visits with friends, information about classes, courses, and professors, a borrowed book that was just super, thousands of calories, and more.

The point of all this is that part of the attack on the student's money problem involved calculating a mean. However, an answer of $2.69 doesn't have much meaning by itself. Interpretations and comparisons are called for.

7.2.1 The Mean for a Simple Frequency Distribution

The steps for finding the mean from a simple frequency distribution follow, but first (looking only at the data and not at the summary statistics at the bottom) estimate the mean of the SWLS scores in the space at the bottom of Table 3.

The first step in calculating the mean from a simple frequency distribution is to multiply each score in the X column by its corresponding f value, so that all the people who make a particular score are included. Next, sum the fX values and divide the total by N. (N is the sum of the f values.) The result is the mean. In terms of a formula,

μ or $\overline{X} = \frac{\Sigma fX}{n}$

Table 3: Calculating the mean of the simple frequency distribution of the Satisfaction With Life Scale scores

SWLS score (X)	f	fX	SWLS score (X)	f	fX
35	2	70	22	5	110
34	1	34	21	3	63
33	2	66	20	5	100
32	4	128	19	3	57
31	2	62	17	2	34
30	5	150	16	2	32
29	7	203	15	1	15
28	6	168	13	2	26
27	10	270	12	1	12
26	9	234	11	1	11
25	9	225	10	2	20
24	6	144	9	2	18
23	6	138	5	2	10
				Σ = 100	2400

Your estimate of the Mean _______

For the data in Table 3,

μ or $\overline{X} = \frac{\Sigma fX}{n} = \frac{2400}{100} = 24.00$

How did 24.00 compare to your estimate?

To answer the question of whether 24.00 is or μ, you need more information. Here's the question to ask: Is there any interest in a group larger than these 100? If the answer is no, the 100 scores are a population and 24.00 = μ. If the answer is yes, the 100 scores are a sample and 24.00 = .

7.2.2 Characteristics of the Mean

Two characteristics of the mean are important for you to know. Both characteristics will come up again later. First, if the mean of a distribution is subtracted from each score in that distribution and the differences are added, the sum will be zero; that is, Σ (X -) = 0. The statistic, X - , is called a deviation score. To demonstrate to yourself that Σ (X -) = 0, you might pick a few numbers to play with (numbers 1, 2, 3, 4, and 5 are easy to work with). In addition, if you know the rules that govern algebraic operations of summation (Σ) notation, you can prove the relationship Σ (X -) = 0.

Second, the mean is the point about which the sum of the squared deviations is minimized. If we subtract the mean from each score, square each deviation, and add the squared deviations together, the resulting sum will be smaller than if any number other than the mean had been used; that is, Σ (X -)^2 is a minimum. You can demonstrate this relationship for yourself by playing with some numbers.

Example 1:

A commercial fish farm wants to advertise and as part of their promotion plan they want to tell customers how much their typical fish weighs. To keep things simple for the moment, suppose they catch fish having weights 1.1, 2.3, 1.7, 0.9 and 3.1 pounds. The fish farm does not want to report all five weights to the public but rather one number that conveys the typical weight among the five fish caught. For these five fish, a measure of the typical weight is the sample mean,

= (1.1+2.3+1.7+0.9+3.1) = 1.82.

Example 2:

You sample ten married couples and determine the number of children they have. The results are 0, 4, 3, 2, 2, 3, 2, 1, 0, 8. The sample mean is = (0+4+3+2+2+3+2+1+0+8)/10 = 2.5. Of course, nobody has 2.5 children. The intention is to provide a number that is centrally located among the 10 observations with the goal of conveying what is typical. The sample mean is frequently used for this purpose, in part because it greatly simplifies technical issues related to methods covered in subsequent chapters. In some cases, the sample mean suffices as a summary of data, but it is important to keep in mind that for various reasons, it can highly unsatisfactory. One of these reasons is illustrated next.

Example 3:

Imagine an investment firm is trying to recruit you. As a lure, they tell you that among the 11 individuals currently working at the company, the average salary, in thousands of dollars, is 88.7. However, on closer inspection, you find that the salaries are

30,25,32,28,35,31,30,36,29,200,500,

where the two largest salaries correspond to the vice president and president, respectively. The average is 88.7, as claimed, but an argument can be made that this is hardly typical because the salaries of the president and vice president result in a sample mean that gives a distorted sense of what is typical. Note that the sample mean is considerably larger than 9 of the 11 salaries.

Example 4:

A scholar conducted a study, a portion of which dealt with the attitudes of post graduate students towards number of jobs in career . Among other things, the student were asked how many jobs they desired over the next 30 years. The responses of 105 males are shown in table 4. The sample mean is = 64.9. But this is hardly typical because 102 of the 105 males gave a response less than the sample mean.

Table 4: Responses by Males in the Job Attitude Study

6	1	1	3	1	1	1	1	1	1	6	1	1	1	4
5	3	9	1	1	1	5	12	10	4	2	1	1	4	45
8	5	0	1	150	13	19	2	1	18	3	1	3	1	11
1	2	1	1	1	12	1	1	2	6	1	1	1	1	4
1	150	6	40	4	30	10	1	1	0	3	4	1	4	7
1	10	0	19	1	9	1	1	1	5	0	1	1	15	4
1	4	1	1	11	1	1	30	12	6000	1	0	1	1	15

Outliers are values that are unusually large or small. In the last example, one participant responded that he wanted 6,000 jobs over the next 30 years, which is clearly unusual compared to the other 104 students. Also, two gave the response 150, which again is relatively unusual. An important point made by these last two examples is that the sample mean can be highly influenced by one or more outliers. That is, care must be exercised when using the sample mean because its value can be highly atypical and therefore potentially misleading. Also, outliers are not necessarily mistakes or inaccurate reflections of what was intended. For example, it might seem that nobody would seriously want 6,000 jobs, but a documentary on the job hoppers made it clear that such individuals do exist. Moreover, similar studies conducted within a wide range of countries confirm that generally a small proportion of individuals will give a relatively extreme response.

7.3 Median

The median is the point that divides a distribution of scores into two parts that are equal in size. To find the median of the Student Center expense data, arrange the daily expenditures from highest to lowest, as shown in Table 5. Because there are 14 scores, the halfway point, or median, will have seven scores above it and seven scores below it. The seventh score from the bottom is \$2.40. The seventh score from the top is \$2.50. The median, then, is halfway between these two scores, or \$2.45. Remember, the median is a hypothetical point in the distribution; it may or may not be an actual score.

Table 5: Data of Table 2 Arranged in Descending Order

X	
\$8.50	
6.78	
4.47	
4.20	7 scores
3.81	
3.25	
2.50	
	Median = \$2.45*
2.40	
1.75	
0.00	
0.00	7 scores
0.00	
0.00	
0.00	

*(The halfway point between two numbers is the mean of the two numbers. Thus, (\$2.40 + \$2.50)/2 =\$2.45.)

What is the interpretation of a median of \$2.45? The simplest interpretation is that on half the days our student spends less than \$2.45 in the Student Center and on the other half he spends more.

What if there had been an odd number of days in the sample? Suppose the student chose to sample half a month, or 15 days. Then the median would be the eighth score. The eighth score has seven scores above and seven below. For example, if an additional day was included, during which \$3.12 was spent, the median would be \$2.50. If the additional day's expenditure was zero, the median would be \$2.40. The reasoning you just went through can be expressed in formula form. The formula for finding the location of the median is:

Median location = $= \frac{N+1}{2}$

The location may be at an actual score (as in the second example) or a point between two scores (the first example).

Example 5:

Seven individuals are given a test that measures depression. The observed scores are

34,29,55,45,21,32,39.

Because the number of observations is n=7, which is odd, m =(7+1)/2=4.

Putting the observations in order yields 21,29,32,34,39,45,55.

The fourth observation is X(4) = 34, so the sample median is M = 34.

Example 6:

We repeat the last example, only with six test scores 29,55,45,21,32,39.

Because the number of observations is n = 6, which is even, m = 6/2 = 3.

Putting the observations in order yields 21,29,32,39,45,55.

The third and fourth observations are X(3) = 32 and X(4) = 39, so the sample median is

M = (32+39)/2 = 35.5.

Example 7:

Consider again the data in example 3 dealing with salaries. We saw that the sample mean is 88.7. In contrast, the sample median is M = 31, providing a substantially different impression of the typical salary earned. This illustrates that the sample median is relatively insensitive to outliers, for the simple reason that the smallest and largest values are trimmed away when it is computed. For this reason, the median is called a resistant measure of location. The sample mean is an example of a measure of location that is not resistant to outliers.

Example 8:

As previously noted, the sample mean for the job attitude data in table 4 is = 64.9. But the median is M =1, which provides a substantially different perspective on what is typical.

7.3.1 The Median for a Simple Frequency Distribution

The formula for finding the location of the median that you used earlier works for a simple frequency distribution, too.

Median location =

Thus, for the scores in Table 3,

Median location = = = 50.5

To find the 50.5th position, begin adding the frequencies in Table 3 from the bottom (2+2+2+1+. . .). The total is 43 by the time you include the score of 24. Including 25 would make the total 52—more than you need. So the 50.5th score is among those nine scores of 25. The median is 25.

Suppose you start the quest for the median at the top of the distribution rather than at the bottom. Again, the location of the median is at the 50.5th position in the distribution. To get to 50.5, add the frequencies from the top (2+1+ 2+ . . .). The sum of the frequencies of the scores from 35 down to and including 26 is 48. The next score, 25, had a frequency of 9. Thus, the 50.5th position is among the scores of 25. The median is 25. Calculating the median by starting from the top of the distribution produces the same result as calculating the median by starting from the bottom.

7.3.2 The Sample Mean Versus the Sample Median

How do we choose between the mean and median? It might seem that because the median is resistant to outliers and the mean is not, use the median. But the issue is not this simple. Indeed, for various reasons, both the mean and median can be highly unsatisfactory. What is needed is a good understanding of their relative merits. To complicate matters, even when the mean and median have identical values, it will be seen that for purposes beyond merely describing the data, the choice between these two measures of location can be crucial. It is also noted that although the median can better reflect what is typical, in some situations its resistance to outliers can be undesirable.

Example 9:

Imagine someone invests \$200,000 and reports that the median amount earned per year, over a 10-year-period, is \$100,000. This sounds great, but now imagine that the earnings for each year are: \$100,000, \$200,000, \$200,000, \$200,000, \$200,000, \$200,000, \$200,000, \$300,000, \$300,000, \$-1,900,000. So at the end of 10 years this individual has earned nothing and in fact lost the \$200,000 initial investment. (The sample mean is 0.) Certainly the long-term total amount earned is relevant in which case the sample mean provides a useful summary of the investment strategy that was followed.

7.4 Mode

The third central tendency statistic is the mode. As mentioned earlier, the mode is the most frequently occurring score—the score with the highest frequency. For the Student Center expense data, the mode is \$0.00. Table 4 shows the mode most clearly. The zero amount occurred five times, and all other amounts occurred only once. When a mode is given, it is often helpful to tell the percentage of times it occurred. You will probably agree that "The mode was \$0.00, which was observed on 36 percent of the days" is more informative than "The mode was \$0.00."

The mode of numbers is the number that occurs most frequently. If two numbers tie for most frequent occurrence, the collection has two modes and is called bimodal.

7.5 Comparing Measures of Central Tendency

Comparing Measures of Central Tendency

Example 10:

On an interview for a job, the interviewer tells you that the average annual income of the company's 25 employees is \$60,849. The actual annual incomes of the 25 employees are shown below. What are the mean, median, and mode of the incomes? Was the person telling you the truth?

\$17,305,	\$478,320,	\$45,678,	\$18,980,	\$17,408,
\$25,676,	\$28,906,	\$12,500,	\$24,540,	\$33,450,

$12,500,	$33,855,	$37,450,	$20,432,	$28,956,
$34,983,	$36,540,	$250,921,	$36,853,	$16,430,
$32,654,	$98,213,	$48,980,	$94,024,	$35,671

The mean of the incomes is

$$\text{Mean} = \frac{17{,}305+478{,}320+45{,}678+\ldots\ldots\ldots\ldots\ldots 35{,}671}{25} \quad \frac{1{,}521{,}225}{25} = \$\ 60{,}849$$

To find the median, order the incomes as follows:

$12,500,	$12,500,	$16,430,	$17,305,	$17,408,
$18,980,	$20,432,	$24,540,	$25,676,	$28,906,
$28,956,	$32,654,	$33,450,	$33,855,	$34,983,
$35,671,	$36,540,	$36,853,	$37,450,	$45,678,
$48,980,	$94,024,	$98,213,	$250,921,	$478,320

From this list, you can see that the median (the middle number) is $33,450. From the same list, you can see that $12,500 is the only income that occurs more than once. So, the mode is $12,500. Technically, the person was telling the truth because the average is (generally) defined to be the mean. However, of the three measures of central tendency Mean: $60,849 Median: $33,450 Mode: $12,500 it seems clear that the median is most representative. The mean is inflated by the two highest salaries.

7.6 Choosing a Measure of Central Tendency

Which of the three measures of central tendency is the most representative? The answer is that it depends on the distribution of the data and the way in which you plan to use the data.

For instance, in above Example, the mean salary of $60,849 does not seem very representative to a potential employee. To a city income tax collector who wants to estimate 1% of the total income of the 25 employees, however, the mean is precisely the right measure.

Example 11:

Which measure of central tendency is the most representative of the data shown in each frequency distribution?

a. Number	Tally	b. Number	Tally	c. Number	Tally
1	7	1	9	1	6
2	20	2	8	2	1
3	15	3	7	3	2
4	11	4	6	4	3
5	8	5	5	5	5
6	3	6	6	6	5
7	2	7	7	7	4
8	0	8	8	8	3
9	15	9	9	9	0

a. For this data, the mean is 4.23, the median is 3, and the mode is 2. Of these, the mode is probably the most representative.

b. For this data, the mean and median are each 5 and the modes are 1 and 9 (the distribution is bimodal). Of these, the mean or median is the most representative.

c. For this data, the mean is 4.59, the median is 5, and the mode is 1. Of these, the mean or median is the most representative.

7.7 When to Use the Mean, Median and Mode?

A common question is: Which measure of central tendency should I use? The general answer is, given a choice, use the mean. Sometimes, however, the data give you no choice. Here are three considerations that limit your choice.

7.7.1 Scale of Measurement

A mean is appropriate for ratio or interval scale data, but not for ordinal or nominal distributions. A median is appropriate for ratio, interval, and ordinal scale data, but not for nominal data. The mode is appropriate for any of the four scales of measurement.

For an ordinal scale such as class standing in college, either median or mode makes sense. The median would probably be sophomore, and the mode would be freshman.

7.7.2 Open-Ended Class Intervals

Even if you have interval or ratio data, there is a circumstance in which you cannot calculate a mean. This circumstance is when the class interval with the highest (or lowest) scores is open-ended. As an example, age data are sometimes reported with the highest category being "75 and over." In such cases, there is no midpoint, and therefore you cannot compute a mean. Medians and modes are appropriate measures of central tendency when one or both of the extreme class intervals are open-ended.

7.7.3 Skewed Distributions

Even if you have interval or ratio data and all class intervals have midpoints, using a mean is not recommended if the distribution is severely skewed. The example 7 demonstrates why the mean gives an erroneous impression for severely skewed distributions.

Table 6: Data characteristics and recommended central tendency statistic

	Recommended statistic		
Data characteristic	**Mean**	**Median**	**Mode**
Nominal scale data	No	No	Yes
Ordinal scale data	No	Yes	Yes
Interval scale data	Yes	Yes	Yes
Ratio scale data	Yes	Yes	Yes
Open-ended category(ies)	No	Yes	Yes
Skewed distribution	No	Yes	Yes

7.8 Empirical Relation between Mean, Median and Mode

A distribution in which the values of mean, median and mode coincide (i.e. mean = median = mode) is known as a symmetrical distribution. Conversely, when values of mean, median and mode are not equal the distribution is known as asymmetrical or skewed distribution. In moderately skewed or asymmetrical distribution a very important relationship exists among these three measures of central tendency. In such distributions the distance between the mean and median is about one-third of the distance between the mean and mode, Karl Pearson expressed this relationship as:

Mode = mean - 3 [mean - median]

Mode = 3 median - 2 mean

and Median = 1/3[mode +2 mean]

Knowing any two values, the third can be computed.

Example 12:
Given median = 20.6, mode = 26
Find mean.
Mode = 3 Median – 2 Mean
Mean = ½[3Median – Mode]
Mean = ½[3(20.6) – (26)]
Mean = ½ [35.8]
Mean = 17.9

7.9 Weighted Mean

Sometimes several sample means are available from the same or similar populations. In such cases, a weighted mean,is the best estimate of the population parameter, μ If every sample has the same N, you can compute a weighted mean by adding the means and dividing by the number of means. If the sample means are based on N's of different sizes, however, you cannot use this procedure. Here is a story that illustrates the right way and the wrong way to calculate a weighted mean.

In one of the colleges, a student with a cumulative grade point average (GPA) of 3.25 in the middle of the junior year was eligible to enter a program to "graduate with honors." Discovering this rule after sophomore year, one student decided to figure out if he had a chance to qualify. Calculating a cumulative GPA seemed easy enough to do: Four semesters had produced GPAs of 3.41, 3.63, 3.37, and 2.16. Given another GPA of 3.80, the sum of the five semesters would be 16.37, and dividing by 5 gave an average of 3.27, well above the required 3.25.

Graduating with honors seemed like a great ending for college, so he embarked on a goal-oriented semester— a GPA of 3.80 (a B in language and an A in everything else). And, at the end of the semester he had accomplished the goal. Unfortunately, "graduating with honors" was not to be.

There was a flaw in his method of calculating his cumulative GPA. His method assumed that all of the semesters were equal in weight, that they had all been based on the same number of credit hours. His calculations based on this assumption are shown on the left side of Table 6.

Unfortunately, all five semesters were not the same; the semester with the GPA of 2.16 was based on 19 hours, rather than the usual 16 or so. Thus, that semester should have been weighted more heavily than semesters with fewer hours.

The formula for a weighted mean is

$$\bar{X}_W = \frac{N1\bar{X}1 + N2\bar{X}2 + \cdots + Nk\bar{X}k}{N1 + N2 + \cdots + Nk}$$

Where,

$\bar{X}$ = the weighted mean

$\bar{X}_{1,} \bar{X}_{2,} \bar{X}_3$ = sample means

N_1, N_2, N_k = sample sizes

K = number of samples

Table 7: Two methods of calculating a weighted mean from five semesters' GPAs; the method on the left is correct only if all semesters have the same number of credit hours

Flawed method	Correct method		
Semester GPA	Semester GPA	Credit hours	GPA X hours
3.41	3.41	17	58
3.63	3.63	16	58
3.37	3.37	19	64
2.16	2.16	19	41
3.80	3.80	16	61
Σ = 16.37		Σ = 87	Σ = 282

$\bar{X}_W = \frac{16.37}{5} = 3.27$ $\bar{X}_W = \frac{282}{87} = 3.24$

The right side of Table 6 shows the steps for a weighted mean, which is required to correctly calculate a cumulative GPA. Each semester's GPA is multiplied by its number of credit hours. These products are summed and that total is divided by the sum of the hours. As you can see from the numbers on the right, the actual cumulative GPA was 3.24, not high enough to qualify for the honors program.

More generally, to find the mean of a set of means, multiply each separate mean by its *N,* add these products together, and divide the total by the sum of the *N*'s. As an example, three means of 2.0, 3.0, and 4.0, calculated from the scores of samples with *N*'s of 6, 3, and 2, produce a weighted mean $\bar{X}_w$ of 2.64. Do you agree?

7.10 Summary

Measures of central tendency, particularly the mean and the median, are some of the most used and useful statistics for researchers. They each provide important information about an entire distribution of scores in a single number. For example, we know that the average height of a man in the India is five feet eight inches. This single number is used to summarize information about millions of men in this country. But for the same reason that the mean and median are useful, they can often be dangerous if we forget that a statistic such as the mean ignores a lot of information about a distribution, including the great amount of variety that exists in many distributions. Without considering the variety as well as the average, it becomes easy to make sweeping generalizations, or stereotypes, based on the mean.

7.11 KeyWords

- **Distribution:** A collection, or group, of scores from a sample on a single variable. Often, but not necessarily, these scores are arranged in order from smallest to largest.
- **Mean:** The arithmetic average of a distribution of scores.
- **Median split:** Dividing a distribution of scores into two equal groups by using the median score as the divider. Those scores above the median are the "high" group whereas those below the median are the "low" group.
- **Median:** The score in a distribution that marks the 50th percentile. It is the score at which 50% of the distribution falls below and 50% fall above.
- **Mode:** The score in the distribution that occurs most frequently.
- **Negative skew:** In a skewed distribution, when most of the scores are clustered at the higher end of the distribution with a few scores creating a tail at the lower end of the distribution.
- **Outliers:** Extreme scores that are more than two standard deviations above or below the mean.
- **Skew:** When a distribution of scores has a high number of scores clustered at one end of the distribution with relatively few scores spread out toward the other end of the distribution, forming a tail.
- **Weighted mean** Overall mean calculated from two or more samples with different N's.

7.12 Self Assessment Test

1. For each of the following situations, tell which measure of central tendency is appropriate and why.

   a. As part of a study on prestige, an investigator sat on a corner in a high-income residential area and classified passing automobiles according to color: black, gray, white, silver, green, and other.

   b. In a study of helping behavior, an investigator pretended to have locked himself out of his car. Passersby who stopped were classified on a scale of 1 to 5 as (1) very helpful, (2) helpful, (3) slightly helpful, (4) neutral, and (5) discourteous.

   c. In a study of household income in a city, the following income categories were established: \$0–\$20,000, \$20,001–\$40,000, \$40,001– \$60,000, \$60,001–\$80,000, \$80,001–\$100,000, and \$100,001 and more.

   d. In a study of per capita income in a city, the following income categories were established: \$0–\$20,000, \$20,001–\$40,000, \$40,001–\$60,000, \$60,001–\$80,000, \$80,001–\$100,000, \$100,001–\$120,000, \$120,001–\$140,000, \$140,001–\$160,000, \$160,001–\$180,000, and

   \$180,001–\$200,000.

e. First admissions to a state mental hospital for 5 years were classified by disorder: schizophrenic, delusional, anxiety, dissociative, and other.

f. A teacher gave her class an arithmetic test; most of the children scored in the range 70–79. A few scores were above this, and a few were below.

2. Which of the following distributions is bimodal?
   a. 10, 12, 9, 11, 14, 9, 16, 9, 13, 20

b. 21, 17, 6, 19, 23, 19, 12, 19, 16, 7

c. 14, 18, 16, 28, 14, 14, 17, 18, 18, 6

3. A senior psychology major performed the same experiment on three groups and obtained means of 74, 69, and 75 percent correct. The groups consisted of 12, 31, and 17 participants, respectively. What is the overall mean for all participants?

Answers for Self Assessment Test

1. a. The mode is appropriate because nominal variable events are being observed.

b. The median or mode is appropriate because helpfulness is measured with an ordinal variable.

c. The median and mode are appropriate for data with an open-ended category.

d. The median and mode are appropriate measures. It is conventional to use the median for income data because the distribution is often severely skewed. (About half of the frequencies are in the \$0–\$20,000 range.)

e. The mode is appropriate because these are nominal data.

f. The mean is appropriate because the data are not severely skewed.

2. It may help to arrange the scores in descending order. Only distribution c has two modes. They are 14 and 18.

3. This problem calls for a weighted mean.

$$74 \times 12 = 888$$

$$69 \times 31 = 2139$$

$$\frac{7 \times 17}{\sum 60} = \frac{1275}{\sum 4302}$$

$$\bar{X}_w = \frac{4302}{60} = 71.7$$

7.13 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

Unit - 8 : Measures of Variation and Skewness

Unit Structure

8.0 Objectives
8.1 Introduction
8.2 Range
8.3 Interquartile Range
8.4 Variance and Standard Deviation
8.5 Deviation Scores
8.6 Skewness
8.7 Types of Skewness
8.8 Summary
8.9 Key Words
8.10 Self Assessment Test
8.11 References

8.0 Objectives

After going through this unit, you should be able to:

- Explain the concept of variability
- Find and interpret the range of a distribution
- Find and interpret the interquartile range of a distribution
- Distinguish among the standard deviation of a population, the standard deviation of a sample used to estimate a population standard deviation, and the standard deviation used to describe a sample
- Calculate the variance of a distribution
- Explain the concept of skewness

8.1 Introduction

Measures of central tendency, such as the mean and the median provide useful information. But it is important to recognize that these measures are limited and, by themselves, do not provide a great deal of information. There is an old saying that provides a caution about the mean: "If your head is in the freezer and your feet are in the oven, on average you're comfortable."

The value of knowing about variability is illustrated by the story of two brothers who went water skiing on Christmas Day (the temperature was about 35°F). On the average each skier finished his turn 5 feet from the shoreline (where one may step off the ski into only 1 foot of very cold water). This bland average of the two actual stopping places, however, does not convey the excitement of the day.

The first brother, determined to avoid the cold water, held onto the towrope too long. Scraped and bruised, he finally stopped rolling at a spot 35 feet up on the rocky shore. The second brother, determined not to share the same fate, released the towrope too soon. Although he swam the 45 feet to shore very quickly, his lips were very blue. No, the average stopping place of 5 feet from the shoreline doesn't capture the excitement of the day. To get the full story, you should ask about variability.

Here are some other situations in which knowing the variability is important.

1. You are an elementary school teacher who has been assigned a class of fifth graders whose mean IQ is 115, well above the IQ of 100 that is the average for the general population. Because children with IQs of 115 can handle more complex, abstract material, you plan many sophisticated projects for the year.

   Caution: If the variability of the IQs in the class is small, your projects will probably succeed. If the variability is large, the projects will be too complicated for some, but for others, even these projects will not be challenging enough.

2. Suppose that your temperature, taken with a thermometer under your tongue, is 97.5°F. You begin to worry. This is below even the average of 98.2°F.

   Caution: There is variability around that mean of 98.2°F. Is 97.5°F below the mean by just a little or by a lot? Measuring variability is necessary if you are to answer this question.

3. Having graduated from college, you are considering two offers of employment, one in sales and the other in management. The pay is about the same for both. Using the library to check out the statistics for salespeople and managers, you find that those who have been working for 5 years in each type of job also have similar averages. You conclude that the pay for the two occupations is equal.

   Caution: Pay is more variable for those in sales than for those in management. Some in sales make much more than the average and some make much less, whereas the pay of those in management is clustered together. Your reaction to this difference in variability might help you choose.

   The information that measures of variability provide is completely independent of that provided by the mean, median, and mode. Table 1 shows three distributions, each with a mean of 15. As you can see, however, the actual scores are quite different. Fortunately, measures of variability reveal the differences in the three distributions, giving you information that the mean does not.

Table 1: Illustration of Three Different Distributions

	X1	X2	X3
	25	17	90
	20	16	30
	15	15	15
	10	14	0
	5	13	- 60
Mean	15	15	15

One aspect of most sets of data is that the values are not all alike; indeed, the extent to which they are unlike, or vary among themselves, is of basic importance in statistics. Consider the following examples:

In a hospital where each patient's pulse rate is taken three times a day, that of patient A is 72, 76, and 74, while that of patient B is 72, 91, and 59. The mean pulse rate of the two patients is the same, 74, but observe the difference in variability. Whereas patient A's pulse rate is stable, that of patient B fluctuates widely.

A supermarket stocks certain 1-pound bags of mixed nuts, which on the average contain 12 almonds per bag. If all the bags contain anywhere from 10 to 14 almonds, the product is consistent and satisfactory, but the situation is quite different if some of the bags have no almonds while others have 20 or more.

Measuring variability is of special importance in statistical inference. Suppose, for instance, that we have a coin that is slightly bent and we wonder whether there is still a fifty-fifty chance for heads. What If we toss the con 100 times and get 28 heads and 72 tails? Does the shortage of heads-only 28 where we might have expected 50-imply that the count is not "fair?" To answer such questions we must have some idea about the magnitude of the fluctuations, or variations, that are brought about by chance when coins are tossed 100 times.

8.2 Range

The range is simply the difference between the largest score (the maximum value) and the smallest score (the minimum value) of a distribution. This statistic gives researchers a quick sense of how spread out the scores of a distribution are, but it is not a particularly useful statistic because it can be quite misleading. For example, in our depression survey we may have 1 student score a 1 and another score a 20, but the other 98 may all score 10. In this example, the range will be 19 (20 – 1 = 19), but the scores really are not as spread out as the range might suggest. Researchers often take a quick look at the range to see whether all or most of the points on a scale, such as a survey, were covered in the sample. The range of a quantitative variable is the highest score minus the lowest score.

Range = $X_H - X_L$

Where, X_H = highest score

X_L = lowest score

Conceptually, the range is easy to understood, its calculation is very easy, and there is a natural curiosity about the smallest and largest values. Nevertheless, it is not a very useful measure of variation – its main shortcoming being that it does not tell us anything about the dispersion of the values that fall between the two extremes. For example, each of the following three sets of data

Set A:	5	18	18	18	18	18	18	18	18	18
Set B:	5	5	5	5	5	18	18	18	18	18
Set C:	5	6	8	9	10	12	14	15	17	18

has a range of 18 -5 = 13, but their dispersions between the first and last values are totally different.

In actual practice, the range is used mainly as a "quick and easy" measure of variability; for instance, in industrial quality control it is use to keep a close check on raw materials and products on the basis of small samples taken at regular intervals of time.

8.3 Interquartile Range

The next measure of variability, the interquartile range, tells the range of scores that make up the middle 50 percent of the distribution. To find the interquartile range, you must have the 25th percentile score and the 75th percentile score. The 10th percentile score has 10 percent of the distribution below it; it is near the bottom. The 95th percentile score is near the top; 95 percent of the scores in the distribution are smaller. The 50th percentile divides the distribution into equal halves. The median is the point that divides a distribution into equal halves. Thus, the median is the 50th percentile. Finding the 25th and 75th percentile scores involves the same kind of thinking used to find the median. The 25th percentile score is the one that has 25 percent of the scores below it. Look at Table 2, which shows a frequency distribution of 40 scores. To find the 25th percentile score, multiply 0.25 times N (0.25 X 40 = 10). When N = 40, the 10th score up from the bottom is the 25th percentile score. You can see in Table 2 that there are 5 scores of 23 or lower. The 10th score is among the 7 scores of 24. Thus, the 25th percentile is a score of 24.

The 75th percentile score has 75 percent of the scores below it. The easiest way to find it is to work from the top, using the same multiplication procedure, 0.25 times N (0.25 X 40 = 10). The 75th percentile score is the 10th score down from the top of the distribution. In Table 3.8 there are 9 scores of 29 or higher. The 10th score is among the 4 scores of 28, so the 75th percentile score is 28.

TABLE 2: Finding the 25th and 75th percentiles

Score	f	
31	1	
30	3	9 scores
29	5	
28	4	
27	5	
26	4	
25	6	
24	7	
23	3	5 scores
22	2	
	N = 40	

The interquartile range is simply the 75th percentile minus the 25th percentile:

IQR = 75th percentile - 25th percentile

where IQR = interquartile range.

Thus, for the distribution in Table 2, IQR = 28 – 24 = 4.

The interpretation is that the middle 50 percent of the scores have values from 24 to 28. .

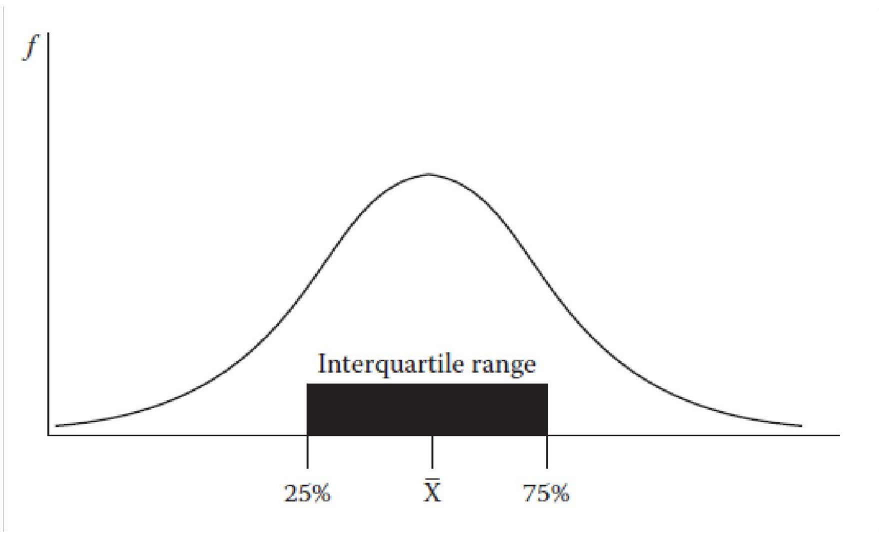


Figure 1: Interquartile Range

Example 1:

Consider the 10 values 50, 50, 50, 50, 50, 50, 50, 50, 50, 50. The interquartile range is zero. If we decrease the first value to 20 and increase the last to 80, the interquartile range is still zero because it measure the variability of the central half of the data, while ignoring the upper and lower fourth of the observations. Indeed, no matter how small we make the first value, and no matter how much we increase the last value, the interquartile range remains zero.

8.4 Variance and Standard Deviation

The variance provides a statistical average of the amount of dispersion in a distribution of scores. Because of the mathematical manipulation needed to produce a variance statistic, variance, by itself, is not often used by researchers to gain a sense of a distribution. In general, variance is used more as a step in the calculation of other statistics (e.g., analysis of variance) than as a stand-alone statistic. But with a simple manipulation, the variance can be transformed into the standard deviation, which is one of the statistician's favorite tools.

The best way to understand a standard deviation is to consider what the two words mean. Deviation, in this case, refers to the difference between an individual score in a distribution and the average score for the distribution. So if the average score for a distribution is 10 (as in our previous example), and an individual child has a score of 12, the deviation is 2. The other word in the term standard deviation is standard. In this case, standard means typical, or average. So a standard deviation is the typical, or average, deviation between individual scores in a distribution and the mean for the distribution. This is a very useful statistic because it provides a handy measure of how spread out the scores are in the distribution. When combined, the mean and standard deviation provide a pretty good picture of what the distribution of scores is like.

In a sense, the range provides a measure of the total spread in a distribution (i.e., from the lowest to the highest scores), whereas the variance and standard deviation are measures of the average amount of spread within the distribution. Researchers tend to look at the range when they want a quick snapshot of a distribution, such as when they want to know whether all of the response categories on a survey question have been used (i.e., did people use all 5 points on the 5-point Likert scale?) or they want a sense of the overall balance of scores in the distribution. Researchers rarely look at the variance alone, because it does not use the same scale as the original measure of a variable, although the variance statistic is very useful for the calculation of other statistics (such as analysis of variance). The standard deviation is a very useful statistic that researchers constantly examine to provide the most easily interpretable and meaningful measure of the average dispersion of scores in a distribution.

8.5 Deviation Scores

A deviation score is a raw score minus the mean of the distribution, whether the distribution is a sample or a population.

Deviation Score = $X - \bar{X}$ or $X - \mu$

Raw scores that are greater than the mean have positive deviation scores, raw scores that are less than the mean have negative deviation scores, and raw scores that are equal to the mean have a deviation score of zero.

Table 3: The Computation of Deviation Scores from Raw Scores

Name	Score	$X - \bar{X}$	Deviation score
Selene	14	14 – 8	6
Ian	10	10 – 8	2
Luke	8	8 – 8	0
Zachary	5	5 – 8	–3
Stephen	3	3 – 8	–5
	$\Sigma X = 40$		$\Sigma (X - \bar{X}) = 0$*

$$\bar{X} = \frac{\Sigma X}{N}$$

$$\bar{X} = \frac{40}{5} = 8$$

*Notice that the sum of the deviation scores is always zero. Add the deviation scores; if the sum is not zero, you have made an error.

Table 3 provides an illustration of how to compute deviation scores for a small sample of data. In Table 3, first computed the mean, 8, and then subtracted it from each score. The result is deviation scores, which appear in the right-hand column. A deviation score tells you the number of points that a particular score deviates from, or differs from, the mean. In Table 3, the X - value for Selene, 6, tells you that she scored six points above the mean. Luke, 0, scored at the mean, and Stephen, -5, scored five points below the mean.

There are two central issues that need to be addressed when considering the formulas for calculating the variance and standard deviation of a distribution: (1) whether to use the formula for the sample or the population, and (2) how to make sense of these formulas.

It is important to note that the formulas for calculating the variance and the standard deviation differ depending on whether you are working with a distribution of scores taken from a sample or from a population. The reason these two formulas are different is quite complex and requires more space than allowed in a short book like this. Briefly, when we do not know the population mean, we must use the sample mean as an estimate. But the sample mean will probably differ from the population mean. Whenever we use a number other than the actual mean to calculate the variance, we will end up with a larger variance, and therefore a larger standard deviation, than if we had used the actual mean. This will be true regardless of whether the number we use in our formula is smaller or larger than our actual mean. Because the sample mean usually differs from the population mean, the variance and standard deviation that we calculate using the sample mean will probably be smaller than it would have been had we used the population mean. Therefore, when we use the sample mean to generate an estimate of the population variance or standard deviation, we will actually underestimate the size of the true variance in the population because if we had used the population mean in place of the sample mean, we would have created a larger sum of squared deviations, and a larger variance and standard deviation. To adjust for this underestimation, we use n - 1 in the denominator of our sample formulas. Smaller denominators produce larger overall variance and standard deviation statistics, which will be more accurate estimates of the population parameters.

Table 4: Variance and Standard Deviation Formulas

	Population	Estimate Based on a Sample
Variance	$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$	$s^2 = \frac{\Sigma(X-\bar{X})^2}{n-1}$
	Where Σ = to sum X = a score in the distribution μ = the population mean N= the number of cases in the population	Where Σ = to sum X= a score in the distribution $\bar{X}$ = the sample mean n= the number of cases in the sample
Standard Deviation	$\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}$	$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}}$
	Where, Σ = to sum X = a score in the distribution μ = the population mean N= the number of cases in the population n	Where, Σ = to sum X= a score in the distribution = the sample mean = the number of cases in the sample

The formulas for calculating the variance and standard deviation of a population and the estimates of the population variance and standard deviation based on a sample are presented in Table 4. As you can see, the formulas for calculating the variance and the standard deviation are virtually identical. Because both require that you calculate the variance first, we begin with the formulas for calculating the variance (see the upper row of Table 4). This formula is known as the deviation score formula.

When working with a population distribution, the formulas for both the variance and the standard deviation have a denominator of N, which is the size of the population. In the real world of research, particularly social science research, we usually assume that we are working with a sample that represents a larger population. For example, if I study the effectiveness of my new reading program with a class of second graders, as a researcher I assume that these particular second graders represent a larger population of second graders, or students more generally.

Because of this type of inference, researchers generally think of their research participants as a sample rather than a population, and the formula for calculating the variance of a sample is the formula more often used. Notice that the formula for calculating the variance of a sample is identical to that used for the population, except the denominator for the sample formula is n-1.

How much of a difference does it make if we use N or n-1 in our denominator? Well, that depends on the size of the sample. If we have a sample of 500 people, there is virtually no difference between the variance formula for the population and for the estimate based on the sample. After all, dividing a numerator by 500 is almost the same as dividing it by 499. But when we have a small sample, such as a sample of 10, then there is a relatively large difference between the results produced by the population and sample formulas.

To illustrate, suppose that you are calculating a standard deviation. After crunching the numbers, you find a numerator of 100. You divide this numerator by four different values depending on the sample size and whether you divide by N or n-1. The results of these calculations are summarized in Table 5. With a sample size of 500, subtracting 1 from the denominator alters the size of the standard deviation by less than one one-thousandth. With a sample size of 10, subtracting 1 from the denominator increases the size of the standard deviation by nearly 2 tenths. Note that in both the population and sample examples, given the same value in the numerator, larger samples produce dramatically smaller standard deviations. This makes sense because the larger the sample, the more likely each member of the sample will have a value near the mean, thereby producing a smaller standard deviation.

Table 5: Effects of Sample Size and n-1 on Standard Deviation

	N = 500	N = 10
Population	$\sigma = \sqrt{}(100/500) = .44721$	$\sigma = (100/10) = 3.16$
Sample	$s = (100/499) = .44766$	$s = (100/9) = 3.33$

The second issue to address involves making sense of the formulas for calculating the variance. In all honesty, there will be very few times that you will need to use this formula. Thankfully, all computer statistics and spreadsheet programs, and many calculators, compute the variance and standard deviation for us. Nevertheless, it is mildly interesting and quite informative to examine how these variance formulas work.

To begin this examination, keep in mind that the variance is simply an average of a distribution. To get an average, you need to add up all of the scores in a distribution and divide this sum by the number of scores in the distribution, which is n (remember the formula for calculating the mean). With the variance, however, you need to remember that you are not interested in the average score of the distribution. Rather, you are interested in the average difference, or deviation, between each score in the distribution and the mean of the distribution.

To get this information, we have to calculate a deviation score for each individual score in the distribution (see Figure 2). This score is calculated by taking an individual score and subtracting the mean from that score. If you compute a deviation score for each individual score in the distribution, then you can sum the deviation scores and divide by n to get the average, or standard deviation, right?

The problem here is that, by definition, the mean of a distribution is the mathematical middle of the distribution. Therefore, some of the scores in the distribution will fall above the mean (producing positive deviation scores), and some will fall below the mean (producing negative deviation scores). When you add these positive and negative deviation scores together, the sum will be zero. Because the mean is the mathematical middle of the distribution, you will get zero when you add up these deviation scores no matter how big or small our sample, or how skewed or normal our distribution. And because you cannot find an average of zero (i.e., zero divided by n is zero, no matter what n is), we need to do something to get rid of this zero.

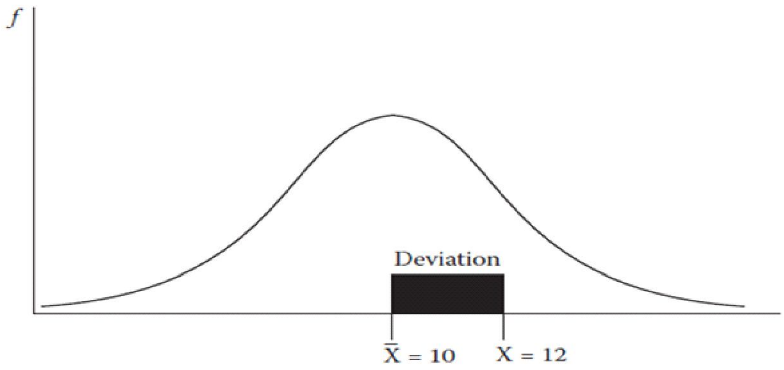


Figure - 2: Deviation

The solution statisticians came up with is to make each deviation score positive by squaring it. So, for each score in a distribution, you subtract the mean of the distribution and then square the deviation. If you look at the deviation score formulas in Table 3, you will see that all that the formula is doing with $(X - \mu)^2$ is to take each score, subtract the mean, and square the resulting deviation score. What you get when you do this is the all-important squared deviation, which is used all the time in statistics. If you then put a summation sign in front, you have $\Sigma(X - \mu)^2$. What this tells us is that after we produce a squared deviation score for each case in your distribution, you then need to add up all of these squared deviations, giving you the sum of squared deviations, or the sum of squares (SS). Once this is done, you divide by the number of cases in your distribution, and you get an average, or mean, of the squared deviations. This is your variance.

The final step in this process is converting the variance into a standard deviation. Remember that to calculate the variance, you have to square each deviation score. You do this to avoid getting a sum of zero in your numerator. When you square these scores, you change your statistic from your original scale of measurement (i.e., whatever units of measurement were used to generate our distribution of scores) to a squared score. To reverse this process and give us a statistic that is back to your original unit of measurement, you merely need to take the square root of your variance. When you do this, you switch from the variance to the standard deviation. Therefore, the formula for calculating the standard deviation is exactly the same as the formula for calculating the variance, except you put a big square root symbol over the whole formula. Notice that because of the squaring and square rooting process, the standard deviation and the variance are always positive numbers.

Table 6 presents some imaginary (but true-to-life) data on boxes of cookies sold by six Girl Scouts. These data are used to illustrate the calculation of the standard deviation, s. If these data were a population, the value of 's' would be identical.

Table 6: Deviation-Score Formula to Compute 's' for Cookie Sales

Boxes of cookies	Deviation scores	
X	**$X - \bar{X}$**	**$(X - \bar{X})^2$**
28	18	324
11	1	1
10	0	0
5	–5	25
4	–6	36
2	–8	64
$\Sigma X = 60$	$\Sigma (X - \bar{X}) = 0$	$\Sigma (X - \bar{X})^2 = 450$

$\bar{X} = \frac{\Sigma X}{n} = 60/6 = 10$ boxes

$s = \sqrt{[\Sigma (X - \bar{X})^2/N]} = \sqrt{(450/6)} = \sqrt{(75)} = 8.66$ boxes

The numbers of boxes sold are listed in the X column of Table 6. The rest of the arithmetic needed for calculating s is also given. To compute s by the deviation score formula, first find the mean. Obtain a deviation score for each raw score by subtracting the mean from the raw score. Square each deviation score and sum the squares to obtain $\sum\left(X-\bar{X}\right)^2$. Divide $\sum\left(X-\bar{X}\right)^2$ by N. Take the square root. The result is s = 8.66 boxes.

Now, what does s = 8.66 boxes mean? How does it help your understanding? The 8.66 boxes is a measure of the variability in the number of boxes the six Girl Scouts sold. If s was zero, you would know that each girl sold the same number of boxes. The closer s is to zero, the more confidence you can have in predicting that the number of boxes any girl sold was equal to the mean of the group. Conversely, the further s is from zero, the less confidence you have. With S = 8.66 and = 10, you know that the girls varied a great deal in cookie sales.

This interpretation of the standard deviation has not given you any more information than the range does. (A range of zero means that each girl sold the same number of boxes, and so forth.) The range, however, has no additional information to give you—the standard deviation does.

Now look again at Table 6 and the formula for s. Notice what is happening. The mean is subtracted from each score. This difference, whether positive or negative, is squared and these squared differences are added together. This sum is divided by N and the square root is found. Every score in the distribution contributes to the final answer, but they don't all contribute equally.

Notice the contribution made by a score such as 28, which is far from the mean; its contribution to $\sum\left(X-\bar{X}\right)^2$ is large. This makes sense because the standard deviation is a yardstick of variability. Scores that are far from the mean cause the standard deviation to be greater. Take a moment to think through the contribution to the standard deviation made by a score near the mean.

Example 2:

For various reasons, a high fiber diet is thought to promote good health. Among cereals regarded to have high fiber, is there much variation in the actual amount of fiber contained in one cup? For 11 such cereals, the amount of fiber (in grams), written in ascending order, is

7.5,8.0,8.0,8.5,9.0,11.0,19.5,19.5,28.5,31.0,36.0.

The sample mean is = 17, so the deviation scores are

-9.5,-9.0,9.0,-8.5,6,2.5,2.5,1.5,14.0,19.0

Deviation scores reflect how far each observation is from the mean, but often it is convenient and desirable to find a single numerical quantity that summarizes the amount of variation in our data. An initial suggestion might be to simply average the deviation scores. That is, we might use

$\frac{1}{n}\Sigma\,(Xi-\bar{X})$

But using the rules for summation already described, it can be seen that this average difference is always zero, so this approach is unsatisfactory. Another possibility is to use the average of the absolute deviation scores:

$\frac{1}{n} \Sigma |Xi - \bar{X}|$

This is reasonable, but it makes certain theoretical developments difficult. It turns out that using the squared differences instead greatly reduces certain mathematical problems.

That is, use what is called the sample variance, which is

$s^2 = \frac{1}{n-1} \Sigma (Xi - \bar{X})^2$

In other words, use the average squared difference from the mean. The sample standard deviation is the (positive) square root of the variance, s.

Example 3:

Imagine you sample 10 adults (n = 10), ask each to rate the performance of the president on a 10-point scale, and that their responses are: 3,9,10,4,7,8,9,5,7,8.

The sample mean is $\bar{X} = 7$, $\Sigma (Xi - \bar{X})^2 = 48$, so the sample variance is $s^2 = 48/9 = 5.33$. Consequently, the standard deviation is $s = \sqrt{(5.33)} = 2.31$.

Another way to summarize the calculations is as follows.

i	Xi	$Xi - \bar{X}$	$\left(Xi - \bar{X}\right)^2$
1	3	–4	16
2	9	2	4
3	10	3	9
4	4	–3	9
5	7	0	0
6	8	1	1
7	9	2	4
8	5	–2	4
9	7	0	0
10	8	1	1
Σ		0	48

The sum of the observations in the last column is $\sum\left(Xi - \bar{X}\right)^2 = 48$. So again, $s^2 = 48/9 = 5.33$.

The other formula (raw-score), which is algebraically equivalent, involves far fewer decimals. It also produces answers more quickly, especially if you are working with a calculator.

The formula is $S \; or \; \sigma = \sqrt{\frac{\sum X^2 - (\sum X)^2}{N}}$

where ΣX^2 = sum of the squared scores

$(\Sigma X)^2$ = square of the sum of the raw scores

N = number of scores

Although this formula may appear more forbidding, it is actually easier to use than the previous formula because you don't have to compute deviation scores. The numbers you work with will be larger, but your calculator won't mind.

Table 7 shows the steps for calculating s by this raw-score formula. The data are for boxes of cookies sold by the six Girl Scouts. You can put the arithmetic of Table 7 into words: Square the sum of the values in the X column and divide the total by N. Subtract this quotient from the sum of the values in the X^2 column. Divide this difference by N and find the square root. The result is s. Notice that the value of s in Table 7 is the same as the one you calculated in Table 6. In this case, the mean is an integer, so the deviation scores introduced no rounding errors.

Table 7: Using the raw-Score Formula to Compute *S* for Cookie Sales by a Sample of Six Girl Scouts

Boxes of cookies	
X	X^2
28	784
11	121
10	100
5	25
4	16
2	4
$\Sigma X = 60$	$\Sigma X^2 = 1050$

Note: $(\Sigma X)^2 = (60)^2 = 3600$

$$S \, or \, \sigma = \sqrt{\frac{\sum X^2 - (\sum X)^2}{N}}$$

$$\sqrt{\{[\sum X^2 - (\sum X)^2 / N] / N\}} = \sqrt{\{[1050 - (60)^2 / 6] / 6\}} = \sqrt{(\{1050 - 600\} / 6} = \sqrt{(450/6)} = \sqrt{75} = 8.66$$
boxes

ΣX^2 and $(\Sigma X)^2$: Did you notice the difference in these two terms when you were working with the data in **Table 3.12**?

If so, congratulations. You cannot calculate a standard deviation correctly unless you understand the difference. Re-examine **Table 3.12** if you aren't sure of the difference between ΣX2 and (ΣX)2.

Note: The range is usually two to five times greater than the standard deviation when N = 100 or less. The range (which can be calculated quickly) will tell you if you made any large errors in calculating a standard deviation.

8.5.1 Why Variance?

If the variance is a difficult statistic to understand, and rarely examined by researchers, why not just eliminate this statistic and jump straight to the standard deviation? There are two reasons. First, you need to calculate the variance before you can find the standard deviation anyway, so it is not more work. Second, the fundamental piece of the variance formula, which is the sum of the squared deviations, is used in a number of other statistics, most notably analysis of variance (ANOVA). When you learn about more advanced statistics such as ANOVA, factorial ANOVA, and even regression, you will see that each of these statistics uses the sum of squares, which is just another way of saying the sum of the squared deviations. Because the sum of squares is such an important piece of so many statistics, the variance statistic has maintained a place in the teaching of basic statistics.

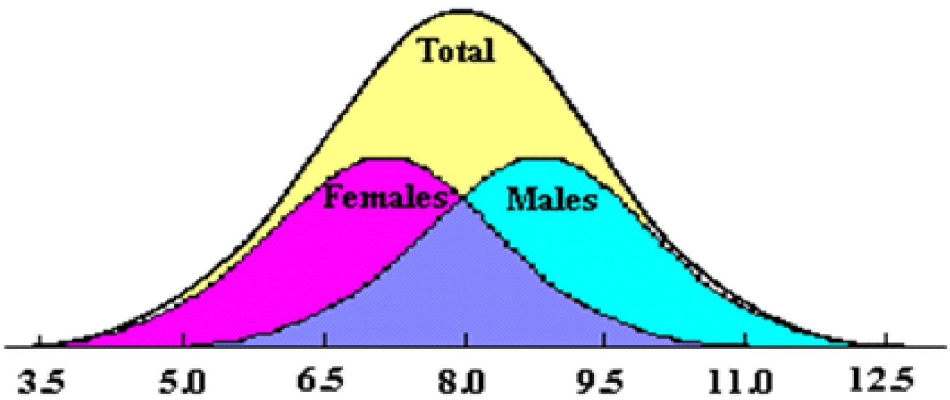
8.6 Skewness

A fundamental task in many statistical analyses is to characterize the location and variability of a data set. A further characterization of the data includes skewness. Measure of Dispersion tells us about the variation of the data set. Skewness tells you the amount and direction of variation of the data set (departure from horizontal symmetry) relative to a standard bell curve. Why do you care? One application is testing for normality: many statistics inferences require that a distribution be normal or nearly normal. A normal distribution has skewness and excess kurtosis of 0, so if your distribution is close to those values then it is probably close to normal.

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the center point.

8.7 Types of Skewness

Teacher expects most of the students get good marks. If it happens, then the cure looks like the normal curve below (Figure 3):



But for some reasons (e. g., lazy students, not understanding the lectures, not attentive etc.) it is not happening. So we get another two curves.(Figure 4)

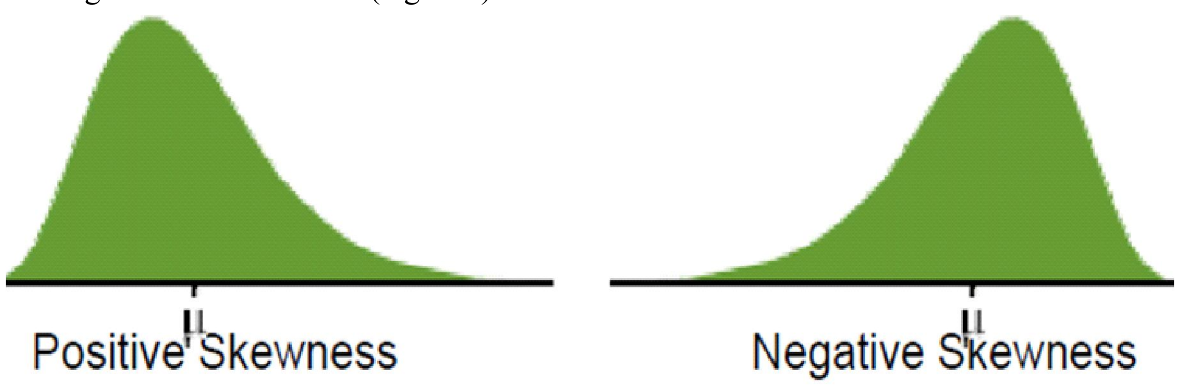
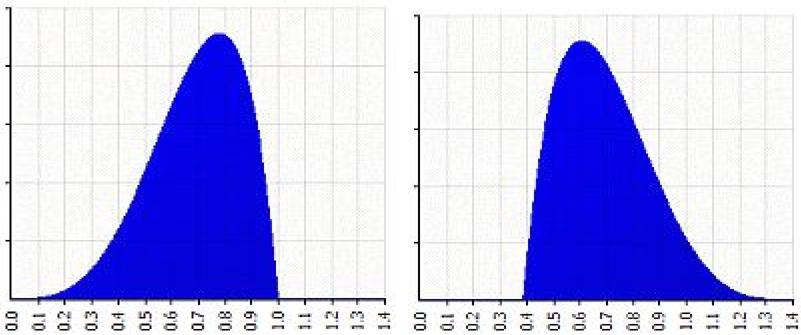


Figure 4: Positive and Negative Skewness

The first one is known as positively skewed and the second one is known as negatively skewed curve.

The first thing you usually notice about a distribution's shape is whether it has one mode (peak) or more than one. If it's unimodal (has just one peak), like most data sets, the next thing you notice is whether it's symmetric or skewed to one side. If the bulk of the data is at the left and the right tail is longer, we say that the distribution is skewed right or positively skewed; if the peak is toward the right and the left tail is longer, we say that the distribution is skewed left or negatively skewed.

Look at the two graphs below (Figure 5). They both have $\mu = 0.6923$ and $\sigma = 0.1685$, but their shapes are different.



Beta(α=4.5,β=2) 1.3846 – Beta (α=4.5,β=2)

skewness = 0.5370 skewness = +0.5370

-0.5370

Figure 5: Measure of Skewness

The first one is moderately skewed left: the left tail is longer and most of the distribution is at the right. By contrast, the second distribution is moderately skewed right: its right tail is longer and most of the distribution is at the left.

You can get a general impression of skewness by drawing a histogram, but there are also some common numerical measures of skewness. Some authors favor one, some favor another. The mean and standard deviation have the same units as the original data, and the variance has the square of those units. However, the skewness has no units: it's a pure number.

8.7.1 Measures of Skewness

1. Karl Pearson coefficient of Skewness

Sk = 3(mean - median) / Standard Deviation = $3\left(\bar{X} - Me\right)\backslash\sigma$

2. Bowley coefficient of skewness or Quartile measure of skewness $\dfrac{Q_3 + Q_1 - 2Me}{Q_3 - Q_1}$

Interpretation:

1. If Sk = 0, then the frequency distribution is normal and symmetrical.

2. If S> k 0, then the frequency distribution is positively skewed.

3. If S<k 0, then the frequency distribution is negatively skewed.

8.8 Summary

Measures of variation, such as the variance, standard deviation, and range, are important descriptive statistics. They provide useful information about how spread out the scores of a distribution are, and the shape of the distribution. Skewness tells you the amount and direction of variation of the data set (departure from horizontal symmetry) relative to a standard bell curve.

8.9 Key Words

- **Interquartile Range (IQR)**: The difference between the 75th percentile and 25th percentile scores in a distribution.
- **Range:** The difference between the largest score and the smallest score of a distribution.
- **Squared Deviation:** The difference between an individual score in a distribution and the mean for the distribution, squared.
- **Standard Deviation:** The average deviation between the individual scores in the distribution and the mean for the distribution.
- **Sum of Squared Deviations, Sum of Squares:** The sum of each squared deviation for all of the cases in the sample.
- **Variance:** The sum of the squared deviations divided by the number of cases in the population, or by the number of cases minus one in the sample.
- **Skewness:** The degree to which a distribution of scores deviates from normal in terms of asymmetrical extension of the tails.

8.10 Self Assessment Test

1. Find the range for the two distributions.

   a. 17, 5, 1, 1

   b. 0.45, 0.30, 0.30

2. The temperatures listed are averages for March, June, September, and December. Calculate the mean and standard deviation for each city. Summarize your results in a sentence.

San Francisco, CA	54°F	59°F	62°F	52°F
Albuquerque, NM	46°F	75°F	70°F	36°F

3. Compute s for the three sets of scores.

   a. 7, 6, 5, 2

   b. 14, 11, 10, 8, 8

   c. 107, 106, 105, 102

Answers for Self Assessment Test

1. a. Range = 17 - 1 = 16

   b. Range = 0.45 - 0.30 = 0.15

2.

City	Mean	Standard deviation
San Francisco	56.75°F	3.96°F
Albuquerque	56.75°F	16.24°F

Interpretation: Although the mean temperature is the same for the two cities, Albuquerque has a wider variety of temperatures. (If your mean scores are 56.69°F for San Francisco and 56.49°F for Albuquerque, you are very alert. These are weighted means based on 30 days in June and September and 31 days in March and December.)

3. a. = 5; s = 1.87

   b. = 10.2; s = 2.23

   c. = 105; s = 1.87

8.11 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

Unit - 9 : Tests of Hypothesis I

Unit Structure:

9.0 Objectives

9.1 Introduction

9.2 Point Estimation

9.3 Interval Estimation

9.4 Confidence Limits

9.5 Hypothesis

9.6 Types of Errors in Testing Hypothesis

9.7 Level of Significance

9.8 Critical Regions

9.9 One-Tailed and Two-Tailed Tests

9.10 Process of Testing Hypothesis

9.11 Testing Hypothesis about Population Mean

9.12 Testing Hypothesis about the Difference between Two Means

9.13 Test of Hypothesis Concerning Attributes

9.14 Testing Hypothesis about a Population Proportion

9.15 Testing Hypothesis about the Difference between Two Proportions

9.16 Summary

9.17 Key Words

9.18 Self Assessment Test

9.19 References Books

9.0 Objectives

After studying this unit, you should be able to understand

- The testing of Hypothesis
- The estimation theory i.e. point estimation and interval estimation
- The statistical hypothesis, null hypothesis & alternate hypothesis
- The types of errors in testing hypothesis
- The level of significance & critical regions
- One-tailed and two-tailed tests
- Process of testing hypothesis
- Z-test about population mean
- Z-test about population proportion

9.1 Introduction

The inductive inference may be termed as the logic of drawing statistically valid conclusions about the population characteristics on the basis of a sample drawn from it in a scientific manner. The problems are to generalise the results of the sample to the population, to find how far these generalisations are valid and also to estimate the population parameters along with the degree of confidence. The answers to these problems are given by a very important branch of Statistics known as the Statistical inference which may be classified into: a) Estimation Theory and b)Testing of Hypothesis

Estimation of population parameters like mean, correlation coefficient, variance, proportions etc. from the corresponding sample statistics is one of the very important problems of statistical inference. The estimation of population parameters is imperative in making business decisions. The theory of estimation was founded by Prof. R.A. Fisher and is divided into two groups namely: a) Point Estimation and b) Interval Estimation

9.2 Point Estimation

A particular value of a statistic which is used to estimate a given parameter is known as a point estimator of the parameter. A good estimator is one which is as close to the true value of the parameter as possible. A good estimator should satisfy the following criteria:

a) Unbiasedness

A statistic $t = t(x_{1,} x_{2,} x_{3,}$ ————————————, Xn), a function of the sample observations $x_{1,} x_{2,} x_{3,}$ — ————————————, X_n is said to be an unbiased estimate of the corresponding population parameter θ, if E (t) = θ i.e. if the mean value of the sampling distribution of the statistic is equal to the parameter.

For example mean $\bar{X}$ is an unbiased estimate of the population mean μ.

b) Consistency

A statistic $t = t(x_{1,} x_{2,} x_{3,}$ ————————————, x_n) based on a sample of size n is said to be a consistent estimator of the parameter θ if it converges in probability to θ

if $\quad t_n \to \theta \; as \; n \to \infty$

Symbolically, lim $P(t_n \to \theta) = 1$

$$P(t_n \to \theta) = 1$$
$$n \to \infty$$

Sample mean x is a consistent estimator of the population mean, sample proportion 'p' is a consistent estimator of population proportion P.

c) Efficiency

If there is more than one consistent estimators of a parameter θ then the efficiency is the criterion which enables us to choose between them by considering the variances of the sampling distributions of the estimators. Thus, if t_1 and t_2 are consistent estimators of a parameter θ such that

Var (t_1) = Var (t_2), for all n.

Then t_1 is said to be more efficient than t_2.

d) Sufficiency

A statistic $t = t(x_{1,} x_{2,} x_{3,}$————————————, x_n) is said to be a sufficient estimator of parameter θ if it contains all the information in the sample regarding the parameter. In other words, a sufficient statistic utilises all the information that a given sample can furnish about the parameter. Properties of sufficient estimators are given below:

- It is always consistent.

- It may or may not be unbiased.

- A minimum variance unbiased estimator for a parameter exists iff there exists a sufficient estimator for it.

- If a sufficient estimator exists for some parameter then it is also most efficient estimator.

9.3 Interval Estimation

In point estimation, a single value of a statistic is used as an estimate of the population parameter. But even the best possible point estimate may deviate enough from the true parameter value to make the estimate unsatisfactory. Thus to get a statistic 't' from a given random sample, the problem arises 'Can we make some reasonable probability statements about the unknown parameter θ in the population from which the sample has been drawn?'

The answer to this problem is given by Neyman, a pioneered of Interval Estimation. This consists in the determination of two C_1 and C_2 such that

- $P [c_1 < \theta < c_2$, for given value of t] = 1- α

Where α is the level of=significance.

The interval $[c_1 , c_2]$, within which the unknown value of the parameter θ is expected to lie is known as Confidence Limits and 1- α is called the confidence coefficient .

For example, α = 0.01 or 0.05 gives the 99% or 95% confidence limits.

If 't' is the statistic used to estimate the parameter θ then

(1- α) confidence limits for θ = t± S.E (t) X $t_{\alpha/2}$

Where $t_{\alpha/2}$ is the critical value of t at level of significance α for a two-tailed test. Thus the computation of confidence limits for a parameter θ involves the following steps.

a) Compute the appropriate sample statistic 't'.

b) Obtain the Standard Error of the sampling distribution of the statistic t i.e. S.E. (t).

c) Choose appropriate confidence coefficient (1- α), depending on the precision of the estimate.

For large samples, the distribution of the standardised variate corresponding to the sampling distribution of the statistic t will be

$$Z = \frac{t - E(t))}{S.E.(t)}$$

9.4 Confidence Limits

Confidence Limits for Mean μ

If we consider a large random sample of size from an infinite population, with mean μ and variance σ^2 then

$$Z = \bar{X} - \mu / \sigma\sqrt{n}$$

Confidence Limits for Standard Deviation

In a random sample of size n from an infinite population with variance σ^2, the variance of the sampling distribution of standard deviation is given by

$$Var(s) = \sigma^2 \backslash 2n$$

Confidence Limits for Difference of Means

If $\bar{X}1$ and $\bar{X}2$ are the sample means based on large random samples of size n_1 and n_2 from two infinite population with means μ_1 and μ_2 and standard deviation σ_1 and σ_2 respectively then

$E(\bar{X}_1 - \bar{X}_2) = E\bar{X}_1) - E\bar{X}_2) = \mu_1 - \mu_2$

and $Var(\bar{X}1 - \bar{X}2) = Var(\bar{X}1) + Var\ \bar{X}2)$

[Var (ax+by) = a^2 Var (x) + b^2 Var (y) , if x and y are independent]

This implies

$Var(\bar{X}_1 - \bar{X}_2) = \sigma_1^2/n_1 + \sigma_2^2/n_2$

$S.E.(\bar{X}_1 - \bar{X}_2) = \sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}$

Thus for large samples

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - E\ \bar{X}_1 - \bar{X}_2)}{S.E.(\bar{X}_1 - \bar{X}_2)} = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$$

9.5 Hypothesis

The inductive inference is based on deciding about the characteristics of the population on the basis of sample study. Such decisions involve the risk of taking inaccurate decisions. E.g. a pharmaceutical concern may be interested to find if a new drug is effective for the particular disease say in reducing blood pressure. Modern theory of probability plays a very important role in decision making, the branch of statistics which helps us in arriving the criteria for such decisions is known as testing of hypothesis. J.Neyman and E.S. Pearson developed this theory of testing of hypothesis and employs statistical techniques to arrive at decisions in certain situations where there is an element of uncertainty.

9.5.1 Statistical Hypothesis

A statistical hypothesis is some statement which may or may not be true about a population. If the hypothesis completely specifies the population then it is known as simple hypothesis otherwise it is composite hypothesis.

9.5.2 Null Hypothesis

Null hypothesis is a hypothesis of no difference. It is denoted by H_0. According to Prof. R.A.Fisher – Null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.

Null Hypothesis is expressed as

$$H_0 : \theta = \theta_0$$

9.5.3 Alternative Hypothesis

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

$H_0 : \mu = \mu_0$

Then, the alternative hypothesis could be :

(i) $H_1 = \mu \neq \mu_0$ (i.e. $\mu > \mu_0$ or $\mu < \mu_0$)

(ii) $H_1 : \mu > \mu_0$

(iii) $H_1 : \mu < \mu_0$

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

9.6 Types of Errors in Testing of Hypothesis

As already stated, the inductive interference consists in arriving at a decision to accept or reject a null hypothesis (H_0) after inspecting only a sample from it. As such, an element of risk – the risk of taking wrong decisions is involved. In any test procedure, the four possible mutually disjoint and exhaustive decisions are:

(i) Reject H_0 when actually it is not true, i.e. when H_0 is false.

(ii) Accept H_0 when it is true.

(iii) Reject H_0 when it is true.

(iv) Accept H_0 when it is false.

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

		Decision from sample	
		Reject H_0	**Accept H_0**
True State	**H_0 True**	**Wrong (Type I Error)**	**Correct**
	H_0 False (H_1 True)	**Correct**	**Wrong (Type II Error)**

Thus, in testing of hypothesis, we are likely to commit two types of errors. The error of rejecting H_0 when H_0 is true is known as Type I error and the error of accepting H_0 when H_0 is false (i.e. is true) is known as Type II Error.

If we write:

$$P[\text{Reject } H_0 \text{ when it is true}] = P[\text{Type I Error}] = \alpha$$

And $P[\text{Accept } H_0 \text{ when it is wrong}] = P[\text{Type II Error}] = \beta$

Then α and β are also called the sizes of type I error and type II error respectively.

Therefore, $\alpha = P_r\,[\text{Type I error}]$

$= P_r\,[\text{Rejecting } H_0 / H_0 \text{ is true}]$

Must be the complement of

$(1 - \alpha) = P_r\,[\text{Accepting } H_0 / H_0 \text{ is true}]$

This probability $(1-\alpha)$ corresponds to the concept of 100 (1 - α) % confidence interval. Our efforts would obviously be to have a small probability of making a type I error. Hence the objective is to construct the test to minimize α.

Similarly, the probability of committing a type II error is designated by β

$\beta = P_r\,[\text{Type II error}]$

$= P_r\,[\text{Accepting } H_0 / H_0 \text{ is false}]$

$1- \beta = P_r\,[\text{Rejecting } H_0 / H_0 \text{ is false}]$

This probability (1- β) is known as the power of a statistical test.

The table shown below gives the probabilities associated with each of the four cells shown in the previous table:

The decision :	**The Null hypothesis** True	False
Accept H_0	**(1 - α)** **Confidence level**	**β**
Reject H_0	**α**	**1 - β** **Power of the test**
Sum	**1.00**	**1.00**

9.7 Level of Significance

The maximum size of type I error which we are prepared to risk is known as the level of significance. It is usually denoted by α and is given by:

$P\,[\text{Rejecting } H_0 \text{ when } H_0 \text{ is true}] = \alpha$

Commonly used levels of significance in practice are 5% (.05) and 1% (.01). If we adopt 5% level of significance, it implies that in 5 samples out of 100, we are likely to reject a correct H_0. In other words, this implies that we are 95% confident that our decision to reject H_0 is correct. Level of significance is always fixed in advance before collecting the sample information.

9.8 Critical Regions

Suppose, we take several samples of the same size from a given population and compute some statistic t, (say x, p etc.). for each of these samples. Let $t_1, t_2 .... t_k$ be the values of the statistics for these samples. Each of these values may be used to test some null hypothesis H_0. Some values may lead to rejection of H_0 while others may lead to acceptance of H_0. These sample statistics $t_1, t_2 .... t_k$ (comprising the sample space), may be divided into two mutually disjoint groups, one leading to the rejection of H_0 give us a region called Critical Region (C) or Rejection Region (R), while those which lead to the acceptance of H_0 give us a region called Acceptance Region (A). Thus, if the statistics $t \approx C$, H_0 is rejected and if $t \approx A$, H_0 is accepted.

The sizes of type I and type II errors in terms of the critical region are defined below:

α = P [Rejecting H_0 when H_0 is true]

= P [Rejecting $H_0 \backslash H_0$]

= P [t õC \ H_0]

β = P [Accept H_0 when H_0 it is wrong]

= P [Accept H_0 when H_1 is true]

= P [Accept $H_0 \backslash H_1$]

= P [$t \approx A \backslash H_1$]

Where C is the critical (rejection) region, A is acceptance region and C”A = Ø, C”A = S (sample space)

Example 1 We need to test whether a coin is perfect or not, we toss it 5 times. The null hypothesis of perfectness is rejected if and only if more than 4 heads are obtained. We will found:

i. Critical region

ii. Probability of type I error and

iii. Probability of type II error when the corresponding probability of getting a head is 0.2

Solution: Let A denote the no. of heads obtained when the coin is tossed 5 times

H_0 : The coin is perfect i.e. unbiased i.e. $H_0 : p = 1/2$

Under H_0, $X \approx B\,(n = 5,\ p = ½)$

Therefore $P\,(X = x \backslash H_{0-}) = {}^5C_x\, p^x q^{n-x} = {}^5C_x (1/2)^x (1/2)^{5-x}$

$P\,(X = x \backslash H_{0-}) = {}^5C_x (1/2)^x = {}^5C_x\, x\ 1/32;\quad x = 0,1,.........,5$

(i) Critical Region or Rejection Region

Reject H_0 if more than 4 heads are obtained.

Critical region = {X > 4} = { X = 5}

(ii) Probability of type I error (α) is given by:

$\alpha =$ P [Reject $H_0 \setminus H_0$] = P[X = 5 \ H_0] = 1/32 x $^5C_5$ = 1/32 = .03125

(iii) The probability of type II error (β) is given by

β = P [Accept $H_0 \setminus H_0$] = 1 - P [Reject $H_0 \setminus H_0$]

= 1 – P [X = 5 \ p = 0.2] = 1 – $[^5C_x\, p^x\, q^{5-x}]_{x=5,\, p=0.2}$

= 1- $^5C_5\,(0.2)^5$ = 1 - .00032 = 0.99968

Example 2 We need to test whether a coin is perfect or not, we toss it 5 times. The null hypothesis of perfectness is accepted if and only if only 3 heads are obtained. Then the power of the test corresponding to the alternative hypothesis that probability of head is 0.4, is

i. 272 / 3125

ii. 2853 / 3125

iii. 56 / 3125

iv. None of the above

Solution: Let A denote the no. of heads obtained when n = 5 (i.e. the coin is tossed 5 times)

And let p = probability of a head in a random toss of coin

Null Hypothesis H_0 : p = 1 /2

Alternative Hypothesis H_1 : p = 0.4 Critical Region : X > 3

Power of the test for testing H_0 against H_1 is given by:

1 - β = P [Reject H_0 when H_1 is true] = P [Reject H_0 À H_1]

= P (X > 3 / p = 0.4) = $^5C_r(0.4)^r(0.6)^{5-r}$ [Because X – B (n = 5, p = 0.4 under H_1)]

= $^5C_r\,(0.4)^4 + (0.6) + 5 \times (4/10)^4 \times 6/10 + (4/10)^5$

= $(2/5)^4\,[3 + 2/5]$ = 16 x 17 / 625 x 5 = 272 / 3125

Hence the correct answer is part (i)

9.9 One-Tailed and Two-Tailed Tests

In any test, critical region is represented by a portion of the area under the probability curve ofthe sampling distribution of the test statistics.

A test of any statistical hypothesis where the alternative hypothesis is one-tailed (right-tailed or left-tailed) is called a one-tailed test. For example – a test for testing the mean of a population

$H_0 : \mu = \mu_0$

-Against the alternative hypothesis

$H_1 : \mu > \mu_0$ (Right tailed) or $H_1 : \mu < \mu_0$ (Left tailed),

Is a single tailed test. In the right tailed test ($H_1 : \mu > \mu_0$) the critical region lies entirely in the right-tail of the sampling distribution of x bar while for the left-tailed test ($H_1 : \mu < \mu_0$), the critical region is entirely in the left tail of the distribution of x bar.

A test of statistical hypothesis where the alternative hypothesis is two-tailed such as:

$H_0 : \mu = \mu_0$

Against the alternative hypothesis

$H_1 : \mu \neq \mu_0$ ($\mu > \mu_0$ and $\mu < \mu_0$)

Is known as two-tailed test and in such a case the critical region is given by the portion of the area lying in both the tails of the probability curve of the test statistics. In a particular problem, whether one-tailed or two tailed test is to be applied depends entirely on the nature of the alternative hypothesis is one-tailed, we apply one-tailed test.

For example, suppose that there are two population brands of bulbs, one manufactured by standard process (with mean life μ_1) and the other manufactured by some new technique (with mean life μ_2). If we want to test if the bulbs differ significantly, then our null hypothesis is $H_0 : \mu = \mu_0$ and alternative hypothesis will be $H_0 : \mu_1$ "' μ_2, thus giving us a two-tailed test. Similarly, for testing if the product of the new process is inferior to that of standard process, we have:

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 < \mu_2$

Thus giving us a left-tailed test. Similarly, for testing if the product of the new process is inferior to that of standard process, we have:

$H_0 : \mu_1 = \mu_2$ and $H_1 : \mu_1 > \mu_2$

Thus giving us a right-tailed test. Thus the decision about applying a two-tailed test or a single tailed (right or left) test will depend on the problem under study.

9.10 Process of Testing Hypothesis

Process of hypothesis testing can be explained in following steps:

a) Hypothesis Set-up

The first step in testing hypothesis is to establish the hypothesis to be tested. Since statistical hypothesis are usually assumptions about the value of some unknown parameter, the hypothesis specifies a value or range of values for the parameter. The conventional approach to testing hypothesis is not to construct single hypothesis about the population parameter but rather to setup two different hypothesis. These are:

- Null Hypothesis
- Alternate Hypothesis

The rejection of null hypothesis means that the differences have statistical significance and the acceptance of the null hypothesis means that the differences are due to chance.

b) Set-up a suitable significance level

Having set up a hypothesis, the next step is to select a suitable level of significance α. The confidence with which an observer rejects or accepts null hypothesis depends on the significance level adopted. In practice, we take 5% or 1% level of significance. 5% level of significance means that there are about 5 chances out of 100 that we would reject the null hypothesis when it should be accepted. 1% level of significance means that there is only one chance out of 100 that we would reject the null hypothesis when it should be accepted.

c) **A suitable test statistic determination**

The third step is to determine a suitable test statistic and its distribution.

$$\text{Test statistic} = \frac{\text{Sample statistic} - \text{Hypothesised population parameter}}{\text{Standard error of the sample statistic}}$$

d) **Critical region determination**

Before the sample is taken, it is important to specify which values of the test statistic will lead to the rejection of H_0 and which lead to acceptance of H_0. The former is called critical region. Thus establishing a critical region is similar to determining a $100(1 - \alpha)\%$ confidence interval.

e) **Computations**

The fifth step in testing hypothesis is the performance of various computations from a random sample of size n for test statistic obtained in step (3). Then we need to see whether sample result falls in the critical region or in the acceptance regions.

f) **Decision Making**

Finally we may draw statistical conclusions and the management may take decisions. The conclusion comprises either accepting the null hypothesis or rejecting it. The decision will depend on whether the computed value of the test criterion falls in the region of rejection or the region of acceptance.

9.11 Testing Hypothesis about Population Mean

a) We will first take the hypothesis testing concerning the population parameter μ by considering the two-tailed test.

$H_{0\,:}\ \mu = \mu_0$ (μ_0 is hypothesis value of μ)

Since the best unbiased estimator of μ is the sample mean $\bar{X}$ (X bar), therefore we will focus our attention on the sampling distribution of x bar. From central limit theorem, we know

$$\bar{X} \text{ - } N(\mu, \sigma_{\bar{X}})$$

$$z = (x \text{ bar} - \mu) / \sigma_{\bar{X}} \qquad \text{(if } \sigma \text{ is known)}$$

where $\sigma_{\bar{X}} = \sigma / \sqrt{n} = s / \sqrt{n}$ (if σ is known for large samples)

If the calculated value of $z < -z_{\alpha/2}$ or $> z_{\alpha/2,}$ the null hypothesis is rejected

b) If the hypothesis involves a right-tailed test. For example:

$H_{0\,:}\ \mu < \mu_0$ and $H_{1\,:}\ \mu > \mu_0$

For the calculated value $z > z_\sigma$ null hypothesis is rejected.

c) If the hypothesis involves a left-tailed test. For example:

$H_{0\,:}\ \mu\ e \geq \mu_0$ and $H_{1\,:}\ \mu < \mu_0$

For the calculated value $z < -z_\sigma$ null hypothesis is rejected.

Example 1:

The mean lifetime of a sample of 100 bulbs produced by a company is found to be 1580 hours with standard deviation of 90 hours. Test the hypothesis that the mean lifetime of the bulbs produced by the company is 1600 hours.

Solution:

Null Hypothesis : There is no significant difference between sample mean and population mean.

$$H_{0:}\ \mu = \mu_0 \quad \text{and} \quad H_{1:}\ \mu \ \sqrt{\ }\ \mu_0$$

$$z = (\text{x bar} - \mu) / \sigma_{\bar{X}}$$

$$\sigma_{\bar{X}}\ {}^{r} = s/\sqrt{n}$$

$$z = \frac{1580-1600}{90/\sqrt{100}}$$

$$= -2.22$$

The critical value is $z = \pm 1.96$ for a two-tailed test at 5% level of significance. We reject Null hypothesis because value of z falls in the rejection region.

Example 2:

A sample of 400 students is found to have a mean height 171.38cms. Can it be reasonably regarded as a sample from a large population of mean height 171.17 cms and standard deviation of 3.30 cms?

Solution

Null Hypothesis : There is no significant difference between sample mean and population mean height.

$$H_0 : \mu = \mu_0 \quad \text{and} \quad H_1 : \mu \neq \mu_0$$

$$z = (\bar{X} - \mu) / \sigma_{\bar{X}}$$

$$\sigma_{\bar{X}} = s/\sqrt{n}$$

$$z = \frac{171.38-171.17}{3.30/\sqrt{400}}$$

$$= 1.31$$

$z <$ critical value of $z = 1.96$ at 5% level of significance.

We accept Null hypothesis. Hence there is no significant difference between the sample mean height and population mean height.

Example 3:

Consider the following data obtained from a sample of 1000 labourers, calculate the standard error of mean.

Weekly Earnings(Rs. Hundred)	No. of Labourers
0-10	50
10-20	100
20-30	150
30-40	200
40-50	200
50-60	100
60-70	100
70-80	100

Is it likely that the samples have come from the population with average weekly earnings of Rs 4200?

Solution

Weekly earnings	X	f	(X-45)/10 d	fd	fd^2
0-10	5	50	-4	-200	800
10-20	15	100	-3	-300	900
20-30	25	150	-2	-300	600
30-40	35	200	-1	-200	200
40-50	45	200	0	0	0
50-60	55	100	1	100	100
60-70	65	100	2	200	400
70-80	75	100	3	300	900
		n=1000		Σfd=-400	$\Sigma fd^2 = 3900$

$\bar{X} = A + \Sigma fd * i/n$

$= 45 - 400*10/1000$

$= 41$

$s = \sqrt{[\Sigma fd^2/n - (\Sigma fd/n)^2]} * i$

$= \sqrt{[3900/1000 - (-400/1000)^2]} * 10$

$= 19.34$

$\sigma_{\bar{X}} = s/\sqrt{n}$

$= 19.34/\sqrt{1000}$

$= 19.34/31.62$

$= 0.612$

$z = (\bar{X} - \mu) / \sigma_{\bar{X}}$

$= (41-42)/0.612$

$= -1.634$

Since z < critical value is $z = \pm 1.96$ at 5% level of significance, it is not significant and hence there is no significant difference between the sample average and the population average weekly earnings.

9.12 Testing Hypothesis about the Difference between Two Means

The test statistic for testing the difference between two population means when the populations are normally distributed, is given below

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$$

The null hypothesis is $H_0: \mu_1 - \mu_2 = 0$

Then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$$

At 5% level of significance, the critical value of z for two-tailed test $= \pm 1.96$. If the computed value of z is greater than 1.96 or less than -1.96 then reject H_0 otherwise accept H_0. If σ_1^2 and σ_2^2 are not known then for large samples, s_1^2 and s_2^2 can be used.

Example 1:

Intelligence test given to two groups of males and females gave the following information:

	Mean Score	S.D.	Number
Males	70	12	100
Females	75	10	50

Is the difference in the mean scores of males and females statistically significant?

Soluation:

Null Hypothesis: The difference in the mean scores of males and females is not significant

i.e. $\mu_1 = \mu_2$

$\bar{X}_1 = 75 \quad \bar{X}_2 = 70$

$s_1^2 = 100 \; s_2^2 = 144$

$n_1 = 50 \;\; n_2 = 100$

$$z = \frac{\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}}$$

$$= \frac{\bar{X}_1 - \bar{X}_2)}{\sqrt{(s_1^2 / n_1 + s_2^2 / n_2)}}$$

= (75 -70)/ “3.44

= 2.695

Since z > critical value z = 2.58 at 1% level of significance therefore the hypothesis is rejected. Hence the difference in the mean score of males and females is statistically significant.

Example 2:

In a manufacturer company, information about a product given by two suppliers is

	Mean	S.D.	Size
Supplier A	1300	82	100
Supplier B	1288	93	100

Which brand of product you are going to purchase if you desire to take a risk of 5%?

Solution:

Null Hypothesis: The difference in the brand of product is not significant

i.e. $\mu_1 = \mu_{2.}$

$\bar{X}_1 = 1300 \quad \bar{X}_2 = 1288$

$n_1 = 100 \;\; n_2 = 100$

$$z = \frac{(x_1 - x_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}}$$

$$z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{(s_1^2 / n_1 + s_2^2 / n_2)}}$$

$= (1300 - 1288) * 10 / \sqrt{[(82)^2 + (93)^2]}$

$= 0.968$

Since z < critical value z = 1.96 at 5% level of significance therefore the hypothesis is accepted. Hence the quality of two brands does not differ significantly.

9.13 Test of Hypothesis Concerning Attributes

As distinguished from variables where quantitative measurement of a phenomenon is possible, in case of attributes we can find out presence or absence of a certain characteristic. For example attribute employment, people are classified as employed and unemployed. The binomial type of problem can be formed with such data.

$z = (x - np) / \sqrt{npq}$

where np is mean and "npq is standard deviation of binomial probability model.

Example:

In 600 throw of dice, odd points appeared 360 times. Check it whether dice is fair at 5% level of significance.

Solution:

$p = q = ½$

$n = 600$, $np = 300$

$z = (x - np) / \sqrt{npq} = (360 - 300) / \sqrt{600 * 1/2 * 1/2} = 4.9$

Since z > critical value z = 1.96 at 5% level of significance. The hypothesis is rejected. Hence the dice does not seem to be fair.

9.14 Testing Hypothesis about a Population Proportion

The population parameter of interest is population proportion 'pi' (π). If the sample size is large, then sample proportion p will be approximately normally distributed. Then

$$z = (\kappa - \theta) / \sigma_{\hat{e}} \sim N(0,1)$$

The null hypothesis is that there is no significance difference between the sample proportion and population proportion, i.e. $H_0 : p = \pi$

Since the sample proportion p is unbiased estimator of π,

$$z = (p - \pi) / \sigma_p \text{ ; where } \sigma_p = \sqrt{[\pi(1 - \pi)]} / \sqrt{n}$$

Therefore, the statistic

$z = (p - \pi) / [\sqrt{\{\pi(1 - \pi)\}} / \sqrt{n}]$

Example 1:

A mall surveyor of a mall concludes that 60% of the shoppers entering in the mall are not doing any purchasing. A Also, a sample of 50 shoppers was surveyed and result showed that 35 of them left without buying anything. Are these sample results consistent with the claim of the sales clerk? Use a level of significance of 0.05.

Solution

The null hypothesis is

$H_0 : \pi = 0.60$

The sample proportion p = 35 / 60 = 0.70

Using the z statistic, we have:

$z = (p - \pi) / [\sqrt{\{\pi(1 - \pi)\}} / \sqrt{n}]$

$= (0.70 - 0.60) / [\sqrt{(0.6)(0.4) / 50}]$

$= 1.45$

The critical value of z is 1.64 at 5% level of significance.

Since, the computed value of z = 1.45 is less than the critical value of z = 1.64, therefore, the null hypothesis cannot be rejected. Hence, based on this sample data, we cannot reject the claim of the sales clerk.

Example 2 :

A dice is thrown 49152 times and of these 25145 yielded either 4 or 5 or 6. Is this consistent with the hypothesis that the dice must be unbiased.

Solution

Null Hypothesis : Dice is unbiased.

Coming 4 or 5 or 6 is considered as success.

n = 49152

p = 25145/49152 = 0.512

$\pi = 0.5$

$z = (p - \pi) / [\sqrt{\{\pi(1 - \pi)\}} / \sqrt{n}] = z = (p - \pi) / [\sqrt{\{p(1 - p)\}} / \sqrt{n}]$

$= (0.512 - 0.5)/[\sqrt{(0.512 \sqrt{0.488/49152})}]$

$= 6.0$

Since computed z > critical value z = 3.0. Therefore null hypothesis is rejected. Hence the dice is certainly biased.

9.15 Testing Hypothesis about the Difference between Two Proportions

Let p_1 and p_2 be the sample proportions obtained in large samples of size n_1 and n_2 drawn from respective populations having proportions π_1 and π_2. We can test null hypothesis H_0.

The sampling distribution of differences in proportion $p_1 - p_2$ is normally distributed with mean

$\mu_{p1-p2} = \pi_1 - \pi_2$

and standard deviation $\sigma_{p1-p2} = \sqrt{[(\pi_1(1-\pi_1)/n_1) + (\pi_2(1-\pi_2)/n_2)]}$

Therefore $z = (p_1 - p_2) - (\pi_1 - \pi_2) / \sqrt{[(\pi_1(1-\pi_1)/n_1) + (\pi_2(1-\pi_2)/n_2)]}$

If the null hypothesis is true, p_1 and p_2 are two independent unbiased estimators of the same parameter $\pi_1 = \pi_2 = \pi$. The pooled estimate of π is the weighted mean of the two sample proportions i.e.

$p = (n_1p_1 + n_2p_2)/(n_1 + n_2)$

Then $z = (p_1 - p_2)/\sigma_{p1-p2}$

Where $\sigma_{p1-p2} = \sqrt{p(1-p)(1/n_1 + 1/n_2)}$

Example 1:

A new product is to be advertised. Company considers two different television advertisements for promotion. Management thinks that advertisement A is more effective than advertisement B. Two test market areas with virtually identical consumer characteristics are selected: advertisement A is used in one area and advertisement B is used in other area. In a random sample of 60 customers who saw advertisement A, 18 tried the product. In a random sample of 100 customers who saw advertisement B, 22 tried the product. Does this indicate the advertisement A is more effective than advertisement B if a 5% level of significance is used?

Soluation :

We suppose that the null hypothesis be that there is no significant difference in the effectiveness of the two advertisements A & B, i.e. $H_0 = \pi_1 = \pi_2$

The appropriate statistic to be used is:

$$z = [(p_1 - p_2)(\pi_1 - \pi_2)] / (\sigma_{p1-p2})$$

$$= (p_1 - p_2) / \sqrt{\{(p(1-p)(1/n_1 + 1/n_2)\}} \qquad (\pi_1 = \pi_2)$$

Where $p_1 = x_1 / n_1$

$= 18 / 60 = 0.30, n_1 = 60$

$p_2 = x_2 / n_2$

$= 22 / 100 = 0.22, n_2 = 100$

And $p = (n_1p_1 + n_2p_2) / (n_1 + n_2)$

$= (x_1 + x_2) / (n_1 + n_2) = 40 / 160 = 0.25$

$z = (0.30 - 0.22) / \sqrt{\{(0.25(0.75)(1/60 + 1/100)\}}$

$= 0.08 / 0.071 = 1.13$

Since the calculated value of z = 1.13 is less than the critical value of z = 1.645* at 5% level of significance, therefore, the null hypothesis is accepted. Hence there is no significant difference in the effectiveness of the two advertisements A and B.

Example 2:

A product is to be inspected from two companies resp. 500 nos. of product from factory A are inspected and 12 pcs. are found to be defective. 800 nos. of product from factory A are inspected and 12 pcs. are found to be defective . Does this indicate that 5% level of significance that production at second factory is better than in first factory?

Solution :

We suppose that the null hypothesis be that there is no significant difference in the proportion of defective items in the two factories.

Where $p_1 = x_1 / n_1$

$= 12 / 500 = 0.024,$

$p_2 = x_2 / n_2$

$= 12 / 800 = 0.015$

The appropriate statistic to be used is:

$$z = (p_1 - p_2) / \sqrt{\ } \{(p(1-p)(1/n_1 + 1/n_2)\}$$

And $p = (n_1p_1 + n_2p_2) / (n_1 + n_2)$

$= (x_1 + x_2) / (n_1 + n_2) = (12 + 12) / (500 + 800) = 0.018$

$z = (0.024 – 0.015) /$ “ $\{(0.018(0.982)(0.00325)\}$

$= 0.009 / 0.0076 = 1.184$

Since the calculated value of z is less than the critical value of z = 1.96 at 5% level of significance, therefore, the null hypothesis is holds good. Hence we cannot conclude that the production in the second factory is better than in the first factory.

9.16 Summary

Estimation of population parometars like mean, correlation coefficient, variance, proportion etc. from the corresponding sample statistics is one of the very important problems of statistical inference. A particular value of a statistics which is used to estimate a given parameter is known as a point estimator of the parameter. Internal estimation is another way of estimating. Hypothesis is a tentative solution which is offered for verification. There are two types of hypothesis namely, null and alternate hypothesis. They are so linked that acceptance of one leads to rejection of other and vice-versa. Hypothesis is tested using suitable tests of significance involving a particular procedure. Normally we have to take care of Type – I and Type – II errors and level of significance.

9.17 Key Words

- **Point Estimation:** A particular value of a statistic which is used to estimate a given parameter is known as a point estimator of the parameter.
- **Null Hypothesis:** A hypothesis of no difference is null hypothesis.
- **Alternate Hypothesis:** Hypothesis which is complementary to null hypothesis is alternate hypothesis.

- **Type I Error:** The error of rejecting H_0 when H_0 is true is known as Type I error.
- **Type II Error:** The error of accepting H_0 when H_0 is false (i.e. is true) is known as Type II error.

9.18 Self Assessment Test

1. Describe the concept of point estimation and interval estimation with the help of suitable examples.
2. Discuss briefly the importance of estimation theory in decision making in the face of uncertainty.
3. Describe the important properties of a good estimator.
4. Differentiate between the following
   - Statistic and Parameter
   - Critical Region and Region of Acceptance
   - Null Hypothesis and Alternative Hypothesis.
5. Explain the following:
   - Type I and Type II errors
   - One-tailed and two-tailed tests
6. A coin is tossed 100 times under identical conditions independently giving 30 heads and 70 tails. Test at 1% level of significance whether or not the coin is unbiased. [**Solution:** Coin is biased]
7. An ambulance service claims that it takes, on the average, 8.9 minutes to reach its destination in emergency calls. To check on this claim, the agency which licenses ambulance services has then timed on 50 emergency calls, getting as mean of 9.3 minutes with a standard deviation of 1.8 minutes. At the 5% level of significance, does this constitute evidence that the figure claimed is too low.[**Solution:** Accept Hypothesis]

9.19 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

Unit - 10: Tests of Hypothesis II

Unit Structure:

10.0 Objectives

10.1 Introduction

10.2 Small Samples Versus Large Samples

10.3 Student's t-Distribution

10.4 t-Distribution for Testing Hypothesis for Mean

10.5 T-Test for Significance of Differences of Two Sample Means

10.6 T-Test for Significance of Observed Sample Correlation

10.7 Summary

10.8 Key Words

10.9 Self Assessment Test

10.10 References

10.0 Objectives

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

- 't' test, its assumptions and applications.
- 't' test for the significance of mean.
- 't' test for the significance of the difference between two sample means.
- 't' test for the significance of an observed sample correlation coefficient.

10.1 Introduction

Hypothesis is considered as the principle instrument in research. Its main function is to provide justification for aim to new experiments and observations. In fact, many experiments are carried out with the aim of testing hypothesis. In social science, where direct knowledge of population parameters is rare, hypothesis testing is used as strategy for deciding whether a sample data offer such support for a hypothesis and a generalisation can be made. Decision makers often face situations wherein they are interested in testing hypothesis on the basis of available information and then take decisions on the basis of testing. Thus hypothesis testing enables us to make probability statements about population parameters.

Now question is – what is hypothesis? When one talks about hypothesis, one means a mere assumption to be proved or disapproved. But for a researcher hypothesis is a formal question that he/she intends to resolve. Thus a hypothesis may be defined as a proposition or a set of proposition set forth as an explanation for the occurrence of some specified group of phenomenon. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable

Given below are some of the characteristics of hypothesis:

1. Hypothesis should be clear and precise.
2. Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.
3. Hypothesis should be stated as far as possible in simple terms so that the same is easily understandable by all concerned. But simplicity of hypothesis has nothing to do with significance.
4. Hypothesis should be capable of being tested.
5. Hypothesis should be limited in scope.
6. Hypothesis must be specific. A researcher must remember that narrower hypotheses are more testable generally.
7. Hypothesis should be consistent with most known facts.
8. Hypothesis should be amenable to testing within a reasonable time.
9. Hypothesis must explain the facts. This means that by using the hypothesis plus other known and accepted generalisations, one should be able to deduce the original problem condition.

10.2 Small Samples Versus Large Samples

While considering a sampling technique, the important decision about size of a sample. Different experts have different opinions. For examples some gave a suggestion that sample size should be 5% of the size of population while some have opinion that sample size should be at least 10%. However sample size depends on various factors like cost, degree of accuracy, time etc. Sampling theory does not provide any support for good estimate of sample size but we must consider the following:

- The size of sample should increase as the variation in the individual items increases.
- The greater the accuracy, the larger should be the sample size and vice-versa.

In the previous unit , we discussed z-test and in this unit ,we'll discuss t-test. Now question is when to use z-test and when to use t-test. The table given below will help us in deciding appropriate test.

	σ known	**σ unknown**
Large Sample	**z-test**	**z-test**
Small Sample	**z-test**	**t-test**

Here large sample is considered as e $\leq$ 30. At this level or large size of sample, at most all distributions tend to approximate normal distribution.

10.3 Student's t-Distribution

Let x_i (i = 1,2........,n) be a random sample of size n from a normal population with mean μ and variance σ^2 . Then student's 't' test is defined by statistic:

$$t = (\bar{X} - \mu) / s\sqrt{n}$$

$$= (\bar{X} - \mu) / s\sqrt{(n-1)}$$

Where $\bar{X} = 1/n \sum_{i=1}^{n} (x_i)$ is the sample mean

and $s^2 = 1 / (n-1) \sum_{i=1}^{n} (x_i - \bar{X})^2$

and degree of freedom = (n-1)

Applications of T-test

t -Test use for the following:

i. To test the significance of sample mean.

ii. To test the significance of differences of two sample means.

iii. To test the significance of observed sample correlation.

Assumption for Student t-Test

The following assumptions are made in the student's t test:

i. The parent population from which the sample is drawn is normal.

ii. The sample observations are independent i.e. the sample is random.

iii. The population standard deviation σ is unknown.

10.4 t-Distribution for Testing Hypothesis for Mean

Null hypothesis H_0 : $\bar{X} = \mu$

The sample has been drawn from the population with mean μ or there is no significant difference between the sample mean x and the population mean μ.

If $t_{cal} < t_{table}$

Accept Hypothesis i.e. we say that difference between x bar and μ is not significant. The sample might have been drawn from population with mean μ

If $t_{cal} > t_{table}$

Reject Hypothesis.

i.e. we say that the difference between x bar and μ is significant and sample does not seem to have been drawn from the population with mean μ

Degree of freedom = (n-1)

Example1:

A sales store is selling steel rods. The mean weekly sale of steel bars is 146.3 per store. The company goes for advertising and the mean weekly sale of 22 stores for a particular week increases to 153.7 and shows a standard deviation of 17.2. Find out if the advertising campaign is successful.

Solution:

Given values are :

$N = 22, \qquad \bar{X} = 153.7, \qquad s = 17.2$

Null Hypothesis $H_1 : \mu = 146.3$ rods i.e. the difference between $\bar{X}$ and μ is not significant. In other words, the advertising campaign is not successful.

Alternative Hypothesis $H_1 : \mu > 146.3$ rods, (Right tailed)

Under the null hypothesis, the test statistic is student's t

$t = (\bar{X} - \mu) / s\sqrt{(n-1)} \sim t_{22-1} = t_{21}$

$t = (153.7 – 146.3) / 17.2 / \sqrt{21}$

$= 7.4 \times \sqrt{21} / 17.2$

$= 33.9112 / 17.2 = 1.9716$

Tabulated value of t for 21 d.f. at 5% level of significance for single tailed test is 1.721. Since calculated value is greater than the tabulated value, it is significant and we reject the null hypothesis. Hence, the advertising campaign was successful in promoting sales.

Example 2:

A machine is manufacturing M.S. washers of thickness 0.025 cm for some type of machines. Random sample of 10 washers is selected. The mean of thickness of washers is 0.024 cm. and shows a standard deviation of 0.002 cm. Test the significance of the deviation. Value of t for 9 degrees of freedom at 5% level is 2.262.

Solution:

Given values are :

$n = 10, \qquad \bar{X} = 0.024$ cm, $\qquad s = 0.002$ cm

Null Hypothesis $H_0 : \mu = 0.025$ cm i.e. there is not significant deviation between sample

mean $\bar{X} = 0.024$ and population mean $\mu = 0.025$.

Alternative Hypothesis $H_1 : \mu = 0.025$ cm

Under H_0 Test Statistic is:

$t = (\bar{X} - \mu) / s\sqrt{(n-1)} \sim t_{10-1} = t_9$

$t = (0.024 - 0.025) / 0.002 / \sqrt{9}$

$= -0.001 \times 3 / 0.002$

$= -1.5$

Tabulated $t_{0.05}$ for 9 d.f. = 2.262. Since t < 2.262, it is not significant at 5% level of significance and hence the deviation $(x - \mu)$ is not significant.

Example 3:

A machine is weighing the weights of cartons. Out of those cartons, 10 cartons are randomly selected and weighed. The mean weight of cartons is 11.8 O_z. and shows a standard deviation of 0.15 O_z. Test the significance of the deviation of sample mean from the intended weight of 12 O_z. Value of Nu = 9 and $t_{0.05}$ = 2.26.

Solution:

Given values are :

$n = 10, \quad \bar{X} = 11.8\ O_z, \quad s = 0.15\ O_z$

Null Hypothesis $H_0 : \mu = 12\ O_z$ i.e. there is not significant deviation between sample mean x = 11.8 O_z and population mean $\mu = 12.0\ O_z$.

Alternative Hypothesis $H_1 : \mu \sqrt{\ } 12.0\ O_z$ (Two-tailed)

Test statistic. Under H_0 Test Statistic is:

$t = (\bar{X} - \mu) / \sqrt{(n-1)}\ s^2 \sim t_{10-1} = t_9$

$t = (11.8 - 12) / 0.15 / \sqrt{9}$

$= -0.2 \times 3 / 0.15$

$= -4.0$

Tabulated $t_{0.05}$ for 9 d.f. = 2.26. Since calculated t is much greater than tabulated t, it is highly significant. Hence null hypothesis is rejected at 5% level of significance and we conclude that the sample mean differs significantly from the mean $\mu = 12.0\ O_z$

Example 4:

A soap manufacturing company is selling soaps through a large number of retail shops. The mean weekly sale of soaps is 140 dozens per shop. The company goes for advertising and the mean weekly sale of 26 stores for a particular week increases to 147 dozens per shop and shows a standard deviation of 16. Find out if the advertising campaign is successful.

Solution:

Given values are :

$n = 26, \quad \bar{X} = 147$ dozens, $\quad s = 16$ dozens

Null Hypothesis $H_0 : \mu = 140$ dozens i.e. the difference between $\bar{X}$ and μ is just due to fluctuations of sampling. In other words, the advertising campaign is not successful.

Alternative Hypothesis $H_1 : \mu > 147$ dozens, (Right tailed)

Test Statistic. Under the null hypothesis H_0, the test statistic is:

$$t = (\bar{X} - \mu) / s\sqrt{(n-1)} \sim t_{26-1} = t_{25}$$

$$t = (147 - 140) / 16 / \sqrt{25}$$

$$= 7 \times 5 / 16$$

$$= 2.19$$

Tabulated value of t for 25 d.f. at 5% level of significance for single (right) tailed test is 1.798 i.e. $t_{25}(0.05) = 1.708$ [From table VIII]. Since calculated value of t is greater than the tabulated value, it is significant and we reject the null hypothesis at 5% level of significance. Hence, the increase in sales cannot be attributed to fluctuations of sampling and we conclude that the advertising campaign was successful in promoting sales.

Example 5:

A store of a company is taking the material and packing it into a bag. The bag is weighed by machine and readings of 10 samples are taken as follows:

50, 49, 52, 44, 45, 48, 46, 45, 49, 45

Test if the average bag can be taken to be 50 kg.

Solution:

Null Hypothesis $H_0 : \mu = 50$ kg, i.e. the average packing is 50 kg

Alternative Hypothesis $H_1 : \mu = 50$ kg (Two tailed)

Calculations for Sample Means and S.D.

X	**d-48**	**d^2**
50	2	4
49	1	1
52	4	16
44	-4	16
45	-3	9
48	0	0
46	-2	4
45	-3	9
49	1	1
45	-3	9
Total	-7	69

$\bar{X}$ = A + Σd / n = 48 + (-7) / 10

= 48 – 0.7

= 47.3

$s^2 = 1 / (n-1) [\Sigma d^2 – (\Sigma d)^2 / n]$

$= 1/9 [69 – (-7)^2 / 10] = \{(69-4.9) / 9\}$

= 7.12

Under H_0, the test statistic is:

$t = (\bar{X} - \mu) / s\sqrt{(n-1)}$

$t = (47.3 – 50.0) / \sqrt{(7.12 / 10)}$

$= -2.7 / \sqrt{0.712}$

= -3.2

Tabulated value of $t_{0.05}$ for 9 d.f. is 2.262 [From table VIII]. Since calculated t is greater than the tabulated value, it is significant. Hence, H_0 is rejected at 5% level of significance and we conclude that the average packing cannot be taken to be 50 kg.

Example 6:

The mean 53 is obtained from a random sample of size of 16. The sum of squares of deviations from mean is 150. Can this sample be regarded as taken from the population having 56 as mean? Also obtain 95% confidence limits for the mean.

Solution:

Given values are :

n = 16, $\bar{X}$ = 53, $\Sigma(x - \bar{X})^2 = 150$

Null Hypothesis $H_0 : \mu = 56$

against Hypothesis $H_1 : \mu = \sqrt{147}$ (Two-tailed)

$s^2 = 1\{ \Sigma(x - \bar{X})^2 \} / (n-1) = 150 / 15 = 10$

Test Statistic. Under the null hypothesis H_0, the test statistic is:

$t = (\bar{X} - \mu) / s\sqrt{(n-1)} \sim t_{16-1} = t_{15}$

$t = (53 - 56) / 3.162/\sqrt{16}$

= 4 x (-3) / 3.162

= -3.795

The critical tabulated value of t for two-tailed test at 5% level of significance is 2.131 i.e. $t_{15}(0.025) = 2.131$.

Since calculated value of t is greater than the critical value 2.131, it is significant and we reject the null hypothesis at 5% level of significance and we conclude that with 95% confidence that the given sample is not taken from the population with mean 56.

10.5 T-Test for Significance of Differences of Two Sample Means

Let us now try to understand how t test may be used to test be significance of differences of two sample means.

Null Hypothesis H_0: Two sample means x and y do not differ significantly. i.e.samples have been drawn from the population with sample mean.

$$t = ((\bar{X} - \bar{Y})) / S\sqrt{(1/n_1 + 1/n_2)}$$

$$When\, x = 1/n \sum_{i=1}^{n_1} X_1$$

$$When\, y = 1/n \sum_{j=1}^{n_1} Y_j$$

$$and\, s^2 = 1/n(n_1 + n_2 - 2)\left[\sum_{i=1}^{n_1} (X_1 - \bar{X}) + \sum_{j}^{n_1} (Y_1 - \bar{Y})^2\right]$$

$$S^2 = (n_1 s_x^2 + n_2 s_y^2)/(n_1 + n_2 - 2)$$

If $t_{cal} < t_{table}$ ACCEPT H_0

If $t_{cal} > t_{table}$ REJECTT H_0

Example 1:

A random sample of 20 daily workers of village X was found to have average daily earning of rs. 44 with sample variance 900. Another sample of 20 daily workers from village Y was found to earn on an average Rs. 30 per day with sample variance 400. Test whether the workers in village X are earning more than those in village Y.

Solution:

Suppose the daily earning of workers in the villages X & Y be denoted by x and y respectively. Given that

$n_1 = n_2 = 20$

$\bar{X} = 44$, $\bar{Y} = 30$

$s_x^2 = 900$, $s_y^2 = 900$

Null Hypothesis H_0: $\mu_x = \mu_y$ i.e. there is no significant difference in the average daily earnings of the workers in villages X and Y.

Alternative Hypothesis H_1: $\mu_x > \mu_y$ (Right–tailed)

$t = (\bar{Y} - \bar{Y}) / S\sqrt{(1/n_1 + 1/n_2)}$

$S^2 = (n_1 s_x^2 + n_2 s_y^2) / (n_1 + n_2 - 2)$

$= [(20*900 + 20*400)/38]$

$= 648.21$

$t = (44 - 30)/\sqrt{[648.21(1/20 + 1/20)]}$

$= 14/\sqrt{64.821}$

$= 1.7389$

Degree of freedom $= n_1 + n_2 - 2 = 20 + 20 - 2 = 38$

t_{table} for d.f. = 38 for right tailed test is 1.645 at 5 % level of significance.

Since $t_{cal} > t_{table}$ Reject Null Hypothesis.

H_0 is rejected i.e H_1 is accepted.

It is concluded that the workers in village X are earning more than those in village Y with 95% confidence.

Example 2:

A gymnasium is referring and using some products for increasing weights of weight gainers, Product A is used for some particular group(weight gainers)X and Product B is used for some particular group(weight gainers)Y, their subsequent weights are shown below:

Weights(kg):

Group X	42,	39,	48,	60,	41		
Group Y	38,	42,	56,	64,	68,	69,	62

Do you agree with the claim that medicine B increases the weight significantly?

Solution:

We suppose that weights of group X people be denoted by x and the weights of group X people be denoted by y.

Null Hypothesis $H_0 : \mu_1 = \mu_2$ kg, i.e. there is no significant difference between the Product A and Product B in regards they effect on increase in weight

Alternative Hypothesis $H_1 : \mu_1 < \mu_2$ (left-tailed) i.e. Product B increases the weight significantly

A = Avg. of weights (X group) = 46 = $\bar{X}$ = 230 /5 = $\Sigma x / n_1$

B = Avg. of weights (Y group) = 57 = $\bar{Y}$ = 399 / 7 = $\Sigma y / n_2$

Calculations for Sample Means and S.D.

Product A			Product B		
x	d = x–A=x-46	d^2	y	D = y–B=y-57	D^2
42	-4	16	38	-19	361
39	-7	49	42	-15	225
48	2	4	56	-1	1
60	14	196	64	7	49
41	-5	25	68	11	121
			69	12	144
			62	5	25
Total 230	0	290	Total	0	926

$\Sigma(X - \bar{X})^2 = [\Sigma d^2 - (\Sigma d)^2 / n_1]$

$= 290$

$\Sigma(Y - \bar{Y})^2 = [\Sigma d^2 - (\Sigma d)^2 / n_1]$

$= 926$

$$S^2 = 1/(n_1 + n_2 - 2)\left[\sum\left(X - \bar{X}\right)^2 + \sum\left(Y - \bar{Y}\right)^2\right]$$

$= [(290 + 926) / 10]$

$= 121.6$

Under H_0, the test statistic is:

$t = (\bar{X} - \bar{Y}) / s\sqrt{(1/n_1 + 1/n_2)} \sim t_{n1 + n2 - 2} = t_{10}$

$t = (46 - 57) / \sqrt{\{121.6 \times (1/5 + 1/7)\}}$

$= -11 / \sqrt{(121.6 \times 12/35)}$

$= -11 / \sqrt{6.457} = -1.7$

Tabulated value of t for 10 d.f. at 5% level of significance for left tailed test is -1.81 [From table VIII]. Since calculated t is less than the tabulated value, it is not significant. Hence, H_0 is accepted at 5% level of significance and we conclude that the Product A & B do not differ significantly as regards their effect on increase in weight.

Example 3:

A company is processing a product and testing machine gives the nicotine content in milligrams of two samples of tobacco as follows:

Sample A 24, 27, 26, 21, 25

Sample B 27, 30, 28, 31, 22, 36

Can it be said that two samples come from normal populations having the same mean?

Solution:

Null Hypothesis $H_0 : \mu_1 = \mu_2$ kg, i.e. the two samples have been drawn from the normal populations with the same mean.

Alternative Hypothesis $H_1 : \mu_1 \neq \mu_2$

Calculations for Sample Means and S.D.

Sample A			Sample B		
x	d = x–A=x-25	d^2	y	D = y–B=y-30	D^2
24	-1	1	27	-3	9
27	2	4	30	0	0
26	1	1	28	-2	4
21	-4	16	31	1	1
25	0	0	22	-8	64
			36	6	36
Total	-2	22	Total	-6	114

$\bar{X} = A + \Sigma d / n_1 = 25 + (-2) / 5$

$= 25 - 0.4$

$= 24.6$

$\Sigma(X - \bar{X})^2 = [\Sigma d^2 - (\Sigma d)^2 / n_1]$

$= 22 - 4/5$

$= 21.2$

$\bar{Y} = B + \Sigma d / n_2 = 30 + (-6) / 6$

$= 30 - 1$

$= 29$

$\Sigma(y - \bar{y})^2 = [\Sigma d^2 - (\Sigma d)^2 / n_1]$

$= 114 - 36/6$

$= 108$

$S^2 = 1/(n_1 + n_2 - 2)[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2]$

$= [(21.2 + 108) / 9]$

$= 14.36$

Under H_0, the test statistic is:

$t = (\bar{x} - \bar{y}) / s\sqrt{(1/n_1 + 1/n_2)} \sim t_{n1 + n2 - 2} = t_9$

$t = (24.6 - 29.0) / \sqrt{\{14.36 \times (1/5 + 1/6)\}}$

$= -4.4 / \sqrt{(14.36 \times 11/30)}$

$= -4.4 / \sqrt{5.2653} = -1.92$

Tabulated value of t for 9 d.f. at 5% level of significance for two tailed test is 2.262 [From table VIII]. Since calculated t is less than the tabulated value, it is not significant. Hence, H_0 is accepted at 5% level of significance and we conclude that the samples come from normal populations with the same mean.

Example 4:

A small shaft manufacturing workshop is manufacturing shaft on two machines. The average no. of shafts produced on machines per day are 200 and 250 with standard deviations 20 and 25 resp. on the basis of 25 days production Can you regard both the machines equally efficient at 1% level of significance?

Solution:

Given values are :

$n_1 = 25$,	$x_1 = 200$,	$s_1 = 20$
$n_2 = 25$,	$x_2 = 250$,	$s_2 = 25$

Null Hypothesis H_0. $\mu_1 = \mu_2$ i.e. both the machines are equally efficient.

Alternative Hypothesis $H_1 : \mu_x \neq \sqrt{\ } \mu_y$

Under the null hypothesis H_0, the test statistic is:

$t = (\bar{x} - \bar{y}) / s\sqrt{(1/n_1 + 1/n_2)} \sim t_{n1 + n2 - 2} = t_{48}$

We have:

$S^2 = (n_1 s_1^2 + n_2 s_2^2) / (n_1 + n_2 - 2)$

$= (25 \times 400 + 25 \times 625) / (25 + 25 - 2)$

$= 25625 / 48 = 533.85$

$t = (\bar{x} - \bar{y}) / s\sqrt{(1/n_1 + 1/n_2)}$

$= (200 - 250) / \sqrt{\{533.85(1/25 + 1/25)\}}$

$= -50 / \sqrt{(533.85 \times 0.08)}$

$= -50 / \sqrt{42.708}$

$= -7.65$

Tabulated value of t for 48 d.f. for two tailed test is 2.58 (approx.) since d.f. is very large. Since calculated value of t is greater than the tabulated value, it is highly significant and we reject the null hypothesis, machines cannot be considered as equally efficient at 1% level of significance.

Example 5:

A bicycle manufacturing company is selling cycles through a large number of retail shops. They are selling some by products of cycle also through retail shops. One of the products is analyzed for quality. The mean of two random samples of size 9 & 7 are 196.42 and 198.82 resp. The sum of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the samples be considered to have been drawn from the same normal population (apply t-test) .

Solution:

Given values are :

$n_1 = 9, \quad \bar{x} = 196.42, \quad \Sigma(x - \bar{x})^2 = 26.94$

$n_2 = 7, \quad \bar{y} = 198.82, \quad \Sigma(y - \bar{y})^2 = 18.73$

Null Hypothesis. The samples have been drawn from the same normal population, i.e.

$H_0: \mu_x = \mu_y$

Alternative Hypothesis $H_1 : \mu_x \neq \mu_y$ (Two tailed)

Under the null hypothesis H_0, the test statistic is:

$t = (\bar{x} - \bar{y}) / s\sqrt{(1/n_1 + 1/n_2)} \sim t_{n1+n2-2} = t_{14}$

We have:

$S^2 = 1/(n_1 + n_2 - 2)[\Sigma(x - \bar{x})^2 + \Sigma(y - \bar{y})^2]$

$= (26.94 + 18.73) / (9+7-2)$

$= 45.67 / 14 = 3.26$

$t = (\bar{x} - \bar{y}) / s\sqrt{(1/n_1 + 1/n_2)}$

$= (196.42 – 198.82) / \sqrt{\{3.26(1/9 + 1/7)\}}$

$= -2.40 / \sqrt{(3.26 \times 0.254)}$

$= -2.64$

Tabulated value of t for 14 d.f. for two tailed test is 2.15. Since calculated value of t is greater than the tabulated value, it is significant and we reject the null hypothesis at 5% level of significance the samples cannot be considered to have come from the same normal populations.

10.6 T-Test for Significance of Observed Sample Correlation

Let us how try to understand how t test is used to test the significance of observed sample correlation.

If r is the observed correlation coefficient in a sample of n pairs of observations from a bivariate normal population. Then

$$t = r\sqrt{(n-2)} / \sqrt{(1-r^2)}$$

Null Hypothesis H_0: $r = 0$ i.e. Population correlation coefficient is zero.

If $t_{cal} < t_{table}$ ACCEPT H_0

If $t_{cal} > t_{table}$ REJECT H_0

Example 1:

There are 18 pair of observations taken from a sample testing from a bivariate normal population, significant at 5% level of significance. Find out least value of significance.

Solution:

in is given as 18. The observed value of sample correlation coefficient r will be significant at 5% level of significance if

$$t = r\sqrt{(n-2)} / \sqrt{(1 – r^2)} > t_{n-2}(0.025)$$

From Table VIII, we have $t_{16}(0.025) = 2.12$

Substituting in (*), we get

$=> r\sqrt{(18-2)} / \sqrt{(1 – r^2)} > 2.12$

$=> 4r / \sqrt{(1 – r^2)} > 2.12$

Squaring and transposing, we get:

$=> 16 r^2 > (2.12)^2 (1 - r^2)$

$=> 16 r^2 > 4.4944 (1 - r^2)$

$=> (16 + 4.4944) r^2 > 4.4944$

$=> r > 4.4944 / 20.4944 = 0.4683$

Example 2: There are 625 pair of observations taken from a sample testing and a coefficient of correlation of 0.2 is derived.

Find out:

i. Whether the value is significant or no?

ii. What are the 95% and 99% confidence limits for the correlation coefficient in the population?

Solution:

Under the null hypothesis $H_0 : \rho = 0$ i.e. the value of $r = 0.2$ is not significant against $H_1 : \rho \neq 0$, the test statistics is :

$$t = r\sqrt{(n-2)} / \sqrt{(1 - r^2)} > t_{n-2}$$

$$= \{0.2 \text{ x } \sqrt{(625 - 2)}\} / \sqrt{(1 - 0.04)}$$

$$= (0.2 \text{ x } \sqrt{623}) / \sqrt{0.96}$$

$$= 5.09$$

d.f. = 625 – 2 = 623. Since d.f. is greater than 60, the significant value of t are same as in case of normal distribution, viz.,

$t_{0.05} = 1.96$ and $t_{0.01} = 2.58$ for two tailed test.

Since calculated t is much greater than these values, it is significant. Hence $H_0 = \rho$, is rejected and we conclude that the sample correlation is significant of correlation in the population.

95% Confidence limits for ρ (Population Correlation Cofficient):

$$R \pm 1.96 \text{ S.E. } (r) = r \pm 1.96 \text{ x } (1 - r^2) / \sqrt{n}$$

$$= 0.2 \pm 1.96 \text{ x } 0.96 / \sqrt{625}$$

$$= 0.2 \pm 1.96 \text{ x } 0.0384$$

$$= 0.2 \pm 0.075 = (0.125, 0.275)$$

99% Confidence limits for ρ

$$= 0.2 \pm 2.58 \text{ x } 0.0384 = 0.2 \pm 0.099$$

$$= (0.101, 0.299)$$

Example 3:

There are 27 pair of observations taken from a sample testing from a bivariate normal population and a coefficient of correlation of 0.42 is derived. Is it likely that the variables in the population are uncorrelated?

Solution:

n is given as 27. Sample correlation coefficient r = 0.42

$$t = r\sqrt{(n-2)} / \sqrt{(1 - r^2)} > t_{n-2}(0.025)$$

Null Hypothesis $H_0 : \rho = 0$ i.e. the variables are uncorrelated in the population.

Alternative Hypothesis $H_1 : \rho\ 0$ (Two - Tailed)

Test Statistic. Under H_0, the test statistic is

$$t = r\sqrt{(n-2)} / \sqrt{(1 - r^2)} \sim t_{n-2} = t_{25}$$

$$t = 0.42 \sqrt{25} / \sqrt{(1 - (0.42)^2)} = (0.42 \times 5) / 0.908$$

$$= 2.31$$

The tabulated value of t for 25 d.f. and 5% level of significance for a two tailed test is 2.06. Since calculated value of t is greater than the tabulated value, it is significant and we reject the null hypothesis (ρ=0) at 5% level of significance and we conclude that variables are correlated in the population.

10.7 Summary

There are different tests of significance for testing hypothesis. Z- test is normally used for large samples and t-test is used for small samples when σ is unkown. t- Test can be used to test significance of sample mean. It can also be used to test the significance of difference of two sample means. Further, it may be used to test the significance of observed sample correlation.

10.8 Key Words

- **Null Hypothesis:** A hypothesis of no difference is null hypothesis.
- **Alternate Hypothesis:** Hypothesis which is complementary to null hypothesis is alternate hypothesis.
- **Universe:** An aggregate of objects under study is called population or universe.
- **Sample:** finite subset of universe is called a sample.
- **Sample Size:** The number of members in a sample is called sample size.

10.9 Self Assessment Test

1. " Discovery of Student's 't' is regarded as a landmark in the history of statistics." Elaborate.
2. Explain the t-test for testing the significance of the difference between two sample means. Also state assumptions involved.
3. Explain the t-test for testing the significance of an observed sample correlation.
4. Prices of shares of TCS company on the different days were found to be

66 65 69 70 69 71 70 63 64 68

Discuss whether the mean price of the shares in the month is 65 at 5% level of significance.

[solution t = 2.825, Significant.]

5. A salesman is eypected to affect an average sales of Rs. 3500. Given a sample that a particular salesman had made

Rs. 3700 2500 3400 5200 3000 2000

To check whether his work is below standard or not at 5% level of significance.

[solution t = - 0.44, Not Significant.]

6. A random sample of 27 pairs of observations from a normal population gave a correlation coefficient of 0.6.Is this significant of correlation in the population.

[solution t = 3.75, Significant.]

7. A correlation coefficient between marks in two successive tests for 12 students gave a value of 0.65.Can it be regarded as significant at 1% level .

[solution t = 2.7046, Not Significant.]

8. Following data were obtained when sample of two types of tubes were tested.

	Type I	**Type I**
Sample Size	8	7
Sample Means	1234hrs	1036hrs
Sample S.D.	36hrs	40hrs

Is the difference in the means sufficient to warrant that TYPE I is superior to TYPE II.

9. A company is interested in knowing if there is difference in the average salary received by worker in two division. Accordingly sample of 12 workers in the first division and 10 workers in the second division are selected at random.worker's salaries are known to be normally distributed and the standard deviation are about the same. The table value of t at 5% level of significance for 20 d.f. is 2.086.

[solution t = 2.2, Significant.]

10.10 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

Unit – 11 : Chi-Square Test

Unit Structure:

11.0 Objectives
11.1 Introduction
11.2 Degree of Freedom
11.3 Chi-Square Test for Independence of Attribute
11.4 Chi-Square Test as a Test of Goodness of Fit
11.5 Conditions for Chi-Square Test
11.6 Yates Correction
11.7 Limitations
11.8 Analysis of Variance (Anova)
11.9 One-Way Anova
11.10 Two-Way Anova
11.11 Summary
11.12 Key Words
11.13 Self Assessment Test
11.14 References

11.0 Objectives

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

- Chi-square test for independence of attributes
- Chi-square test as goodness of fit
- Yates correction
- Analysis of Variance
- One-way ANOVA
- Two-way ANOVA

11.1 Introduction

The inductive inference may be termed as the logic of drawing statistically valid conclusions about the population characteristics on the basis of a sample drawn from it in a scientific manner. The sampling distribution of the statistics

$$t = t(x_1, x_2, \text{———————————}, x_n)$$

in its standard form viz.

$$Z = (t - E(t))/S.E.(t)$$

Is not normal and hence the normal test cannot be applied. In this case, i.e. for small samples, we apply the exact sample tests which depend on the magnitude of n e.g. Chi-square test.

Chi-Square Distribution

The square of a standard normal variable is called a Chi-square variate with 1 degree of freedom, abbreviated as d.f.. Thus if X is a random variable following normal distribution with mean μ and standard deviation σ, then $(X - \mu)/\sigma$ is a standard normal variate.

$(X - \mu)^2/\sigma^2$

Is a chi-square variate with 1 d.f.

If $X_1, X_2, X_3,$ __________ $, X_v$ are v independent random variables following normal distribution with mean μ_1, $\mu_2, \mu_3,$ __________ $, \mu_v$ and standard deviations $\sigma_1, \sigma_2, \sigma_3,$ ____________ $, \sigma_v$ respectively then the variate

$$\chi^2 = (X_1 - \mu_1)^2 / \sigma_1^2 + (X_2 - \mu_2)^2 / \sigma_2^2 + \text{———————} + (X_v - \mu_v)^2 / \sigma_v^2$$

$$= \sum_{i=1}^{\nu} (X_i - \mu_i)^2 / \sigma_i^2$$

which is the sum of the squares of v independent standard normal variates, follows Chi-square distribution with v d.f.

- **Chi-Square Test (X^2 Test)**

The quantity χ^2 describes the magnitude of discrepancy between theory and observation was given by Prof. Karl Pearson in 1900.

If $\chi^2 = 0$, the observed and expected frequency completely coincides. The greater the value of $\chi 2$, greater the discrepancy between observed and expected frequencies.

If O_i (i=1,2,__________,n) is a set of observed or experimental frequencies and E_i (i=1,2,__________,n) is the corresponding set of expected or theoretical or hypothetical frequencies then Karl Pearson's chi-square given by

$$X^2 = \sum_{i=1}^{n_1} (O_1 - \bar{E})^2 / E_1$$

With degree of freedom (d.f) $\nu = (n-1)$

11.2 Degree of Freedom

The number of degree of freedom is total number of observations less the number of independent constants imposed on the observations.

Degree of freedom=(Number of frequencies) – (Number of Independent constraints on them)

e.g. $\nu = n-k$

where k is the number of independent constants imposed in a set of data of n observations.

If we are given n frequencies ($O_1, O_2,$ ____________ $, O_n$) subject to the linear constraint

$$O = \sqrt{\ }\ E = \sqrt{\ }\ N$$

Then for the application of χ^2 - test $\nu = (n-1)$

Degree of Freedom for r X s Contingency Table

For $r \times s$ contingency table,

The total number of frequencies is $n = r \times s = rs$.

These n frequencies are subjected to the following linear constraints:

a) r rows totals (A_1), (A_2),————————————, (A_r) are fixed.

b) s columns totals (B_1), (B_2),————————————, (B_s) are fixed.

Thus, there are r + s constraints on the cell frequencies but these r + s constraints are not independent since we have:

$$= \sum_{i=1}^{s} (A_i) = \sum_{j=1} (B_i) = N$$

Thus the number of independent constraints is:

$$k = r+s-1$$

Hence, for r X s contingency table, d.f. are:

$$\begin{aligned} \nu &= n-k \\ &= rs - (r + s - 1) \\ &= rs - r - s + 1 \\ &= (r-1) \times (s-1) \end{aligned}$$

Hence, for 2 X 2 contingency table d.f= (2-1) X (2 -1) = 1

2 X 2 Contingency Table

Under the null hypothesis of independence of attributes, the values of χ^2 for the 2 X 2 contingency table

			Total
	a	**b**	**a+b**
	c	**d**	**c+d**
Total	**a+c**	**b+d**	**N=a+b+c+d**

is given by

$$\chi^2 = N(ad-bc)^2/((a+c)(b+d)(a+b)(c+d))$$

11.3 Chi-Square Test for Independence of Attribute

Let us consider two attributes A and B. A divided into r classes $A_1, A_2, A_3,$ ————————, A_r and B divided into s classes $B_1, B_2, B_3,$ ————————, B_s. Such a classification is known as manifold classification.

B\A	A_1	A_2	--------------	A_i	A_r	Total
B_1	$(A_1\ B_1)$	$(A_2\ B_1)$		$(A_i\ B_1)$	$(A_r\ B_1)$	B_1
B_2	$(A_1\ B_2)$	$(A_2\ B_2)$		$(A_i\ B_2)$	$(A_r\ B_2)$	B_2
|						|
|						|
B_j	$(A_1\ B_j)$	$(A_2\ Bj)$		$(A_i\ B_j)$	$(A_r\ B_j)$	B_j
|						|
|						|
B_s	$(A_1\ B_s)$	$(A_2\ B_s)$		$(A_i\ B_s)$	$(A_r\ B_s)$	B_s
Total	A_1	A_2		A_i	A_r	N

The problem is to test if two attributes A & B under consideration are independent or not.

Observed frequency= A_i / N . B_i /N   for i=1,2,——————,r & j=1,2,———————,s

Expected frequency = $A_i B_j$/N

Hypothesis (H_0): Two attributes are independent.

$\chi^2_{cal} < \chi^2_{table}$ Accept hypothesis

$\chi^2_{cal} > \chi^2_{table}$ Reject hypothesis

Value of χ^2 is seen from the table at 5% level of significance.

With the help of χ^2 -test, we can find whether or not to attributes are associated.

We take the null hypothesis that there is no association between attributes i.e. we assume that two attributes are independent.

a) If the calculated value of χ^2 is less than the table value at specified level of significance, the hypothesis holds good i.e. the attributes are independent.

b) If the calculated value of χ^2 is more than the table value then attributes are not independent.

Example 1:

1. In a survey of 200 girls of which 75 were intelligent, 40 had educated mothers, while 85 of the unintelligent girls had uneducated mothers. Do these figures support the hypothesis that educated mothers have intelligent girls? (Value of χ^2 for 1 d.f is 3.841).

Solution:

The given data can be organized in the form of a 2 X 2 contingency table given below

	Intelligent girls	**Un-intelligent girls**	**Total**
Educated mothers	40	125-85=40	40+40=80
Un-educated mothers	120-85= 35	85	200-80=120
Total	75	200-75=125	200

Null Hypothesis:

The Two attributes i.e. education of mothers and intelligence of girls are independent. It means education of mothers does not have any effect on the intelligence of the girls.

Expected frequencies are

E (40) = 75 X 80/200 = 30

E (35) = 75 X 120/200 = 45

E (40) = 125 X 80/200 = 50

E (85) = 125 X 120/200 = 75

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
40	30	10	100	3.33
35	45	-10	100	2.22
40	50	-10	100	2.00
85	75	10	100	1.33

$$\chi^2 = \sqrt{} [(O-E)^2/E]$$

$$= 3.33 + 2.22 + 2.0 + 1.33$$

$$= 8.88$$

d.f. = (2-1) X (2-1) = 1

$\chi^2_{table} = 3.841$ (Given)

$\chi^2_{cal} > \chi^2_{table}$ at 5% level of significance and hence null hypothesis is rejected. It means education of mothers has a significant effect on the intelligence of the girls.

Example 2:

Out of a sample of 120 persons in a village, 76 persons were administered a new drug for preventing malaria and out of them, 24 persons were attacked by malaria. Out of those who were not affected by malaria. Prepare a 2 X 2 table showing actual and expected frequencies and use chi-square test for finding out whether the new drug is effective or not. At 5% level for one degree of freedom, the value of chi-square is 3.84.

Solution:

The given data can be organized in the form of a 2 X 2 contingency table given below:

New drug	**Effect of Malaria Attacked**	**Effect of Malaria not attacked**	**Total**
Administered	**24**	**76-24=52**	**76**
Not administered	**44-12= 32**	**12**	**120-76=44**
Total	**24+32=56**	**52+12=64**	**120**

Null Hypothesis:

The Two attributes i.e. attack by malaria and administration of the new drug is independent. It means new drug is not effective in controlling malaria.

Expected frequencies are

E (24) = 56 X 76/120 = 35.47

E (32) = 56 X 44/120 = 20.53

E (52) = 64 X 76/120 = 40.53

E (12) = 44 X 64/120 = 23.47

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
40	35.47	-11.47	131.5609	3.709
35	20.53	11.47	131.5609	6.408
40	40.53	11.47	131.5609	3.246
85	23.47	-11.47	131.5609	5.605

$$\chi^2 = \sqrt{\ } [(O-E)^2/ E]$$

$$= 3.709 + 6.408 + 3.246 + 5.605$$

$$= 18.968$$

d.f. = (2-1) X (2-1) = 1

$\chi^2_{table} = 3.84$ (Given)

$\chi^2_{cal} > \chi^2_{table}$ at 5% level of significance and hence null hypothesis is rejected. It means that the new drug is definitely effective in controlling the malaria.

Example 3:

A sample of 400 students of post-graduate and 400 students of under-graduate classes was taken to know their opinion about autonomous colleges. 290 of the under-graduate and 310 of the post-graduate students favoured the autonomous status. Present these things in a table and use chi-square test at 5% level that the opinion regarding autonomous status of colleges is independent of the level of classes of students. Table value at 5% level is 3.84 for 1 d.f.

Solution:

The given data can be organized in the form of a 2 X 2 contingency table given below

Class	No. of students favouring	No. of students opposing	Total
Under Graduate	290	400-290=110	400
Post Graduate	310	400-310=90	400
Total	600	200	800

Null Hypothesis:

The Two attributes i.e. opinion about autonomous colleges is independent of level of classes.

Expected frequencies are
E (290) = 600 X 400/800 = 300
E (310) = 600 X 400/800 = 300
(110) = 200 X 400/800 = 100
E (90) =200 X 400/800 = 100

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
290	300	-10	100	0.33
310	300	10	100	0.33
110	100	10	100	1.00
90	100	-10	100	1.00

$$\chi^2 = \sqrt{\ } [(O-E)^2/ E]$$

$$= 0.33 + 0.33 + 1.0 + 1.0$$

$$= 2.66$$

d.f. = (2-1) X (2-1) = 1

$\chi^2_{table} = 3.841$ (Given)

$\chi^2_{cal} < \chi^2_{table}$ at 5% level of significance and hence it is not significant. Null hypothesis is accepted. It means opinion about autonomous colleges is independent of the level of classes.

Example 4:

Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligent levels. The results are as follows

Researcher	Below Average students	Average students	Above Average students	Genius students	Total
A	86	60	44	10	200
B	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different? Given χ^2 values for 3 d.f. and 4 d.f. are 7.82 and 9.49 at 5 % level of significance.

Solution:

Null Hypothesis:

The data obtained are independent of the sampling techniques adopted by the two researchers. It means there is no significant difference between the sampling techniques used by the two researchers for collecting the data.

Now we have a 4 X 2 contingency table and

d.f. = (4-1) X (2-1) = 3

Hence there is a need to compute only 3 expected frequencies and the remaining expected frequencies can be obtained by subtraction from marginal totals.

E (86) = 126 X 200/300 = 84

E (60) = 93 X 200/300 = 62

(44) = 69 X 200/300 = 46

EXPECTED FREQUENCIES

Researcher	Below Average students	Average students	Above Average students	Genius students	Total
A	84	62	46	200-192=8	200
B	126-84=42	93-62=31	69-46=23	12-8=4	100
Total	126	93	69	12	300

O	E	O-E	$(O-E)^2$	$(O-E)^2/E$
86	84	2	4	0.048
60	62	-2	4	0.064
44	46	-2	4	0.087
10	8	2	4	0.50
40	42	-2	4	0.095
33	31	2	4	0.129
25	23	2	4	0.174
2	4	-2	4	1.000

Therefore $\chi^2 = \sqrt{\ } (O-E)^2/E = 2.097$

d.f. = (4-1) X (2-1) = 3

χ^2_{table} for 3 d.f. = 7.82 (Given)

$\chi^2_{cal} < \chi^2_{table}$ at 5% level of significance and hence it is not significant. Null hypothesis is accepted. It means sampling techniques adopted by two researchers do not differ significantly.

11.4 Chi-Square Test as a Test of Goodness of Fit

χ^2 test enables us to decide how well the theoretical observations such as binomial, poisson, normal fit observed distribution

a) $\chi^2_{cal} > \chi^2_{table}$ Reject hypothesis

b) $\chi^2_{cal} < \chi^2_{table}$ Accept hypothesis

From table, values are seen at 5% level of significance.

If calculated value of $\div^2$ is less than table value at the specified level of significance, the fit is considered to be good i.e. the divergence between actual and expected frequencies is very less. If calculated value of χ^2 is greater than table value at the specified level of significance, the fit is considered to be poor i.e. the divergence between actual and expected frequencies is very large.

Note: - Number of theoretical and expected frequency should not be less than 10.

For example

232	190
94	87
18	27
5	10

Then, we take

232	190
94	87
23	37

Example 1:

The number of car accidents per month in a certain town were as follows:

12
8
20
2
14
10
15
6
9
4

Are these frequencies in agreement with the belief that accident conditions were same during the 10-month period?

Solution:

Null Hypothesis:

The given frequencies of number of accidents per month in a certain town are consistent with the belief that the accident conditions were same during the 10-month period.

Since the total number of accidents over the 10 months are:

12+8+20+2+14+10+15+6+9+4=100.

These accidents should be uniformly distributed over the 10-month period and hence

The expected number of accidents for each of the 10 months are 100/10 = 10

Month	Observed No. Of accidents	Expected No. Of accidents	(O-E)	$(O-E)^2$	$(O-E)^2/E$
1	12	10	2	4	0.4
2	8	10	-2	4	0.4
3	20	10	10	100	10.0
4	2	10	-8	64	6.4
5	14	10	4	16	1.6
6	10	10	0	0	0.0
7	15	10	5	25	2.5
8	6	10	-4	16	1.6
9	9	10	-1	1	0.1
10	4	10	-6	36	3.6
Total	**100**	**100**	**0**		**26.6**

$\chi^2 = \sqrt{}\ (O-E)^2/E = 26.6$

d.f. = (10-1) = 9

χ^2_{table} = 16.919 for 9 d.f. at 5 % level of significance.

Since $\chi 2_{cal} > \chi^2_{table,}$ it is significant and hypothesis is rejected. Hence accident conditions are not uniform over the 10-month period.

Example 2

A die is rolled 100 times with the following distribution

Number	**1**	**2**	**3**	**4**	**5**	**6**
Observed	**17**	**14**	**20**	**17**	**17**	**15**

At 1 % level of significance, determine whether the dice is uniform.

Solution:

Null Hypothesis: The dice is uniform.

N= 17+14+20+17+17+15 = 100

Probability of getting each face = 1/6

Expected Frequency for each face = 100/6 = 16.67

Number	**Observed Frequency(O)**	**Expected Frequency**	**(O-E)**	**$(O-E)^2$**	**$(O-E)^2/E$**
1	17	16.67	0.33	0.1089	0.0065
2	14	16.67	-2.67	7.1289	0.4276
3	20	16.67	3.33	11.0889	0.6652
4	17	16.67	0.33	0.1089	0.0065
5	=17	16.67	0.33	0.1089	0.0065
6	15	16.67	-1.67	2.7889	0.1673
Total	**100**	**100**	**0**		**1.2796**

Degree of freedom ν = 6-1 = 5

χ^2_{table} = 15.086 for 5 d.f. at 1 % level of significance.

Since $\chi 2_{cal} < \chi^2_{table,}$ it is not significant and hypothesis is accepted. Hence dice may be considered as uniform.

Example 3:

Digit	**Frequency**
0	1026
1	1107
2	997
3	966
4	1075
5	933
6	1107
7	972
8	964
9	853
Total	**10000**

The above table show the distribution of digits in numbers chosen at random from a directory. Test whether the digits occur equally frequently in the directory. Table value of for 9 degree of freedom at 5% level of significance is 16.92.

Solution:

Null Hypothesis: Digits occur equally frequently in the directory.

The expected frequency for each of digit is 10000/10 = 1000

Digits	Observed Frequency (O)	Expected Frequency (E)	(O-E)	$(O-E)^2$	$(O-E)^2/E$
0	1026	1000	26	676	0.676
1	1107	1000	107	11449	11.449
2	997	1000	-3	9	0.009
3	966	1000	34	1156	1.156
4	1075	1000	75	5625	5.625
5	933	1000	-67	4489	4.489
6	1107	1000	107	11449	11.449
7	972	1000	-28	784	0.784
8	964	1000	-36	1296	1.296
9	853	1000	-147	21609	21.609
Total	**100**	**10000**	**0**		**58.542**

$\chi^2 = \sqrt{\ }(O\text{-}E)^2/E = 58.542$

d.f. = (10-1) = 9

$\chi^2_{table} = 16.919$ for 9 d.f. at 5 % level of significance.

Since $\chi^2_{cal} > \chi^2_{table,}$ it is significant and hypothesis is rejected. Hence we conclude that digits cannot be considered to be distributed uniformly in the directory.

11.5 Conditions for Chi-Square Test

The Chi-square test can be used only if the following conditions are satisfied:

1. Total frequency N should be reasonably large i.e. greater than 50.
2. Observations of the sample should be independent.
3. The constraints on the cell frequencies, if any, should be linear such as

$$\sqrt{\ }O = \sqrt{\ }E = N$$

4. No theoretical frequency should be small. Preferably, each theoretical frequency should be larger than 10 but in any case not less than 5. If any theoretical frequency is less than 5 , then we cannot apply χ^2 – test as such. In that case, we use the technique of pooling which consists in adding the frequencies which are less than 5 with the preceding or succeeding frequency so that the resulting sum is greater than 5 and adjust for the degree of freedom accordingly.
5. The given distribution should not be replaced by relative frequencies but the data should be given in original units.

11.6 Yates Correction

If any cell frequency in 2 X 2 table is less than 5, then for application of χ^2-test, it has to be pooled with the preceding or succeeding frequency so that total is greater than 5. This results in the loss of 1 d.f.. Since for 2 X 2 table,

d.f. = (2-1) X (2-1) = 1

The degree of freedom left after adjusting for pooling are

ν = 1-1= 0 which is absurd.

In such a situation when any cell frequency < 5, we apply the correction given by F. Yates in 1934 and popularly known as the Yates correction for continuity. This consists in adding 0.5 to the cell frequency which is < 5 and adjusting the remaining frequencies accordingly since row and column totals are fixed and then applying χ^2- test without pooling.

After applying Yates's correction, the value of χ^2 is given by

$$\chi^2 = [|ad-bc|-N/2]^2 / (a+c)(b+d)(a+b)(c+d).$$

Example

The following information is obtained concerning an investigation of 50 ordinary shops of small size:

	Shops in Towns	Shops in villages	Total
Run by men	17	18	35
Run by women	3	12	15
	20	30	50

Can it be inferred that shops run by women are relatively more in villages than in towns. Use χ^2 test.

Null Hypothesis:

There, two is no difference so far as shops run by men and women in towns and villages.

Expected frequencies are

E (17) = 35 X 20/50 = 14

E (18) = 35 X 30/50 = 21

E (3) = 15 X 20/50 = 6

E (12) = 15 X 30/50 = 9

O	**E**	**O-E**	**$(O-E)^2$**	**$(O-E)^2/E$**
17	**14**	**3**	**9**	**0.64**
18	**21**	**-3**	**9**	**0.43**
3	**6**	**-3**	**9**	**1.50**
12	**9**	**3**	**9**	**1.00**

$$\chi^2 = \sqrt{\ }\ [(O-E)^2/E]$$

$$= 0.64 + 0.43 + 1.50 + 1.00$$

$$= 3.57$$

As one cell frequency is only 3 in the above table, Now applying Yate's correction, we have

$$\chi^2(\text{corrected}) = [|17-14|-0.5]^2/14 + [|18-21|-0.5]^2/21 + [|3-6|-0.5]^2/6 + [|12-9|-0.5]^2/9$$

$$= (2.5)^2/14 + (2.5)^2/21 + (2.5)^2/6 + (2.5)^2/9$$

$$= 0.446 + 0.298 + 1.04\ 0 + 0.694$$

$$= 2.478$$

d.f. = (2-1) X (2-1) = 1

$\chi^2_{table} = 3.841$ (Given)

$\chi^2_{cal} < \chi^2_{table}$ at 5% level of significance and hence null hypothesis is accepted. It means there is no difference between shops run by men and women in villages and towns.

11.7 Limitations

The chi-square test is no doubt a most frequently used test but its correct application is equally an uphill task. There are certain limitations of chi-square test

1. This test cannot be applied if occurrence of one individual observation has effect upon the occurrence of any other observation in the sample under consideration.
2. For small theoretical frequencies, we have to apply Yate's correction.
3. It does not take into consideration frequencies of non-occurrence.
4. Computation procedure is complex.

11.8 Analysis of Variance (Anova)

ANOVA is an extremely useful technique concerning researches in the fields of economics, sociology, business, biology, education and in researches of several other disciplines. This technique is important in all those situations where we want to compare more than two populations such as in comparing the yield of crops from several varieties of seeds, the drinking habits of ten groups of university students and so on.

Prof. R.A. Fisher was the first man to use the term variance who developed a very elaborate theory concerning ANOVA, explaining its usefulness in practical field. After some time, professor Snedecor and many others contributed to the development of this technique. The essence of ANOVA is that the total amount of variation in a set of data is divided into two types i.e. the amount which can be attributed to chance and that amount which can be attributed to specified causes. There may be variation between samples and also within sample items.

Thus by using ANOVA technique, we can investigate any number of factors which influence the dependent variable. If we take only one factor and investigate the differences amongst its various categories having numerous possible values then it is one-way ANOVA. If we take two factors at the same time, then it is two-way ANOVA

11.9 One-Way Anova

The technique involves the following steps

a) Obtain the mean of each sample i.e. $X_1, X_2, X_3, X_4, \_\_\_\_\_\_\_\_\_\_\_\_, X_k$ when there are k samples.

b) Obtain the mean of the sample means as follow.

$$\bar{X} = (X_1 + X_2 + X_3 + \text{————————} + X_k)/k$$

c) Obtain the sum of squares for variance between the samples.

$$\text{SS between samples} = n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2 + \text{————————} + n_k(\bar{X}_k - \bar{X})^2$$

d) Obtain the mean square between samples by dividing result of step (c) by degree of freedom.

MS between samples = SS between samples/(k-1) where (k-1) represents degree of freedom between samples.

e) Obtain the sum of squares for variance with in samples.

$$SS \text{ within samples} = \sqrt{}(X_{1i} - \bar{X}_1)^2 + \sqrt{}(X_{2i} - \bar{X}_2)^2 + \text{———————} + \sqrt{}(X_{ki} - \bar{X}_k)^2$$

$$i = 1, 2, 3, \text{————}$$

f) Obtain the mean square within samples by dividing result of step (e) by degree of freedom.

MS within samples = SS within samples/ (n-k)

Where (n-k) represents degree of freedom within samples.

$n = n_1 + n_2 + n_3 + \text{————————} + n_k$

k = number of samples

g) Sum of squares of deviations for total variances is

SS for total variance = SS between samples + SS within samples

h) F-ratio = MS between samples/MS within samples

This ratio is used to check whether the difference among several means is significant. For this, we will get value of F for given degree of freedom at different level of significance.

If $F_{cal} < F_{table}$ then the difference is taken as insignificant and

If $F_{cal} > F_{table}$ then the difference is taken as significant.

11.9.1 Analysis of Variance Table for One-Way Anova

Source of variation	Sum of squares (SS)	Degree of freedom (d.f.)	Mean Square (MS)	F-ratio
Between samples	$n_1(\overline{X_1} - \overline{X})^2$ + ------------+ $n_k(\overline{X_k} - \overline{X})^2$	k-1	SS between samples/(k-1)	MS between samples/MS within samples
Within samples	$\sum(X_{1i} - \overline{X_1})^2$ + --------------- + $\sum(X_{ki} - \overline{X_k})^2$ i = 1, 2, 3, --------	n-k	SS within samples/(n-k)	
Total	$\sum(X_{ij} - \overline{X})^2$ i = 1, 2, 3, -------- j = 1, 2, 3, --------	n-1		

11.9.2 Short-Cut Method for One-Way Anova

Steps involved in short-cut method are given below

a) Compute $T = \sqrt{\;} X_{ij}$

b) Compute Correction factor = $(T)^2/n$

c) Find sum of squares for total variance

$$\text{Total SS} = \sqrt{\;} X_{ij}^2 - (T)^2/n \qquad i= 1, 2, \text{———} \qquad j=1, 2, \text{———}$$

d) Obtain the sum of squares for variance between the samples.

$$\text{SS between samples} = \sqrt{\;}\ (T_j)^2/n_j - (T)^2/n \qquad j=1, 2, \text{———}$$

e) Obtain the sum of squares within samples.

$$\text{SS within samples} = [\sqrt{\;} X_{ij}^2 - (T)^2/n] - [\sqrt{\;}\ (T_j)^2/n_j - (T)^2/n]$$

After doing all this, we can set the table as described above.

Example

Prepare an analysis of variance table for the following per acre production data for three varieties of corn, each grown on 4 plots and state if the variety differences are significant.

Plot of land	**Per acre production data**		
	Variety of corn		
	A	**B**	**C**
1	**6**	**5**	**5**
2	**7**	**5**	**4**
3	**3**	**3**	**3**
4	**8**	**7**	**4**

Solution:

Null Hypothesis: There is no difference between sample means.

First we calculate mean of each sample

$\bar{X} = (6 + 7 + 3 + 8)/4 = 6$

$\bar{X} = (5 + 5 + 3 + 7)/4 = 5$

$\bar{X} = (5 + 4 + 3 + 4)/4 = 4$

Mean of the sample means $\bar{X} = (\bar{X}_1 + \bar{X}_2 + \bar{X}_3)/k$

$= (6 + 5 + 4)/3$

$= 5$

SS between samples = $n_1(\bar{X}_1 - \bar{X})^2 + n_2(\bar{X}_2 - \bar{X})^2 +$ ———————— $+ n_k(\bar{X}_k - \bar{X})^2$

$= 4(6\text{-}5)^2 + 4(5\text{-}5)^2 + 4(4\text{-}5)^2$

$= 4 + 0 + 4$

$= 8$

SS within samples = $\sqrt{\ }(X_{1i} - X_1)^2 + \sqrt{\ }(X_{2i} - X_2)^2 +$ ———————— $+ \sqrt{\ }(X_{ki} - X_k)^2$

i = 1, 2, 3, 4

$= [(6\text{-}6)^2 + (7\text{-}6)^2 + (3\text{-}6)^2 + (8\text{-}6)^2]$

$+[(5\text{-}5)^2 + (5\text{-}5)^2 + (3\text{-}5)^2 + (7\text{-}5)^2]$

$+[(5\text{-}4)^2 + (4\text{-}4)^2 + (3\text{-}4)^2 + (4\text{-}4)^2]$

$= 14 + 8 + 2$

$= 24$

SS for total variance = SS between samples + SS within sample

$= 8 + 24 = 32$

Source of variation	**SS**	**d.f.**	**MS**	**F-ratio**	**F_{table} at 5% level**
Between samples	**8**	**(3-1) = 2**	**8/2 = 4**	**4/2.67 =1.5**	**F(2,9) = 4.26**
Within samples	**24**	**(12-3)= 9**	**24/9 = 2.67**		
Total	**32**	**(12-1) = 11**			

Now $F_{cal} < F_{table}$ then the difference is taken as insignificant i.e. there is no difference in corn output due to varieties.

Short-cut method

T = 60

n = 12

Correction factor = $(T)^2/n = 60x60/12 = 300$

Total SS = $\sqrt{\ } X_{ij}^2 - (T)^2/n$ i= 1, 2, ———— j=1, 2, ————

$= (6)^2 + (7)^2 + (3)^2 + (8)^2 + (5)^2 + (5)^2 + (3)^2 + (7)^2 + (5)^2 + (4)^2$

$+ (3)^2 + (4)^2 - 300$

$= 332 – 300 = 32$

SS between samples = $\sqrt{\ }(T_j)^2/n_j - (T)^2/n$ j=1, 2, ————

$= (24x24)/4 + (20x20)/4 + (16x16)/4 – 300$

$= 144 + 100 + 64 – 300$

$= 8$

SS within samples = $[\sqrt{\ } X_{ij}^2 - (T)^2/n] – [\sqrt{\ }(T_j)^2/n_j - (T)^2/n]$

$= 32 – 8 = 24$

Same result as we got from direct method.

11.10 Two-Way Anova

This technique is used when the data are classified on the basis of two factors. For example, the agricultural output may be classified on the basis of different varieties of seeds and also on the basis of different varieties of fertilizers used.

Steps involved are given below:

a) Compute $T = \sqrt{\ } X_{ij}$

b) And Compute Correction factor = $(T)^2/n$

c) Find sum of squares for total variance

$$\text{Total SS} = \sqrt{\ } X_{ij}^2 - (T)^2/n \qquad i= 1, 2, \text{———} \qquad j=1, 2, \text{———}$$

d) Take the total of different columns and then obtain the square of each column total and divide such squared values of each column by the number of items in the concerning column and take the total of result thus obtained and subtract correction factor to obtain SS between columns.

e) Take the total of different rows and then obtain the square of each row total and divide such squared values of each row by the number of items in the concerning row and take the total of result thus obtained and subtract correction factor to obtain SS between rows.

f) SS for error = Total SS – (SS between columns + SS between rows)

g) Degree of freedom for total variance = (c.r -1)
Degree of freedom for variance between columns = (c -1)
Degree of freedom for variance between rows = (r -1)
Degree of freedom for residual variance = (c-1) (r -1)

11.10.1 Analysis of Variance Table for Two-Way Anova

	Sum of squares (SS)	Degree of freedom (d.f.)	Mean Square (MS)	F-ratio
Between columns	$\sum (T_j)^2/n_j - (T)^2/n$	(c-1)	SS between columns/(c-1)	MS between columns/MS residual
Between rows	$\sum (T_i)^2/n_i - (T)^2/n$	(r-1)	SS between rows/(r-1)	MS between rows/ MS residual
Error or Residual	Total SS – (SS between columns + SS between rows)	(c-1)(r-1)	SS residual/ (c-1)(r-1)	
Total	$\mathbf{\sum X_{ij}^2 - (T)^2/n}$	**(c.r-1)**		

Both the F-ratios are compared with their corresponding table values for given degree of freedom at a specified level of significance.

If $F_{cal} < F_{table}$ then the difference is taken as insignificant and

If $F_{cal} > F_{table}$ then the difference is taken as significant.

Example:

A farmer applies three types of fertilizers on four separate plots. The table on yield per acre are given below

Fertilizers	**Yield**				
	A	**B**	**C**	**D**	**E**
Nitrogen	**6**	**4**	**8**	**6**	**24**
Potassium	**7**	**6**	**6**	**9**	**28**
Phosphates	**8**	**5**	**10**	**9**	**32**
Total	**21**	**15**	**24**	**24**	**84**

Find out if plots are materially different in fertility as also if the three fertilizers make any material difference in yield.

Solution:

Null Hypothesis:

a) H_0 = The four plots are equally fertile.

b) H_0 = There is no difference among three fertilizers i.e. all fertilizers are equally effective.

Let X_{ij} denote the yield from the jth plot receiving ith fertilizer, i=1, 2, ——; j=1, 2, ——

Number of rows r =3

Number of columns = 4

n = c.r = 3x4 = 12

$T = \sqrt{\ } X_{ij =} 84$

Correction factor = $(T)^2/n$ = 84x84/12 = 588

Sum of squares for total variance

Total SS = $\sqrt{\ } X_{ij}^2 - (T)^2/n$

= (36+16+64+36) + (49+36+36+81) + (64+25+100+81) -588

= 36

SS between columns (plots) = $C_1^2 + C_2^2 + C_3^2 + C_4^2)/r$ – corrector factor

= (441 + 225 + 576 + 576)/3 – 588

= 18

SS between rows (fertilizers) $= R_1^2 + R_2^2 + R_3)/c$ – corrector factor

$= (576 + 784 + 1024)/4 – 588$

$= 8$

SS for error = Total SS – (SS between columns + SS between rows)

$= 36 – 18 – 8$

$= 10$

h) Degree of freedom for total variance $= (3.4 -1) = 11$

Degree of freedom for variance between columns $= (4 -1) = 3$

Degree of freedom for variance between rows $= (3 -1) = 2$

Degree of freedom for residual variance $= (4-1)(3 -1) = 6$

Sources of variation	d.f.	SS	MSS = SS/d.f.	Variance Ratio (F)
Between columns(plots)	**3**	**18**	**6**	**6/1.67 =3.593**
Between rows (fertilizers)	**2**	**8**	**4**	**4/1.67 = 2.395**
Error	**6**	**10**	**1.67**	
Total	**11**	**36**		

$F_{table}(3,6) = 4.76$ and

$F_{table}(2,6) = 5.14$ at 5% level of significance.

a) $F_{cal}(3.593) < F_{table}(4.76)$, we fail to reject H_0 and conclude that the four plots are equally fertile.

b) $F_{cal}(2.395) < F_{table}(5.14)$, we fail to reject H_0 and conclude that the different fertilizers are considered to be equally effective.

11.11 Summary

The chi-square test is an important test amongst the several tests of significance developed by statisticians. Chi-square written as x^2 is a statistical measure used in the context of sampling analysis for comparing a variance to a theoretical variance. Thus the chi-square test is applicable in large number of problems. The test is a technique through the use of which it is possible to: a) test the goodness of fit b)test for independence of attributes. F.Yate's suggested a correction for continuity in x^2 particularly when cell frequencies are small. The rule for correction is to adjust the observed frequency in each cell of a table in such a way as to reduce the deviation of the observed from the expected frequency for that cell by 0.5. The ANOVA technique is extremely useful technique in various fields. The ANOVA technique is important in those situations where we want to compare more than two populations such as comparing the yield of crop from several varieties of seeds, the smoking habits of five groups of college students etc. In such circumstances

one does not want to consider all possible combinations of two populations at a time for that would require a great number of tests. This would consume lot of time and money. Therefore we often use ANOVA techniques and investigate the differences among the means of all the populations simultaneously.

11.12 Key Words

- **Chi-square Test**: A statistical test for large sized data used to test hypothesis.
- **Degree of Freedom:** The number of degree of freedom is total number of observations less the number of independent constants imposed on the observations.
- **ANOVA:** Analysis of variance.
- **Null Hypothesis:** Hypothesis of no difference.

11.13 Self Assessment Test

1. What is x^2-test of goodness of fit? What cautions are necessary while applying this test.
2. What is a contingency table? Describe how the x^2- distribution may be used to test whether the criteria of classification in an $m \times n$ contingency table are independent.
3. What do you understand by Yate's correction? Under what circumstances, it must be applied? Give formula for x^2 with & without Yates' correction.
4. What is two-way classification with one observation per cell in ANOVA. Describe the various steps in carrying out the analysis of variance of a two-way classified data with one observation per cell.
5. In a set of random numbers, the digits 0,1,2,————,9 were found to have the following frequencies:

Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Frequency	**43**	**32**	**38**	**27**	**38**	**52**	**36**	**31**	**39**	**24**

Test whether they are significantly different from those expected on the hypothesis of uniform distribution. Given $x^2_{0.05}$ for 9 d.f. = 16.92)

(Ans. $x^2 = 16.33$(Not significant))

6. Complete the following ANOVA table and test the hypothesis of homogeneity of

a) Blocks

b) Treatments

Sources of Variation	**Sum of Squares**	**d.f.**	**Mean S.S.**	**Variance Ratio (F)**
Blocks	**26.8**	**4**	**-------**	**------**
Treatment	**-------**	**3**	**-------**	**------**
Error			**2.5**	
Total	**85.3**			

(Ans. M.S.S(Blocks) = 6.7 , F(Blocks)=2.68 Not significant

M.S.S(Treatments)= 9.5 , F(Treatments)= 3.8 Significant)

11.14 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

Unit - 12 : Qualitative Research

Unit Structure:

12.0 Objectives

12.1 Introduction

12.2 Content Analysis

12.3 Focused Groups

12.4 Field Surveys

12.5 Application of Qualitative Research

12.6 Attitude Study

12.7 Instruments of Data Collection

12.8 Summary

12.9 Key Words

12.10 Self Assessment Test

12.11 References

12.0 Objectives

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

- The meaning of Qualitative Research
- The need of Qualitative Research
- Techniques of Content Analysis
- Qualitative Research through Focused group and Field surveys
- Application of Qualitative Research in Business
- Questionnaire and other instruments used for Qualitative Research

12.1 Introduction

The meaning of the term quantitative is fairly self-evident and refers to the tradition of research dominant in science since the 17th century, with its emphasis on the measurement and quantification of phenomena as essential steps in the process of enquiry. It will be seen that this emphasis on measurement is also linked to a particular set of philosophical assumptions about the nature of the world and how it works, as well as the understanding of it. Qualitative research is generally presented as an opposing category to quantitative research, but the term itself, however, is rarely explained; if quantitative research is about quantities, what qualities is qualitative research concerned with? Qualitative research is concerned with the quality or nature of human experiences and what these phenomena mean to individuals.

Qualitative research thus tends to start with 'what', 'how' and 'why' type questions rather than 'how much' or 'how many' questions. It is also concerned with examining these questions in the context of everyday life and each individual's meanings and explanations. Qualitative research can thus be broadly described as interpretative and naturalistic, in that it seeks to understand and explain beliefs and behaviours within the context that they occur. Beyond this definition, however, it is important to point out that within the broad tradition of qualitative research there are a number of theoretical orientations. Many, but not all, these orientations derive from the social sciences, and in particular sociology and anthropology, but qualitative research methods are now being used in a wide range of fields and disciplines.

Qualitative research is a type of scientific research. In general terms, scientific research consists of an investigation that: seeks answers to a question, systematically uses a predefined set of procedures to answer the question, collects evidence, produces findings that were not determined in advance, produces findings that are applicable beyond the immediate boundaries of the study

Qualitative research shares these characteristics. Additionally, it seeks to understand a given research problem or topic from the perspectives of the local population it involves. Qualitative research is especially effective in obtaining culturally specific information about the values, opinions, behaviors, and social contexts of particular populations.

The strength of qualitative research is its ability to provide complex textual descriptions of how people experience a given research issue. It provides information about the "human" side of an issue – that is, the often contradictory behaviors, beliefs, opinions, emotions, and relationships of individuals. Qualitative methods are also effective in identifying intangible factors, such as social norms, socioeconomic status, gender roles, ethnicity, and religion, whose role in the research issue may not be readily apparent. When used along with quantitative methods, qualitative research can help us to interpret and better understand the complex reality of a given situation and the implications of quantitative data.

Although findings from qualitative data can often be extended to people with characteristics similar to those in the study population, gaining a rich and complex understanding of a specific social context or phenomenon typically takes precedence over eliciting data that can be generalized to other geographical areas or populations. In this sense, qualitative research differs slightly from scientific research in general.

Quantitative and qualitative research methods differ primarily in: their analytical objectives, the types of questions they pose, the types of data collection instruments they use, the forms of data they produce, the degree of flexibility built into study design

Table – 1 : Comparison of Quantitative and Qualitative Research Approaches

	Quantitative Approaches	**Qualitative Approaches**
General framework	Seek to confirm hypotheses about phenomena	Seek to explore phenomena
	Instruments use more rigid style of eliciting and categorizing responses to questions	Instruments use more flexible, iterative style of eliciting and categorizing responses to questions
	Use highly structured methods such as questionnaires, surveys, and structured observation	Use semi-structured methods such as in-depth interviews, focus groups, and participant observation
Analytical objectives	To quantify variation	To describe variation
	To predict causal relationships	To describe and explain relationships
	To describe characteristics of a population	To describe individual experiences To describe group norms
Question format	Closed-ended	Open-ended
Data format	Numerical (obtained by assigning numerical values to responses)	Textual (obtained from audiotapes, videotapes, and field notes)
Flexibility in study design	Study design is stable from beginning to end	Some aspects of the study are flexible (for example, the addition, exclusion, or wording of particular interview questions)
	Participant responses do not influence or determine how and which questions researchers ask next	Participant responses affect how and which questions researchers ask next
	Study design is subject to statistical assumptions and conditions	Study design is iterative, that is, data collection and research questions are adjusted according to what is learned

The origin and uses of qualitative research given that there are a number of different theoretical orientations within the broad tradition of qualitative research, there are various possible starting points that can be identified corresponding to each of these orientations. Having trained originally in anthropology it is being shamelessly partisan to identify Malinowski as a key figure in developing the methodology of qualitative research, and in particular the research technique known as participant observation. He wrote the following famous statement in the introduction to his ethnography Argonauts of the Western Pacific first published in 1922 (Malinowski, 1962): '. . . the final goal, of which an Ethnographer should never lose sight . . . is, briefly, to grasp the native's point of view, his relation to life, to realize his vision of his world. We have to study man, and we must study what concerns him most intimately, that is the hold which life has on him. In each culture, the values are slightly different; people aspire after different aims, follow different impulses, yearn after a different form of happiness. In each culture, we find different institutions in which man pursues his life-interest, different customs by which he satisfies his aspirations, different codes of law and morality which reward his virtues or punish his defections. To study the institutions, customs, and codes or to study the behaviour and mentality without the subjective desire of feeling by what these people live, of realizing the substance of their happiness is, in my opinion, to miss the greatest reward which we can hope to obtain from the study of man.'

12.2 Content Analysis

Content analysis is a widely used qualitative research technique. Rather than being a single method, current applications of content analysis show three distinct approaches: conventional, directed, or summative. All three approaches are used to interpret meaning from the content of text data and, hence, adhere to the naturalistic paradigm. The major differences among the approaches are coding schemes, origins of codes, and threats to trustworthiness. In conventional content analysis, coding categories are derived directly from the text data. With a directed approach, analysis starts with a theory or relevant research findings as guidance for initial codes. A summative content analysis involves counting and comparisons, usually of keywords or content, followed by the interpretation of the underlying context. The authors delineate analytic procedures specific to each approach and techniques addressing trustworthiness with hypothetical examples drawn from the area of end-of-life care.

Content analysis has a long history in research, dating back to the 18th century in Scandinavia. In the United States, content analysis was first used as an analytic technique at the beginning of the 20th century. Initially, researchers used content analysis as either a qualitative or quantitative method in their studies. Later, content analysis was used primarily as a quantitative research method, with text data coded into explicit categories and then described using statistics. This approach is sometimes referred to as quantitative analysis of qualitative data. More recently, the potential of content analysis as a method of qualitative analysis for health researchers has been recognized, leading to its increased application and popularity (Nandy & Sarvela, 1997).

Qualitative content analysis is one of numerous research methods used to analyze text data. Other methods include ethnography, grounded theory, phenomenology, and historical research. Research using qualitative content analysis focuses on the characteristics of language as communication with attention to the content or contextual meaning of the text. Text data might be in verbal, print, or electronic form and might have been obtained from narrative responses, open-ended survey questions, interviews, focus groups, observations, or print media such as articles, books, or manuals. Qualitative content analysis goes beyond merely counting words to examining language intensely for the purpose of classifying large amounts of text into an efficient number of categories that represent similar meanings. These categories can represent either explicit communication or inferred communication. The goal of content analysis is "to provide knowledge and understanding of the phenomenon under study". Therefore, qualitative content analysis is defined as a research method for the subjective interpretation of the content of text data through the systematic classification process of coding and identifying themes or patterns.

12.2.1 Uses of content analysis

Ole Holsti (1969) groups uses of content analysis into three basic categories:

- make inferences about the antecedents of a communication
- describe and make inferences about characteristics of a communication
- make inferences about the effects of a communication.

He also places these uses into the context of the basic communication paradigm.

The following table (Table-2)shows uses of content analysis in terms of their general purpose, element of the communication paradigm to which they apply, and the general question they are intended to answer.

Table - 2

Purpose	Element	Question	Use
Make inferences about the antecedents of communications	Source	Who?	• Answer questions of disputed authorship (authorship analysis)
	Encoding process	Why?	• Secure political & military intelligence • Analyse traits of individuals • Infer cultural aspects & change • Provide legal & evaluative evidence
Describe & make inferences about the characteristics of communications	Channel	How?	• Analyse techniques of persuasion • Analyse style
	Message	What?	• Describe trends in communication content • Relate known characteristics of sources to messages they produce • Compare communication content to standards
	Recipient	To whom?	• Relate known characteristics of audiences to messages produced for them • Describe patterns of communication
Make inferences about the consequences of communications	Decoding process	With what effect?	• Measure readability • Analyse the flow of information • Assess responses to communications

12.2.2 Process of a Content Analysis

According to Dr. Klaus Krippendorff six questions must be addressed in every content analysis:

1. Which data are analysed?
2. How are they defined?
3. What is the population from which they are drawn?
4. What is the context relative to which the data are analysed?
5. What are the boundaries of the analysis?
6. What is the target of the inferences?

Qualitatively, content analysis can involve any kind of analysis where communication content (speech, written text, interviews, images ...) is categorised and classified. In its beginnings, using the first newspapers at the end of 19th century, analysis was done manually by measuring the number of lines and amount of space given a subject. With the rise of common computing facilities like PCs, computer-based methods of analysis are growing in popularity. Answers to open ended questions, newspaper articles, political party manifestoes, medical records or systematic observations in experiments can all be subject to systematic analysis of textual data. By having contents of communication available in form of machine readable texts, the input is analysed for frequencies and coded into categories for building up inferences. Robert Philip Weber (1990) notes: "To make valid inferences from the text, it is important that the classification procedure be reliable in the sense of being consistent: Different people should code the same text in the same way". The validity, inter-coder reliability and intra-coder reliability are subject to intense methodological research efforts over long years.

One more distinction is between the manifest contents (of communication) and its latent meaning. "Manifest" describes what (an author or speaker) definitely has written, while latent meaning describes what an author intended to say/write. Normally, content analysis can only be applied on manifest content; that is, the words, sentences, or texts themselves, rather than their meanings. Dermot McKeone (1995) has highlighted the difference between prescriptive analysis and open analysis. In prescriptive analysis, the context is a closely defined set of communication parameters (e.g. specific messages, subject matter); open analysis identifies the dominant messages and subject matter within the text.

A further step in analysis is the distinction between dictionary-based (quantitative) approaches and qualitative approaches. Dictionary-based approaches set up a list of categories derived from the frequency list of words and control the distribution of words and their respective categories over the texts. While methods in quantitative content analysis in this way transform observations of found categories into quantitative statistical data, the qualitative content analysis focuses more on the intentionality and its implications.

12.3 Focus Groups

A focus group is a form of qualitative research in which a group of people are asked about their perceptions, opinions, beliefs and attitudes towards a product, service, concept, advertisement, idea, or packaging. Questions are asked in an interactive group setting where participants are free to talk with other group members. There are many definitions of a focus group in the literature, but features like organised discussion (Kitzinger 1994), collective activity (Powell 1996), social events (Goss & Leinbach 1996) and interaction (Kitzinger 1995) identify the contribution that focus groups make to social research. Powell define a focus group as “a group of individuals selected and assembled by researchers to discuss and comment on, from personal experience, the topic that is the subject of the research”.

Focus groups are a form of group interviewing but it is important to distinguish between the two. Group interviewing involves interviewing a number of people at the same time, the emphasis being on questions and responses between the researcher and participants. Focus groups however rely on interaction within the group based on topics that are supplied by the researcher.

The main purpose of focus group research is to draw upon respondents’ attitudes, feelings, beliefs, experiences and reactions in a way in which would not be feasible using other methods, for example observation, one-to-one interviewing, or questionnaire surveys. These attitudes, feelings and beliefs may be partially independent of a group or its social setting, but are more likely to be revealed via the social gathering and the interaction which being in a focus group entails. Compared to individual interviews, which aim to obtain individual attitudes, beliefs and feelings, focus groups elicit a multiplicity of views and emotional processes within a group context. The individual interview is easier for the researcher to control than a focus group in which participants may take the initiative. Compared to observation, a focus group enables the researcher to gain a larger amount of information in a shorter period of time. Observational methods tend to depend on waiting for things to happen, whereas the researcher follows an interview guide in a focus group. In this sense focus groups are not natural but organised events. Focus groups are particularly useful when there are power differences between the participants and decision-makers or professionals, when the everyday use of language and culture of particular groups is of interest, and when one wants to explore the degree of consensus on a given topic.

Focus groups can be used at the preliminary or exploratory stages of a study; during a study, perhaps to evaluate or develop a particular programme of activities; or after a programme has been completed, to assess its impact or to generate further avenues of research. They can be used either as a method in their own right or as a complement to other methods, especially for triangulation and validity checking.

12.3.1 Types of Focus Groups

There are different variants of focus groups which include:

- **Two-way focus group** - one focus group watches another focus group and discusses the observed interactions and conclusion
- **Dual moderator focus group** - one moderator ensures the session progresses smoothly, while another ensures that all the topics are covered
- **Dueling moderator focus group** - two moderators deliberately take opposite sides on the issue under discussion
- **Respondent moderator focus group** - one and only one of the respondents are asked to act as the moderator temporarily
- **Client participant focus groups** - one or more client representatives participate in the discussion, either covertly or overtly
- **Mini focus groups** - groups are composed of four or five members rather than 6 to 12
- **Teleconference focus groups** - telephone network is used
- **Online focus groups** - computers connected via the internet are used

12.3.2 The Practical Organisation of Focus Groups

Organising focus group interviews usually requires more planning than other types of interviewing as getting people to group gatherings can be difficult and setting up appropriate venues with adequate recording facilities requires a lot of time.

The recommended number of people per group is usually six to ten, but some researchers have used up to fifteen people (Goss & Leinbach 1996) or as few as four (Kitzinger 1995). Numbers of groups vary, some studies using only one meeting with each of several focus groups, others meeting the same group several times. Focus group sessions usually last from one to two hours. Neutral locations can be helpful for avoiding either negative or positive associations with a particular site or building. Otherwise the focus group meetings can be held in a variety of places, for example, people's homes, in rented facilities, or where the participants hold their regular meetings if they are a pre-existing group.

It is not always easy to identify the most appropriate participants for a focus group. If a group is too heterogeneous, whether in terms of gender or class, or in terms of professional and 'lay' perspectives, the differences between participants can make a considerable impact on their contributions. Alternatively, if a group is homogenous with regard to specific characteristics, diverse opinions and experiences may not be revealed. Participants need to feel comfortable with each other. Meeting with others whom they think of as possessing similar characteristics or levels of understanding about a given topic, will be more appealing than meeting with those who are perceived to be different.

Once the types of participant have been decided, locating them is the next challenge. Recruitment of participants can be time consuming, especially if the topic under consideration has no immediate benefits or attractions to participants. It is likely that people with specific interests will have to be recruited by word of mouth, through the use of key informants, by advertising or poster campaigns, or through existing social networks. Incentives, whether expenses, gift vouchers or presents, will usually need to be offered.

12.3.3 Essential Requirement for Focus Groups

Employing Axelrod's (1975) ten essential ingredients for a successful focus group, one would have:

1. **A Clearly Understood Objective**: Is the focus group part of an on-going research project or is it self-contained? Does the research team have a clearly defined subject of study?

2. **Homogeneity Within the Group**: The participants should be homogeneous in relation to the topic under discussion (i.e., all should either have or have not been exposed to the topic of study).

3. **Good Recruiting**: Recruiting should be done to insure homogeneity and a sufficient number of qualified participants.

4. **A Relaxed Atmosphere**: The moderator should insure confidentiality and promote openness.

5. **A Moderator Who Listens**: The moderator must insure that the discussion does not stray too far from the point of interest, yet must not rule out things that may seem unrelated.

6. **A Well Prepared Moderator**: The moderator typically follows an unstructured interview guide.

7. **Free-Flowing Dialogue**: The moderator should begin the discussion by inviting honest and open dialogue and guiding the discussion only when necessary.

8. **Restrained Group Influence**: The moderator should refrain from contributing to the discussion unless necessary.

9. **Skilled Analysis**: The data can be analyzed by either a qualitative, or ethnographic summary; or a quantitative systematic coding via content analysis (Morgan, 1988, p. 64).

10. **Competent Researchers**: The research team should be sure that all necessary details are controlled.

A successful focus group is one in which a variety of responses are generated which are germane to the topic of study. All participants feel free to express opinions and thoughts regarding the topic at hand.

12.3.4 The role of Moderator

Once a meeting has been arranged, the role of moderator or group facilitator becomes critical, especially in terms of providing clear explanations of the purpose of the group, helping people feel at ease, and facilitating interaction between group members. During the meeting moderators will need to promote debate, perhaps by asking open questions. They may also need to challenge participants, especially to draw out people's differences, and tease out a diverse range of meanings on the topic under discussion. Sometimes moderators will need to probe for details, or move things forward when the conversation is drifting or has reached a minor conclusion. Moderators also have to keep the session focused and so sometimes they may deliberately have to steer the conversation back on course. Moderators also have to ensure everyone participates and gets a chance to speak. At the same time moderators are encouraged not to show too much approval, so as to avoid favouring particular participants. They must avoid giving personal opinions so as not to influence participants towards any particular position or opinion.

The role of the moderator is a demanding and challenging one, and moderators will need to possess good interpersonal skills and personal qualities, being good listeners, non-judgmental and adaptable. These qualities will promote the participants' trust in the moderator and increase the likelihood of open, interactive dialogue.

Finally, the degree of control and direction imposed by moderators will depend upon the goals of the research as well as on their preferred style. If two or more moderators are involved in the facilitation of a focus group, agreement needs to be reached as to how much input or direction each will give. It is recommended that one moderator facilitates and the other takes notes and checks the recording equipment during the meeting. There also needs to be consistency across focus groups, so careful preparation with regard to role and responsibilities is required.

12.3.5 Advantages of the Focus Group

The advantages of the focus group are as follows:

1. **Release of inhibition by participants**: A well moderated group encourages full and open expression of perceptions, experiences, attitudes, etc.

2. **Flexibility**: A focus group is typically more flexible than an individual. The moderator "works from a list of topics - listening, thinking, probing, exploring, framing hunches and ideas".

3. **Handling contingencies**: A focus group is amenable to exploring linkages which go untouched in a statistical survey. Moreover, it is possible to explore avenues of importance which may arise other than those listed on a questionnaire.

4. **Time**: Eliciting responses from eight to twelve respondents in a focus group lasting one to two hours is more "time effective" than interviewing the same number individually.

5. **Interpretability of data**: Though the data usually contain a wide range of responses, identification of issues and the reasons participants hold positions on issues is usually clear upon careful analysis. The group often stimulates recall and actuates important but forgotten personal detail.

6. **Provision of basic exploratory information**: When little is known in advance of investigation, the focus group may provide a basis for formulating research questions and hypothesis.

12.3.5 Disadvantages of the Focus Group

The focus group method has some disadvantages:

1. **Cost**: A series of four focus groups could easily cost more, depending on moderator fee, facility rental, recording and transcribing, data analysis and interpretation, and participant incentives.
2. **Subjects' conformity**: Social desirability, or respondents' motivation to provide socially acceptable responses to conform to group norms is somewhat greater in a group than in the anonymous process of survey questionnaire completion.
3. **Biased results**: An analyst should not generalize from focus group results to the larger population from which the respondents were a sample, and it is well to remember that the respondents are volunteers who may be more extroverted, outgoing, and sociable than the "average" individual.

Therefore, the main features of focus group research is paying particular attention to the benefits of interaction and group dynamics which only this method can offer. Practical considerations and the time it takes to conduct focus group research may discourage many from attempting to collect data using this method. Nevertheless those who participate in this kind of research often find the experience rewarding. The process of research can be more collaborative than other forms of study, and so focus group research can be an empowering process for participants, and an exciting challenge for social researchers wanting to gain a different perspective on their field of interest.

12.4 Field Surveys

A structured survey in which interviewers visit the residences, businesses, and/or public places oftargeted consumers and ask a predetermined list of questions that the client needs answers to is called Field survey. The purpose of conducting field surveys is to increase revenue and efficiency by obtaining quantifiable information in a relatively quick time period, which can be reliably aggregated and compared via a targeted group of individuals that live, work, or recreate in specific locations. Although this method may seem somewhat archaic, field surveys can be the most effective and cost efficient method of receiving insightful market intelligence at the grass roots level.

In order to obtain an adequate quantity and quality of information when researching any specific location, it is sometimes necessary to acquire quantitative and qualitative data directly (face-to-face) from consumers and/or businesses. The researcher dedicates its efforts to obtain feedback from consumers within the locations being researched and analyzed when pertinent. Before beginning field surveys, the researcher prepares extensively to find a representative sample of consumers within a targeted location. Survey questions are formulated so that the response rate is maximized.

In sociology the word survey refers to the study of a population through observation of its members, as it has been carried out for ages in censuses. In modern times, most surveys use a sample of members to measure population characteristics. Therefore as per Groves "The survey is a systematic method for gathering information from (a sample of) entities for the purpose of constructing quantitative descriptors ofthe attributes of the larger population of which the entities are members."

The population under researcher study may include the inhabitants of a town or a country, or the members of a specific category like teachers or left-handed tennis players, etc. The point is that the study does not observe social interactions or communications between persons or institutions in a given population, but only characteristics of the individual members involved, e.g. alcohol consumption, political affiliation, preferred color of coat, etc. There is also a qualitative way of defining and investigating variation in populations, however. The qualitative type of survey does not aim at establishing frequencies, means or other parameters but at determining the diversity of some topic of interest within a given population. This type of survey does not count the number of people with the same characteristic (value of variable) but it establishes the meaningful variation (relevant dimensions and values) within that population.

12.4.1 Open (Inductive) Versus Pre-structured (Deductive) Qualitative Surveys

There is a need for distinction between open (or inductive) and pre-structured (or deductive) qualitative surveys. In the open/inductive survey, relevant objects/topics, dimensions (aspects of objects, variables) and categories (values at dimensions) are identified through interpretation of raw data (e.g. interview transcripts). In the pre-structured survey, some main topics, dimensions and categories are defined beforehand and the identification of these matters in the research units is guided by a structured protocol for questioning or observation. In the pre-structured case the diversity to be studied is defined beforehand and the aim of descriptive analysis is only to see which of the predefined characteristics exist empirically in the population under study.

Many qualitative researchers tend to identify qualitative research with induction (open coding), thereby excluding the analysis of pre-structured data. The pre-structured diversity analysis is concerned with diversity as opposed to numerical distribution. As a fictitious example: an observational study on the diversity of consumer styles, in terms of predefined trademarks of clothing, shoes and drinks, and music styles among Rotterdam adolescents, would correctly be classified as a qualitative survey.

Another source of confusion regards the use of quantitative (metric) data in qualitative surveys. The point here is that the qualitative-versus-quantitative nature of data is established in the analysis. It is not inherent ontology but analysis which determines whether a study is qualitative or quantitative. Again, a fictitious example to illustrate this point: a study on body length is a qualitative survey if it searches for the categories (values) of this dimension that are present in a given population and if it uses these metric data as categorical data in further analysis. In other words: a survey is a qualitative survey if it does not count the frequencies of categories (values), but searches for the empirical diversity in the properties of members, even if these properties are expressed in numbers. For example: it may seem hard to imagine the relevance of such a study on the diversity of body length, but this survey could be a relevant part of a comparative study on interpretation and categorization of body images in ethnic subcultures, for example.

12.4.2 The Empirical Cycle in Qualitative and Statistical Surveys

Qualitative research differs from quantitative research based on the iteration of data collection and analysis in one project: the qualitative researcher starts with some data collection, analyzes them, develops a hypothesis about the subject, and then samples new units theoretically (i.e. informed by the hypothesis to be tested) for data collection and so on until a theoretical saturation of concepts (categories) or a full explanation of the phenomenon is reached. This is the logic of both grounded theory in developing and saturating concepts and analytic induction in developing and testing hypotheses. In this type of qualitative research, bothdata collection and the research question develop in interaction with data analysis.

However, many qualitative studies are based on a single one-shot, one-method sample, sometimes for pragmatic reasons (depending on available money and time), other times because of good prior knowledge or even because of the availability of a pre-structured inventory of codes. The one-shot survey involves only one empirical cycle (research question—data collection—analysis—report) in parallel to the typical case of a statistical survey. Due to this parallel, there are different stages of the research process for both the qualitative survey and the statistical survey in parallel. The following table elaborates the comparison between qualitative survey and statistical survey.

Table - 3: Qualitative Survey and Satistical Survey

Steps	**Qualitative Survey**	**Statistical Survey**
1. Defining knowledge aims		
Topic (material object)	any topic	any topic
Aspect (formal object)	diversity	frequency distribution
Empirical domain	any population (collection)	any population (collection)
Unit of data collection	members of population	members of population
Knowledge function	primarily description	primarily description
2. Sampling		
Method of selection	diversity; by purpose	probability; by chance
Criterion for size (N)	saturation, coverage of population diversity	precision of estimate (CI)
3. Data collection		
Measurement level	any	any
Method of collection	any	any
4. Analysis	diversity analysis	distribution analysis
1st-level analysis Unidimensional description	coding data (downward and upward) in objects, dimensions and categories	counting frequencies descriptive statistics estimating parameters
2nd-level analysis Multidimensional description	*case oriented:* combinatory synthesis of diversity: property-space analysis, typology construction *concept oriented:* holistic synthesis by core concept	*unit oriented:* cluster analysis, homogeneity analysis *variable oriented:* correlation, factor-analysis, index construction, scaling
3rd-level analysis Explanation	deterministic explanation: combinatory analysis QCA, pattern analysis	probabilistic explanation: discriminative analysis, regression, LISREL

12.4.3 Sampling

The statistical survey aims at estimating/evaluating the frequencies of characteristics of units in a population. This aim requires a probability sample. In order to establish the statistical reliability of estimates the researcher needs to know the probability for each member of the population to be selected in the sample. Therefore, one needs a full register of population members as a sampling frame. The sample size is determined by the level of accuracy needed in the population estimates, as measured by the confidence interval (CI) and the confidence level (á). For a 100% precision (á =1) the total population should be included.

A qualitative sample should represent the diversity of the phenomenon under study within the target population. This could be achieved by a large random sample, but this method would not be very efficient. It is both logical and more efficient to purposively select a diversity sample with the aim to cover all existing relevant varieties of the phenomenon (saturation). What saturation is depends on the type and degree of diversity that is judged relevant. Take as an example, a study of the colors of coats that students wear. To be sure that all forms of diversity are covered it would be necessary to include the whole population in the sample, because the number of colors that our eyes and brains can distinguish is innumerable. However, for a single study it might be sufficient to distinguish only, say, ten colors. At that low level of detail a small sample may provide sufficient saturation.

In an interview survey with open questions, each answer is unique. Here also all the members of the population under study should be included to guarantee full and detailed coverage. Therefore, in an inductive description of diversity (i.e. without a predefined coding list as in the case of colors), it seems efficient to follow the following steps:

- start with a small sample,
- perform an intermediate analysis to develop categories,
- decide on a strategy to find uncovered categories, i.e. respondents who are not represented in the categories as developed in step b,
- define a rule as to when to stop (e.g. after five interviews without relevant new information).

In a qualitative survey, saturation is an empirical question, not so much a theoretical one, as in Grounded Theory. The goal is not to detail concepts exhaustively for a theoretical domain (i.e. to cover all theoretical possibilities), but to cover relevant (in terms of aims) diversity in an empirically-defined population which may comprise only a small number of units (e.g. a class in school).

12.4.4 Data collection

The data collection method is not limited by the study design in itself, nor is the type of data to be collected. Both statistical and qualitative surveys may collect data by questioning people—which is the most common type of survey—but also by observing interactions or artifacts in any kind of situation.

12.4.5 Analysis

Both statistical and qualitative surveys are descriptive designs in the first place. In most cases researchers want to get more out of it, however, in terms of:

a Relationships between characteristics of the units, i.e. patterns of categories (in a qualitative survey) or correlations between variables (in a statistical survey), respectively, to gain compact multidimensional description of diversity/variance;

b Explanation of diversity/variance in the subject of study by contextual determinants.

Here researchers may switch from a qualitative procedure to a quantitative one, especially when there is a large number of cases in the data. This is very common in market research for instance, under the assumption that correct statistical representation is not important when it comes to relationships between variables. Traditionally, statistical analysis has been classified according to the number of variables involved. In practice, however, most often a dichotomous classification is used by distinguishing univariate versus multivariate analysis. Some types of multivariate analysis are descriptive (e.g. index construction) and others are explanatory (e.g. path analysis and regression analysis).

In the qualitative literature the various levels of analysis are classified in terms of depth ranging from superficial description to theoretical interpretation or in terms of distance from the data as achieved "by cumulative steps of data transformation".

12.5 Application of Qualitative Research

Research methods and techniques in research studies on business are not predominant. There is a wide range of alternatives, making it a challenge to understand the situations in which a specific method or technique can be used. For example, although qualitative research has flourished in many fields of the social sciences & business, at the same time its use has become fragmented and incoherent. On one hand, the proliferation of qualitative research methods can be confusing; on the other hand, it can unveil new and different forms of conducting research. There are so many ways of conducting research that the researcher must carefully analyze the conditions and the resources to extract as much useful information as possible for the purpose of acquiring knowledge.

All business management professionals will resort to research as a source of information at one time or another in the course of their professional career. Research results can be a precious source of information to improve the decision-making process. Research studies do not actually solve problems or make the decisions. They generate information that can guide management decisions and actions. Research has different meanings, depending on the public. However, among several definitions, there seems to be a consensus that research: (i) is an investigation and inquiry process; (ii) is systematic and methodical; (iii) increases knowledge.

One way that qualitative research methods are used is in "formative" research to inform the design of a study or program. Findings from such research help survey designers in many ways: from identifying the most appropriate way to phrase a question to determining which questions to ask and whom to survey. The quality research can be applied in most of the business. However, few applications of quality research are mentioned as follows:

12.5.1 Business Marketing

The qualitative research methods can be applied in the field of marketing, both in terms of methodological considerations and the structure and logic of the narrative. The goal in the quality research may be to understand clients, the way they think, the issues and influences they take into account, and the decisions they make. The reason for researching clients (the demand), their buying behavior, and the considerations affecting their decision-making is to assist providers (the supply side) in developing marketing strategies and construing their market niches.

After a wave of positivist euphoria in the social sciences and unquestionable successes in a number of areas, many researchers started to feel and understand the limits of quantitative thinking. The emergence of deconstructionism created a new philosophical and epistemological context and accelerated this process. Against this background, the deficiencies of positivist thinking become more obvious. Most social scientists, and especially marketers, have embraced positivist and other objectivist approaches to the study of social reality. In a number of cases, these approaches have yielded pretty good results and have helped society to organize a number of actions much better than they had been organized before.

Approaches of the positivist type are based on a well-prepared, systematic and thorough survey of a large number of respondents, selected at random according to particular rules for ensuring a representative sample. Examples can be brought in from opinion polls and various marketing studies. This approach is very good as long as there are enough materials before the start of the study to prepare an adequate questionnaire and to select the proper sampling technique. In addition, the approach also depends on site-specific conditions to carry out the survey and to sustain the integrity of the sample. However, when these prerequisites are difficult to maintain, quantitative research methods often become irrelevant and yield misleading results.

12.5.2 Transport Business

Qualitative survey methods are being used increasingly in research and policy studies to understand traveler perceptions, attitudes and behavior, as a complement to more established quantitative surveys of public perceptions, attitudes and behavior.

It is correct to say that qualitative research can be used to explore the range of issues present within a given population, and that this can guide the design of subsequent quantification. This is particularly relevant in situations that are dynamic or new. An example of this is where a new area of investigation is being considered. A new area of investigation for the Department of Transport lay in understanding the attitudes of drivers to the provision of motorways. In UK, a major qualitative study was undertaken to explore the range of factors that influenced the public's perception of the provision of motorways. This study, which led to further quantification, involved many group discussions with different categories of motorway drivers, different geographic locations and different social groupings.

12.5.3 Health Service Business

Qualitative methods have much to offer those studying health care and health services. However, because these methods have traditionally been employed in the social sciences, they may be unfamiliar to health care professionals and researchers with a biomedical or natural science background. Indeed, qualitative methods may seem alien alongside the experimental and observational quantitative methods used in clinical, biological and epidemiological research.

Misunderstandings about the nature of qualitative methods and their uses have caused qualitative research to be labeled 'unscientific', difficult to replicate or as little more than anecdote, personal impression or conjecture. The growing interest in qualitative methods in health research, and their increasing acceptance in clinical and biomedical arenas, in the last 10 years, suggest that such misunderstandings may be diminishing.

Qualitative researches not only useful as the first stage of quantitative research. It also has a role to play in 'validating' quantitative researcher in providing a different perspective on the same social phenomena. Sometimes it can force a major reinterpretation of quantitative data. For example, one anthropological study using qualitative methods uncovered the severe limitations of previous surveys: Stone and Campbell found that cultural traditions and unfamiliarity with questionnaires had led Nepalese villagers to feign ignorance of abortion and family planning services, and to underreport their use of contraception and abortion when responding to surveys. More often, the insights provided by qualitative research help to interpret or understand quantitative data more fully. Bloor's work on the surgical decision making process built on an epidemiological study of the widespread variations in rates of common surgical procedures and helped to unpack the reasons why these variations occurred. Elsewhere, Morgan and Watkin's research on cultural beliefs about hypertension has helped to explain why rates of compliance with prescribed medications vary significantly amongst and between white and Afro-Caribbean patients.

In light of what was discussed above, the conclusion is that qualitative research involves concepts, methods, techniques and procedures that are combined and become rather unclear to the researcher. On the other hand, qualitative research offers excellent opportunities to develop research for the business sector, especially in the area of behavior studies. Because of the nature of business problems, it seems clear that the potential of qualitative research is yet to be explored to its fullest extent, in terms of application and use and in terms of the discussion of its scope.

12.5.4 Consumer Behaviour

The study of consumer behavior gained momentum in the late 1940s and 1950s, especially as its value was embraced by advertising agencies on behalf of their clients. A lot of that work was proprietary and did not appear in journals—and the Association for Consumer Research and the Journal of Consumer Research did not show up until the 1970s. However, there was the large-scale survey work by Alfred Politz and a variety of so-called motivation or qualitative research studies—many of which were reported on in the trade press. The lively character of the work in that post-World War II period is especially evident in the two volumes on Consumer Behavior edited by Lincoln H. Clark (1954–1955). These volumes report on two conferences held at the University of Michigan that brought together scholars from several research organizations. The disciplines of sociology, psychology, and economics were well-represented, and although there were tables and statistics reported—and the University of Michigan Survey Research Center had a dominating role—the papers had a markedly thoughtful and discursive quality, suggesting some breadth of thinking and conversation that went on about the lives of consumers, their life cycles, and their decision making. Such earlier subjects and methods are like the trunk of a tree that in time grew many branches and leafed out into the variety of sessions and the flowering richness of specific topics apparent in modern journals and in the contemporary conferences of the Association for Consumer Research, the Society for Consumer Psychology, and the International Research Seminar.

Historically, more and more scholars have come to seek that comprehension of the situation, or at least to approximate it. Thus, despite the resistance that still occurs, this aspiration has gained a large number of adherents, more visibility, and a fairly loud voice. The more fully researchers want to understand consumer behavior, the more they are motivated to use methods that allow the interaction of multiple forces to show itself. A consumer event would be intensively scrutinized by a team of thinkers representing every discipline, explaining every possible antecedent and current element with any possible effect on the action at issue. Short of that, we engage in the varieties of research activity called qualitative research. That means going beyond the exponents of behaviorism (unlike behavioralists) who want to limit their data and thinking to explicit and plainly observable acts and events, to stimulus and response; and who focus on the degree to which the phenomena occur and the level of confidence we may have in getting the same results were the study to be repeated. Being more broadly behavioral means the unleashing of all we can do to find out what consumers' lives are like, especially with reference to the situations that interest us.

Getting reports on their actions may seem the easier part, although some skeptical scientists want to watch the behavior rather than be told about it in an interview or a laboratory. That means going out into the field, mingling with the subjects as if being one of them, making detailed notes, creating the methods of case studies, participant observation, autodriving, ethnography, and thick descriptions. We are also challenged to deal with consumers' inner lives, including their introspections, with all the hazards entailed in their self-expression, their truth and lies, their ignorance, uncertainty, face work, contradictions, and mechanisms of defense—all in all, their being complex human beings. Methods arise to gather this sort of information, such as depth interviews, focus groups sessions, and projective techniques of Quality Research. That also means we have to interpret the complexity and make inferences from what is observable to what is underlying and theoretically operative.

Regardless of the long history, it is a sign of the irregular situation of qualitative research that examples of its application still turn up in the business press as if it were some remarkable newcomer. Recently, The Wall Street Journal (1999) reported Chrysler's "first vehicle designed entirely through an unconventional market-research process known as 'archetype research.' (Ah, Carl Jung!). . .overseen by a. . .French-born medical anthropologist named G. Clotaire Rapaille." The research involved poring over focus group protocols in which participants were asked to drift back to their childhoods and jot down the memories invoked by the prototype of the vehicle. A project about Procter & Gamble's Folgers Coffee was also mentioned in which "Dr. Rapaille concluded that aroma sells coffee more than taste does because aroma invokes feelings of home." If this was news to Procter & Gamble, perhaps the "unconventionality" and ever-renewed sense of novelty about qualitative research is due to the persisting naivete´ and capacity for astonishment of corporate personnel.

Nevertheless, in the face of continuing contention, acrimony, and defensiveness on all sides, qualitative researchers have persisted and amplified their numbers. Their stream of work seems to be well established now. It has taken root and a substantial number of scholars are devotedly pursuing the content and methods of qualitative study and publishing their work. Although qualitative research is a subfield of a relatively small field, it has radiated out into the world.

However, looking back and summing up in this field are not far advanced and there is a need for more grand integrations and overviews. As academic consumer research studies were gathering steam in the 1960s, the landmark text by Engel, Kollat, and Blackwell that emerged in 1968 organized its contents by using the common categories that continue today. These are general cultural and environmental forces, reference group memberships, and psychological components that affect consumer decision-making. Recent and contemporary qualitative consumer research has clearly made a place for itself and contributed to a great richness of detail in exploring consumer situations.

12.6 Attitude Study

Consumer researchers often use qualitative techniques to examine consumer attitudes. The attitudinal output from these qualitative methods may reflect group attitude formation processes rather than the individual's attitudes.

Qualitative methods are defined in terms of the information gained from their use subjective, in-depth understandings of the consumer, and the nature or structure of the consumers' attitudes, feelings, and motivations. Qualitative research methods attempt to uncover what people think or how they feel, achieving greater depth and detail of responses, and resulting in close-up descriptions that better realize the subjective nature of the phenomenon studied.

One type of output frequently sought from the various qualitative techniques is consumer's thoughts, e.g., attitudes, beliefs, opinions, and purchase intentions. For example, these techniques have been used to gain insights into the brand attitudes of women, attitudes towards packaging aesthetics, and attitudes towards financial services.

Consumer researchers have reported the use of **four types** of qualitative research techniques to investigate consumer attitudes: open-ended surveys, individual interviews, focus group interviews, and nominal group interviews. These techniques vary in the exact procedures and settings used to collect the information, and in their assumptions of the quality of the output gained.

Open-ended Surveys (or self-administered questionnaires) provide respondents with written instructions and questions, yet do not provide structured responses. Researchers often use surveys to measure consumer attitudes. Surveys are easily and economically administered on a mass basis to a sample of individuals. Additionally, researchers often use unstructured, open-ended response formats to preclude constraining and/or biasing individual's responses.

Individual Interviews involve relatively open-ended, but interviewer guided, discussions of specific topics. Interviews allow the researcher as interviewer to probe responses and redirect questions towards the respondent. Thus, individual interviews can be more adaptive and less structured than surveys.

However, the former are costly when a large number of respondents are desired. Like individual interviews, **focus group interviews** also involve a moderator guided open-ended discussion of specific topics. However, focus groups are interactive as the topics are simultaneously discussed by a small group of individuals. Researchers often use focus groups to gain insight into consumer attitudes because of the convenience in interviewing several respondents simultaneously, and because consumers are more likely to respond in a group interview setting. Thus, focus groups can provide data quicker and more economically than individual interviews

Nominal Group Techniques are highly structured group interviews that restrict spontaneous interaction. Nominal groups differ from focus groups in the added structure to the interview. For example, participant's responses are often written and shared with the group on a cumulative basis rather than spontaneously and openly discussed as found in focus groups. Additionally, vocal evaluation of the output is prohibited the moderator collects written evaluations by each individual. Like focus groups, the nominal group technique purportedly results in higher involvement and response rates compared to non-group techniques.

Although the effects of using qualitative techniques on the quality of attitudinal output has not been empirically examined. Many studies have found that initial tendencies of individual group members intensify or sometimes change with group discussion. Several theories have been suggested to explain this phenomenon, several of which apply to the use of qualitative techniques to measure consumer attitudes. The theories of social facilitation, social impact, social comparison, and persuasive-arguments are as explained below.

12.6.1 Social Facilitation and Social Impact

Social facilitation is perhaps the simplest theory from which predictions can be posited about measuring attitudes across social and nonsocial settings. This theory suggests that the mere presence of others is a sufficient explanation of behavior in groups. The presence of observers or co-actors has been found to result in greater individual effort and performance. However, we are interested in the effect of others on individual attitudes. The presence of others leads individuals to focus attention on themselves and increases self-awareness and thought about one's own attitudes and feelings. Increased self-awareness has been found to lead to polarized attitudes and evaluations. Attitude polarization refers to individuals adopting more extreme attitudinal positions than those previously held & it is a shift or change in degree but not in direction. Thus, the mere presence of others leads to increased self-awareness that increases thought about one's own attitudes, resulting in attitude polarization.

Focus group and nominal group interviews contain 5 to 12 participants, whose presence may lead to increased self-awareness and attitude polarization by each individual. The interviewer is present in an individual interview and thus, this technique also may lead to polarization. However, no other individual need be present for an individual to respond to an open-ended survey.

It is expected that individuals' attitudes to become more polarized when others are present during the thought eliciting procedure. Therefore, the interviewer-directed thought eliciting methods (i.e., focus groups, nominal groups, and individual interviews) will produce more polarized attitudes than using open-ended, self-administered questionnaires. Any polarization in the individual's attitudes that is detected in the thought eliciting procedure should endure only while the individual is subject to the self-awareness and increased thinking about her or his position that results from the presence of others. When others are not present, individuals will fall back on their pre-existing attitude state. The polarized beliefs will not be salient or persist beyond the interview setting. The polarization of attitudes that is due to self-awareness has been found to dissipate over time. Therefore, any attitude polarization that results from the presence of others will be temporary and will not endure beyond the interview situation.

Social impact theory goes beyond the mere presence of others in its predictions. This theory predicts that the impact of others on individuals' beliefs, cognitions, values, and emotions is some power function of the number of other people present. According to social impact theory the effects of social facilitation should increase with the presence of more individuals. Thus, the greater the number of other people present, the greater self-awareness and the more polarized the attitudes. The status of group members would impact on attitude polarization and higher status individuals would cause greater attitude shift than lower status individuals.

Alternate predictions of attitude polarization due to mere presence result from applying social impact theory. Attitude shifts should occur in the group techniques and these shifts should be greater than in individual interviews in which an interviewer is present with the respondent (i.e., a group of two). Moreover, any attitude shift in individual interviews should be greater than shifts in open-ended surveys and could be attributed to the mere presence of the interviewer. Again, attitude polarization due to the impact of the social situation should endure only as long as the individual remains in the presence of others.

In summary, individual attitude polarization should increase as the number of others present increases such that polarization in focus and nominal groups will be greater than that in individual interviews. Additionally, as the number of individuals within the each group increases the positive effect on attitude polarization should increase. Finally, the shift effect should be transitory. Attitude polarization resulting from the presence of others is temporary and it should not endure beyond the interview situation.

12.6.2 Normative Influence Through Social Comparison

Social comparison theory was originally formulated by Festinger to explain the effects of social communication on opinion change in groups. One derivation of the theory has been used to go beyond mere presence theories to explain the polarization of group member attitudes that result from group discussion. Simply, group members' desires to be favorably evaluated leads them to adopt an attitude that is more extreme than the group norm, to the extent that they are aware of this normative position.

There are three necessary conditions for social comparisons to cause attitude polarization: (1) participants must desire to be favorably evaluated, (2) the setting must provide a standard of comparison, and (3) the setting must allow for the evaluations of others. Thus, in settings where others provide a standard of comparison, a desire to be favorably evaluated may motivate individuals to adopt/express more extreme attitudes. The standard of comparison in group interviews is a normative position on an issue that is explicitly stated by the interviewer/researcher or implicitly derived from the statements provided by group members during the discussion. Additionally, to be evaluated (favorably or unfavorably), the individual's views must be expressed and identified to others.

The desire to be favorably evaluated by others is pervasive. This desire is a source of evaluation apprehension an individual's anticipation of positive or negative outcomes when around others serves as an incentive function. According to social impact theory, we might expect this desire for favorable evaluation to be greater when other group members are acquaintances rather than strangers and higher status individuals rather than lower status individuals. Standards of comparison can be personal, objective, or social. It is this latter social standard that is produced in a setting in which others' views are shared. Individuals have been found to compare themselves with similar others in these settings. Both objective and social standards have been found to affect various types of responses.

Applied to the context at hand, the three necessary conditions for social comparisons to cause attitude polarization are only present when focus group interviews are used. Standards for comparison can develop in both focus groups and nominal groups, but are less likely to develop in individual interviews.

In focus groups each individual's attitudes are expressed and can become the focal point of discussion during the group exchange. However, in nominal groups the attitudes of individuals may be elicited but no conscious effort is made to discuss or otherwise scrutinize individuals' contributions. Since the ideas expressed within nominal groups are not openly discussed, the individual participant is less likely to be apprehensive about being evaluated by the other group members. With less emphasis on the evaluation of others, standards of comparison are less likely to affect attitude formation and change in nominal groups. The result should be less shifting in attitudes compared to focus group interviews.

In individual interviews and open-ended surveys, other members are not available for evaluation, consequently no standard of comparison exists, assuming the interviewer or questionnaire does not offer cues to such a standard. Thus, attitudes expressed by participants in individual interviews and open ended surveys should not polarize or shift from what the individual believed prior to the interview or questionnaire, compared to focus group interviews. Therefore, individual attitude polarization should increase under normative pressures. Polarization should be greater among focus group than nominal group participants. As previously noted, group attitude polarization should be more pronounced than polarization in individual interviews and open ended surveys.

Any change in the individual's attitudes that is detected in the interview will endure only as long as normative pressures are acting on the individual. When the norms are no longer governing individuals' behaviors, they will fall back on their pre-existing attitude structures.

12.6.3 Informational Influence Through Persuasive Arguments

An alternate explanation of attitude polarization or change in groups is information influence or persuasive-arguments theory. This theory posits that the exchange of information in groups can lead members to consider facts that they had not previously considered when initially forming their attitudes. This new information could not only lead to attitude polarization, but to attitude change or depolarization. Weekly held attitudes may be more easily changed because the individual is more influenced through given contradictory information provided by others. Evidence for both polarization and depolarization given group discussion has been reported.

Information is shared in both focus groups and nominal groups but is not provided to respondents in individual interviews and open-ended surveys. In focus groups each individual contributes information about her or his feelings. This information is often the focal point of group discussion. Although not interactively discussed in the group, information about the attitude object is also expressed in nominal group interviews. The individual participant is the only source of information in individual interviews and open-ended surveys, assuming the interviewer or questionnaire produce no demand artifacts.

Therefore, attitude polarization should be greater in group interviews because more beliefs, opinions, and feelings are shared and discussed among participants. As opposed to the other theories presented, information influence should result in an enduring change in the individual's attitudes. Attitudes that change because of new or additional information should remain stable outside the interview setting. The new information or beliefs should be salient and persist beyond the interview.

Therefore, individual attitude change or polarization, resulting from shared information, will be greater for group interview techniques than individual interviews or open-ended surveys. Moreover, the attitude change or polarization resulting from information influence is relatively permanent and will endure beyond the interview situation.

The Table-4 summarizes the relationships between the techniques and degree of attitude shift predicted by the theories.

Table - 4: Theoretical Predictions of Attitude shift

Theory	Qualititative Techniques		
	Focus Group	Nominal Group	Individual Interview
Temporary Shift			
Social Facilitation	Polarization*	Polarization	Polarization
Social Impact	Polarization	Polarization	Less polarization
Social comparison	Polarization	No Shift	No Shift
Enduring Shift			
Persuasive Arguments	Polarization or Change	Polarization or Change	No Shift

* Predicted of qualitative technique on attitude shift

12.7 Instruments of Data Collection

The data collection instruments used in qualitative research include case studies, observation (participant and direct), questionnaire, diaries, in-depth interviews, group interviews (focus groups), collection and qualitative analysis of relevant documents, including records, photographs, and video tapes (paper and electronic).

Self-reports and questionnaires frequently are used to obtain data about subjects' feelings, attitudes, or knowledge. There also are a number of well-developed research instruments that measure knowledge level for qualitative research. Common approaches to obtaining this type of information are through written surveys or interviews.

When using **Interviews**, the researcher decides in advance to use either structured or unstructured interviews. Structured interviews are designed by the researcher to direct both the content and focus of the interview. By developing specifically worded questions and asking them in a predetermined order, the researcher ensures a certain level of consistency when conducting the interviews. When a researcher conducts interviews, the responses usually are recorded and later transcribed. Then the researcher uses qualitative research methods to analyze the words of the participants. The researcher often is trying to identify repeating themes or ideas among the responses. Remember that the level of analysis in qualitative research is words, not numbers.

Unstructured Interviews usually start with a broad open-ended statement such as "Please tell me more about your experience undergoing emergency surgery." In this instance, the researcher simply encourages the subject to describe in detail a specific experience. The interviewer's goal is to obtain as much information as possible without determining the direction of the interview. In this situation, the research subject has an opportunity to control the direction of the responses. The questions that are asked during interviews or as part of self-report instruments, scales, or questionnaires must be phrased in a manner that best asks the question.

Developing a **Questionnaire** is a complex and sophisticated process. The researcher must first examine the concept or construct being researched. In the absence of a tested research instrument, the researcher must proceed with item and tool construction. Each item must be carefully worded and easily understood by the reader. The researcher also must be sensitive to factors such as gender, culture, sequence (ie, usually general to specific), and language. Many researchers have leaned the hard way that if the right question is not asked, the desired answers will not be obtained. Most researchers obtain expert consultation if they are faced with instrument or interview development.

Diaries are another research instrument that can be used to collect data in a qualitative format. Research participants may be asked to maintain either structured or unstructured diaries. Using structured diaries, the researcher asks participants to write about a particular event or situation for a certain number of minutes each day. In unstructured diaries, the participants are asked to write about a more general topic for a designated time period each day. After a certain period of time, the diaries are collected and analyzed using qualitative research methods.

When research scales or instruments are used, research subjects are asked to respond in a way that is quantified. For example, a respondent may be asked to answer “yes” or “no” or to choose a number on a numeric scale such as 1 to 4. A familiar scale for many nurses is a pain rating scale of 1 (ie, no pain) to 10 (ie, worst pain you can imagine). Likert-type scales are a common rating technique that is used to rate feelings, opinions, and attitudes. When a Likert-type scale is used, research participants are asked to rate certain statements on a numeric scale that ranges from “strongly disagree” to “strongly agree.”

There are a number of more specialized research methods that may be used. One such method is the **Q-sort**, which is a comparative reading process; this method often is used in instrument development. The **Delphi Technique** is another specialized method that commonly is used to examine opinions or determine priorities. By using the Delphi technique, the researcher develops a questionnaire and then surveys a panel of experts. When those responses are analyzed, they are provided to the participants along with a second survey. The researcher continues this process until there is consensus among participants.

Another common research technique is the use of a **Visual Analog Scale**. Using a visual analog scale, the researcher asks the participant to rate feelings, beliefs, or knowledge on a 100-mm line with bipolar anchors such as “all” and “none.” Each research subject is asked to place a mark on the line to indicate his or her experience with a particular phenomenon. The line is then measured and the results can be quantified for analysis.

12.8 Summary

Qualitative research is a method of inquiry employed in many different academic disciplines, traditionally in the social sciences, but also in market research and further contexts. Qualitative researchers aim to gather an in-depth understanding of human behavior and the reasons that govern such behavior. The qualitative method investigates the why and how of decision making, not just what, where, when. Hence, smaller but focused samples are more often needed than large samples. Qualitative researchers typically rely on the following methods for gathering information: Participant Observation, Non-participant Observation, Field Notes, Reflexive Journals, Structured Interview, Semi-structured Interview, Unstructured Interview, and Analysis of documents and materials.

12.9 Key Words

- **Content Analysis :** A research method for the subjective interpretation of the content of text data
- **Focus group:**A group of people are asked about their perceptions, opinions, beliefs and attitudes towards a product, service, concept, advertisement, idea, or packaging etc for research.
- **Field survey:**A structured survey in which interviewers visit the residences, businesses, and/or public places of targeted consumers and ask a predetermined list of questions that the client needs answers
- **Open-ended surveys:**Open-ended response formats to preclude constraining and/or biasing individual's responses for survey

12.10 Self Assessment Test

1 What is is Qualitative Research? Describe the same with comparison to quantitative research.

2 What is content analysis? Explain the uses of content analysis.

3 Explain the focus group technique used in qualitative research and mention its advantages and disadvantages.

4 What is field survey? Explain the procedure to conduct the same for qualitative research.

5 How consumer behaviour and attitude study conducted through qualitative research?

6 What are the instruments used to conduct qualitative research?

12.11 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.
- .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 - 13 : Correlation and Simple Regression

Unit Structure:

13.0 Objectives
13.1 Introduction
13.2 Correlation
13.3 Correlation Coefficient
13.4 Spearman's Rank Correlation
13.5 Simple Linear Regression
13.6 Estimating Linear Regression
13.7 Summary
13.8 Key Words
13.9 Self Assessment Test
13.10 Refrences

13.0 Objectives

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

- Concept of correlation
- The correlation co-efficient.
- Concept of liner regression.
- Estimation of liner regression
- Difference between correlation and regression

13.1 Introduction

From statistics point of you Correlation and Regression are the tools to measure the amount of similarity and variations of two or more variables. These are used for studying and measuring the extent of relationship between two or more variables. Generally, there are situation where there is a certain degree of association but not of a fixed type between variables which may occurs in pair or groups. For example, consider the heights and weights of 100 students in a class. We want to measure the heights and weights of all the students. We observe that if the height and weight of a student is represented by x and y respectively, then there corresponds a pair of values (x_i,y_i) of the variables x and y for each measurement. Such a distribution in which each individual of the set is made up of two values is called bivariate distribution.

If the changes in the value of one variable appear to be related to the changes in the value of another variable, then the two variables are said to be correlated. For example, pressure and volume of a gas are correlated because with the increase in pressure there is decrease in volume and the vice versa. Other similarly placed situations are:

The amount of yield of grain per acre and the amount of fertilizer per acre in number of agriculture lands.

The series of sales revenue and advertising expenditure of two firms in a particular year.

The series of the heights of the husband and the wives at the time of marriage.

In a bivariate distribution, we intend to find a relationship (if it exist) between two variables under study. Francis Galton used the term 'Regression' in the latter part of nineteenth century. In his published paper 'Regression towards Mediocrity in Hereditary Stature' he concluded that the average height of fathers is greater than the average height of sons. Regression is used to find out some sort of functional relationship between two or more variables. The average relationship between the correlated variables is estimated by Regression.

13.2 Correlation

The main objective of correlation is to study the relation (if it exists) between two or more variables which are so related that a changes in one variable is accompanied by changes in other variable. Regression, on the other hand, aims at making prediction of unknown values of one variable from the known values of other variable.

We define Correlation as a measure of relation between two variables. In other words, to describe and understand the association between two continuous variables (interval or ratio data), we compute Correlation. Thus two variables are said to be correlated if an increase (or decrease) in one variable is accompanied by a increase or decrease (decrease or increase) in the other variable. Generally, we have two types of Correlation:

- **Positive Correlation**

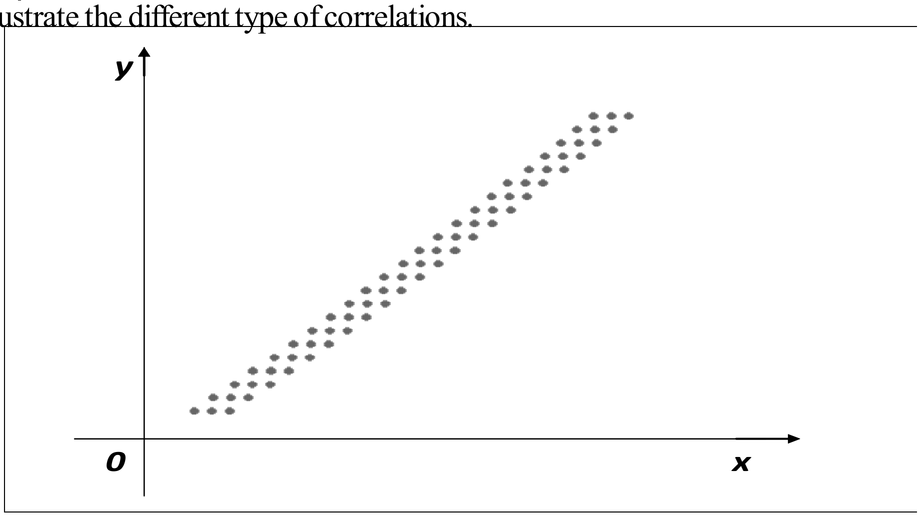
Two variables are said to be positively correlated, if an increase in the value of one is accompanied by an increase in the value of other or a decrease in the value of one is accompanied by a decrease in the value of other i.e. the value of two variables deviate in the same direction. For example, paired variables like supply and demand of commodities, Household income and expenditure, price and supply of commodities is positively correlated.

- **Negative Correlation**

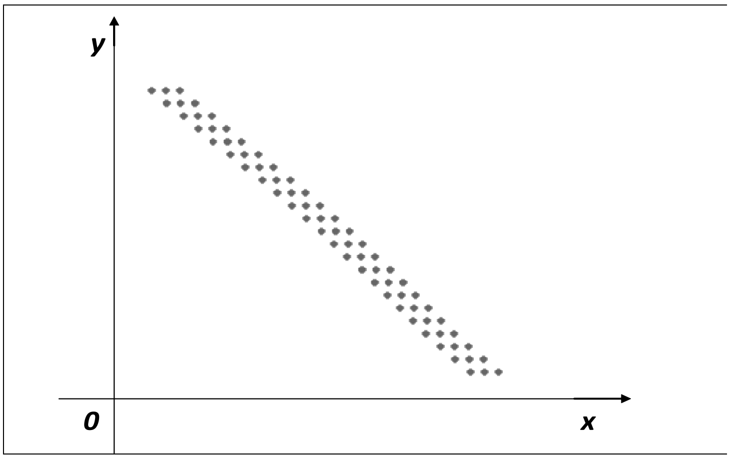
Two variables are said to be negatively correlated, if an increase in the value of one is accompanied by a decrease in the value of other or a decrease in the value of one is accompanied by an increase in the value of other. For example, paired variables like pressure and volume of a gas, current and resistance (voltage being constant), demand and price of commodities are negatively correlated.

- **Scatter Diagram**

Suppose we are given two paired variable values - one is related to the heights of the students while the other is related to the weights of the students. They are plotted on the coordinate plane using x-axis to represent heights and y-axis to represent weights. Then each paired observation shall have one point on the graph. The graphical representation of dots so obtained is called Scatter Diagram. The different graphs shown below illustrate the different type of correlations.



Scatter Diagram for Positive Correlation



Scatter Diagram for Negative Correlation

- **Linear Correlation and Non - Linear Correlation**

The correlation between two variables is said to be linear, if there exists a relationship of the form

$y = a_0 + a_1x$

where, a_0 and a_1 are real numbers

The graph of linear correlation is a straight line.

The correlation between two variables is said to be non - linear, if there exists a relationship of the form of polynomial of order more than one, for example.

$y = a_0 + a_1x + a_2x_2$

where, a_0 , a_1 and a_2 are real numbers.

In the non - linear correlation, a change of one unit in one variable does not correspond to a change of one unit in the other variable. The graph of non - linear correlation is not a straight line

13.3 Correlation Coefficient

The Coefficient of Correlation 'r' is a measure of the degree of relationship between two variables, say, x and y i.e. it measures the degree of association between the two values of related variables given in the data set.

The coefficient of correlation 'r' is given by

$$r = \left[\frac{n\sum xy - \sum x.\sum y}{\sqrt{\{n\sum x^2 - (\sum x)^2\}.\{n\sum y^2 - (\sum y)^2\}}}\right]$$

If the mean of variables x and y, denoted by X and Y are integers, than we can apply the fallowing formula to calculate the coefficient of correlation 'r'

The value of coefficient of correlation 'r' lies between - 1 and + 1. If for a given two sets or data,

r = +1, the two sets or data are said to be perfectly positively correlated.

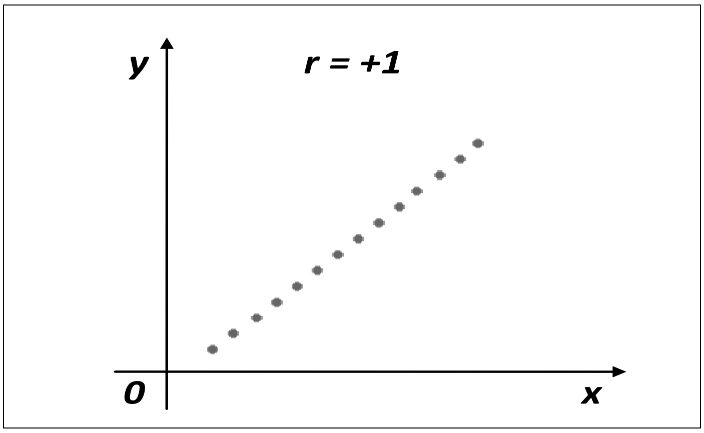
r = - 1, the two sets or data are said to be perfectly negatively correlated.

r = 0, the two sets or data are said to be uncorrelated, that is, there is absence of any linear relationship between the variables. However, there may exist some other form of relationship between them.

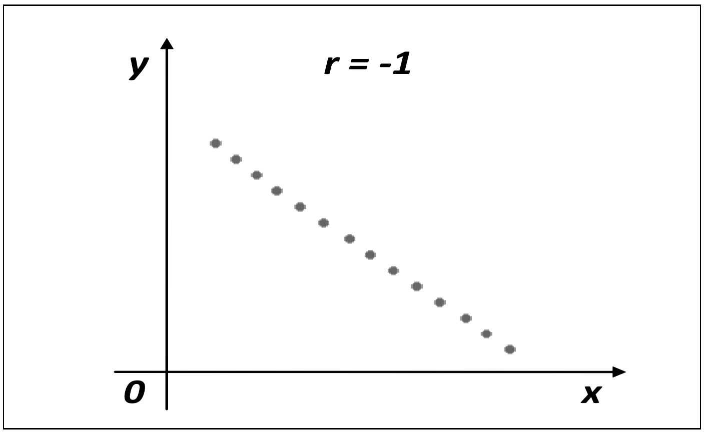
r = 0.8, the two sets or data are said to be strongly correlated.

r = 0.2, the two sets or data are said to be weakly correlated.

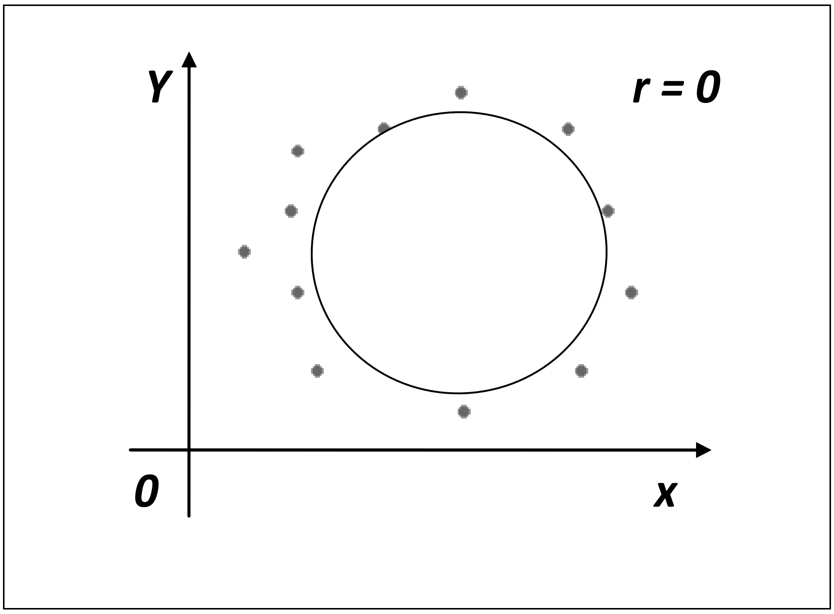
The following figures depict the various types of correlation corresponding to various values of coefficient of correlation:



Perfect Positive Correlation



Perfect Negative Correlation



No Definite Correlation

Properties of Coefficient of Correlation 'r'

The value of 'r' is independent of the origin of reference and the scale of reference i.e. 'r' is not affected by addition or subtraction of a constant to the values of either or both variables. It is also, unaffected by the multiplication or division of the values of either or both variables, by a constant.

The value of 'r' is free from any of the units. In other words, 'r' will have a definite meaning, whether the units of x and y variables are comparable or not. The coefficient of correlation is symmetrical in two variables. The value of 'r' lies between + 1 and - 1.

Example1:

Calculate the coefficient of correlation between X and Y series:

X	1	3	5	7	8	10
Y	8	12	15	17	18	20

Solution:

Table for Calculation

X	Y	X^2	Y^2	XY
1	8	1	64	8
3	12	9	144	36
5	15	25	225	75
7	17	49	289	119
8	18	64	324	144
10	20	100	400	200
$\sum X = 34$	$\sum Y = 90$	$\sum X^2 = 248$	$\sum Y^2 = 1446$	$\sum XY = 582$

The coefficient of correlation 'r' is given by

$$r = \left[\frac{n\sum xy - \sum x.\sum y}{\sqrt{\{n\sum x^2 - (\sum x)^2\}.\{n\sum y^2 - (\sum y)^2\}}}\right] \qquad ---(1)$$

Substituting the values from the Table in the equation (1), we obtain

$$r = \left[\frac{6.(582) - (34).(90)}{\sqrt{\{6.(248) - (34)^2\}.\{6.(1446) - (90)^2\}}}\right]$$

$$r = \left[\frac{(3492 - 3060)}{\sqrt{(332).(576)}}\right]$$

$$r = \left[\frac{432}{437.3}\right] = 0.988 \qquad Ans$$

Example 2:

Calculate the coefficient of correlation between the height of father and son from the following data:

Height of Father(in Inches) (x)	64	65	66	67	68	69	70
Height of Son(in inches)(y)	66	67	65	68	70	68	72

Solution: **Table for Calculation**

Height of Father(in inches)(x)	Height of Son(in inches)(y)	(x – X)	(y – Y)	$(x - X)^2$	$(y - Y)^2$	(x – X)(y- Y)
64	66	-3	-2	9	4	+6
65	67	-2	-1	4	1	+2
66	65	-1	-3	1	9	+3
67	68	0	0	0	0	0
68	70	+1	+2	1	4	+2
69	68	+2	0	4	0	0
70	72	+3	+4	9	16	+12
$\sum x = 469$	$\sum y = 476$	$\sum(x-X)= 0$	$\sum(y-Y)= 0$	$\sum(x-X)^2 = 28$	$\sum(y-Y)^2 = 34$	$\sum(x-X)(y-Y) = 25$

The coefficient of correlation 'r' is given by

$$r = \left[\frac{\sum(x - X).(y - Y)}{\sqrt{\sum(x - X)^2 \sum(y - Y)^2}}\right] \qquad --- (1)$$

Substituting the values from the Table in equation (1), we obtain

$$r = \left[\frac{25}{\sqrt{(28).(43)}}\right] = \mathbf{0.89}$$

13.4 Spearman's Rank Correlation

Ranked Data are defined as data which are arranged in numerical order, usually from largest to smallest and numbered 1,2,3,...... The correlation of such type of data is known as Rank Correlation, as here only ranks are considered. The coefficient of this type of correlation is called Rank Correlation Coefficient, denoted by 'R'.

Suppose we arrange a group of n individuals in order of merit of a certain characteristics. The same group in general give different order of merit for different characteristics. Let us assume that no two individual are equal in either classification. As each of the individual takes the rank values 1,2,3,4,....,n., their arithmetic means are equal. Hence,

$$X = Y = \frac{(n+1)}{2}$$

Thus, $\sum x = \sum y$ and $\sum x^2 = \sum y^2$

Suspose,

$$d = x - y = \{ x - (n + 1) / 2\} - \{ y - (n + 1) / 2\} = X - Y \qquad ----(1)$$

where, d is difference in rank for each pair and X and Y are the deviation from mean. We have

$$\sum X^2 = \sum\{x - (n + 1) / 2\}^2 = \sum x^2 - (n + 1) \sum x + \sum\{(n + 1) / 2\}^2$$

$$= \{n (n + 1)(2n + 1)\} / 2 - \{(n + 1)n(n + 1)\} / 2 + n\{(n + 1) / 2\}^2$$

$$= \{n(n^2 - 1)\} / 12 \qquad ----(2)$$

Similarly, we can obtain,

$$\sum Y^2 = \{n(n^2 - 1)\} / 12 \qquad ----(3)$$

Using equation (1), we obtained

$$\sum d^2 = \sum (X - Y)^2 = \sum X^2 - 2\sum XY + \sum Y^2 \qquad ----(4)$$

On using Equation (2) and (3) in Equation (4), we obtained

$$\sum XY = \{n(n^2 - 1)\} /12 - \{ \sum d^2\} / 2 \qquad ----(5)$$

Thus, using equations (2), (3) and (5) in the following

$$R = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

we obtain,

$$\boldsymbol{R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}}$$

where, n is the total number of individuals.

13.5.1 Properties of Rank Correlation Coefficient:

Rank Correlation Coefficient lies between +1 and - 1, including both the values. If dj stands for the difference in the ranks of the jth individual and if dj = 0, then Rank correlation Coefficient 'R' = 1. If a tie occurs in the ranks of two individuals, they are assigned the average of ranks that they would have obtained. For example, if two individuals are tied up at the fifth rank, each may be ranked (5 + 6) / 2 = 5.5 and next is then ranked 7.

Example 3:

The marks obtained by 8 students in Physics and Chemistry are given in the following Table:

Marks in Physics	52	54	67	82	98	90	69	76
Marks in Chemistry	11	7	23	36	56	37	12	25

Calculate the Rank Correlation Coefficient.

Solution:

After giving ranks to the students in Physics and Chemistry and calculating the difference d in ranks, we obtain the values shown in following Table:

Calculation Table

Marks in Physics(x)	Marks in Chemistry(y)	Rank in Physics	Rank in Chemistry	Difference in Rank (d)	d^2
52	11	8	7	+1	1
54	7	7	8	-1	1
67	23	6	5	+1	1
82	36	3	3	0	0
98	56	1	1	0	0
90	37	2	2	0	0
69	12	5	6	-1	1
76	25	4	4	0	0
Total					$\sum d^2 = 4$

The Rank Correlation Coefficient is given by

$$R=1-\frac{6\sum d^2}{n(n^2-1)}$$

=1- ((6).(4))/(8.(64 - 1))

= 1 - 0.047619

= 0.95238 Ans

13.6 Simple Linear Regression

Sir Francis Galton introduce the term 'Regression' in the end of the 19th century. He used the word 'Regression' to describe the phenomenon that average height of fathers is greater than the average height of their sons. In today's world the term 'Regression' is used to measure the average relationship between the correlated variables. In other words, 'Regression' is the prediction of unknown values of one variable from the known values of other variable.

Regression Analysis is an important technique used to analyze the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed. It helps to estimates the average value of the dependent variable when the independent variables are held fixed. It is one of the most important statistical tools which is extensively used in almost all sciences - Natural, Social and Physical. It is specially used in business and economics to study the relationship between two or more variables that are related causally and for the estimation ofdemand and supply graphs, cost functions, production and consumption functions and so on.

Linear Regression involves a relationship between two variables only. There is a linear regression between the variables under study, if the regression curve is a polynomial of degree one. In other word, it means that if the set of observations for these two variables are plotted, a straight line is drawn through the scatter gram.

The linear relationship between two variables x and y is represented by the following equation, which is a polynomial of degree one.

$$y = a0 + a1\ x,$$

where, a0 and a1 are the constant determining the position of the line.

The slope of the line is denoted by a1 and intercept on the y-axis is given by a0. This relationship is called Linear Regression of y on x. Here a1 is called regression coefficient of y on x. Similarly, the Linear Regression of x on y is given by equation

$x = b0 + b1y$,

where, b0 and b1 are constants. The constant b1 is called regression coefficient of x on y and the constant b0 is the intercept on x- axis.

It may be that all the points do not lie on the straight line equation which represents the linear relationship between two variables. The straight line equation shows the average relationship between the values of two variables. Non- Linear Regression is the ones in which there exists a relationship of degree more than one between the variables under study. The straight line depicting the relationship between two variables is called Line of Regression or Regression Line. A Line of Regression is the one which gives the best fit using Principal of Least Square.

13.6 Estimating Linear Regression

The estimation or prediction of future production, consumption, prices, investments, sales, profits, income etc. are of very great Importance to business professionals. Similarly, population estimates and population projections, GNP, Revenue and Expenditure etc. are indispensable for economists and efficient planning of an economy. Regression estimates the changes in the value of dependent variables with respect to some values of independent variable. These estimation are done by the help of regression lines which are polynomials of degree one, having two variables.

- **Equation of Regression Lines**

In order to obtain the equation of regression line, we will be using the Method of Least Squares. By the help of this method, we will obtain the best fitting line to the set of points on the scatter diagram.

Suppose, the equation of line of best fit of x, be

$$y = a + bx \qquad \text{---- (1)}$$

where, a and b are constants.

Let X and Y be the deviation from the respective means M1 and M2 i.e.

$$X = x - M1 \ , \quad Y = y - M2 \qquad \text{---- (2)}$$

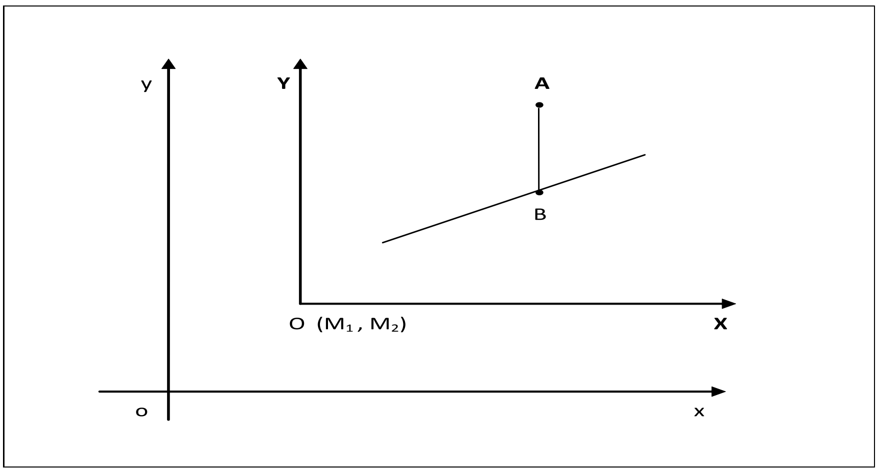
where, M1 and M2 are mean of x- series and y-series, respectively.

On shifting the origin to (M1,M2), let the equation (1) takes the following form:

$$Y = a + bX \qquad \text{---- (3)}$$

Let A(Xr,Yr) be any dot and B be point on the line (see the figure). Then,

$$AB = Yr - a - bXr \qquad \text{---- (4)}$$



If S be the sum of squares of such distances, then

S = ?(Yr - a -bXr) 2, for all the values of r ---- (5)

The principal of Least Squares says that sum of the squares of such distances be minimized. For this we have to choose those values of a and b for which this sum S is minimum. These values of a and b are obtained by equations

Using equation (5) and (6), we obtain,

$$\frac{\partial S}{\partial a} = 0 \quad and \quad \frac{\partial S}{\partial b} = 0 \qquad ---(6)$$

$$\frac{\partial S}{\partial a} = -2\sum(Y - a - bX) = 0 \qquad ---(7)$$

$$\frac{\partial S}{\partial b} = -2\sum X(Y - a - bX) = 0 \qquad ---(8)$$

On simplifying the equations (7) and (8), we obtain,

$\sum Y - na - b\sum X = 0$ and $\sum XY - a\sum X - b\sum X^2 = 0$ ----- (9)

As $\sum X = 0 = \sum Y$, from the equation (9), we get

$$a = 0 \text{ and } \quad b = \frac{\sum XY}{\sum X^2} = r\,\frac{\sigma_Y}{\sigma_X}$$

because, $$r\,\frac{\sigma_Y}{\sigma_X} = \frac{\sum XY}{\sqrt{\sum X^2 . \sum Y^2}} = \sqrt{\frac{\sum Y^2}{\sum X^2}}$$

Hence the Equation of line of best fit is

$$Y = r\frac{\sigma_Y}{\sigma_X}X$$

Going back to the original origin, we obtain the equation of regression line of y on x.

$$\boldsymbol{y - M_2 = r\frac{\sigma_y}{\sigma_x}(x - M_1)} \qquad ---(10)$$

Similarly, we can obtain the equation of regression line of x on y, which is:

$$\boldsymbol{x - M_1 = r\frac{\sigma_x}{\sigma_y}(y - M_2)} \qquad ---(11)$$

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$ are known as regression coefficient of y on x and x on y, respectively.

Example 4

Marks obtained by a student in Physics and Maths (out of 100) are given in the following Table:

Physics(x)	80	45	55	56	58	60	65	68	70	75	85
Maths(y)	82	56	50	48	60	62	64	65	70	74	90

Find the equations of lines of regression.

Solution:

Take, X = x - 65 and Y = y - 70

Calculation Table

Physics(x)	Maths(y)	X=x-65	Y = y-70	X^2	Y^2	XY
80	82	15	12	225	144	180
45	56	-20	-14	400	196	280
55	50	-10	-20	100	400	200
56	48	-9	-22	81	484	198
58	60	-7	-10	49	100	70
60	62	-5	-8	25	64	40
65	64	0	-6	0	36	0
68	65	3	-5	9	25	-15
70	70	5	0	25	0	0
75	74	10	4	100	16	40
85	90	20	20	400	400	400
$\sum x=717$	$\sum y=721$	$\sum X= 2$	$\sum Y= -49$	$\sum X^2=1414$	$\sum Y^2=1865$	$\sum XY=1393$

First of we will calculate the standard deviation of x and y i.e. σx *and* σy and coefficient of correlation 'r'.

Substituting the values from the Calculation Table in the formulas for coefficient of correlation and standard deviation, we obtain,

$$\sigma_x = \sqrt{\frac{\sum X^2}{n} - \left\{\frac{\sum X}{n}\right\}^2} = \sqrt{\frac{1414}{11} - \left\{\frac{2}{11}\right\}^2} = \mathbf{11.34}$$

$$\sigma_y = \sqrt{\frac{\sum Y^2}{n} - \left\{\frac{\sum Y}{n}\right\}^2} = \sqrt{\frac{1865}{11} - \left\{\frac{-49}{11}\right\}^2} = \mathbf{12.24}$$

$$r = \left[\frac{n\sum XY - \sum X.\sum Y}{\sqrt{\{n\sum X^2 - (\sum X)^2\}.\{n\sum Y^2 - (\sum Y)^2\}}}\right]$$

Thus,

$$r = \left[\frac{11 \times 1393 - (2) \times (-49)}{\sqrt{\{11 \times 1414 - (2)^2\}.\{11 \times 1865 - (-49)^2\}}}\right] = \mathbf{0.92}$$

Now as regression coefficient of x on y is bxy and y on x is byx we have

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{(0.92).(11.34)}{(12.24)} = \mathbf{0.85}$$

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{(0.92).(12.24)}{(11.34)} = \mathbf{0.99}$$

If the mean of x-series and y-series are M1 and M2, respectively, then

$$M_1 = 65 + \frac{\sum X}{n} = 65 + \frac{2}{11} = \mathbf{65.2}$$

$$M_2 = 70 + \frac{\sum Y}{n} = 70 + \frac{(-49)}{11} = \mathbf{65.55}$$

We know that the equation of regression line of y on x is given by

$$y - M_2 = r\frac{\sigma_y}{\sigma_x}(x - M_1)$$

Hence, **the equation of regression line of y on x is**

y – 65.55 = (0.99).(x – 65.2)

i.e.

y = (0.99).x + 1.002 **Ans**

Also, the equation of regression line of y on x is

$$x - M_1 = r\frac{\sigma_x}{\sigma_y}(y - M_2)$$

Therefore, **the equation of regression line of x on y is**

x – 65.2 = (0.85).(y – 65.55)

i.e.

x = (0.85).y + 9.48 **Ans**

- **Properties of Regression Coefficient**

The Coefficient of Correlation is the Geometric Mean of the Coefficient of Regression.

We Know that

$b_{yx} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

where, b_{yx} and b_{xy} are coefficient of regression of y on x and x on y, and r is coefficient of correlation

Hence,

$$b_{xy} \times b_{yx} = r\frac{\sigma_x}{\sigma_y} \times r\frac{\sigma_y}{\sigma_x} = r^2$$

i.e.

$$r = \sqrt{b_{xy} \times b_{yx}}$$

2. Arithmetic Mean of coefficient of regression is greater than the coefficient of correlation.

We Know that

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} \text{ and } b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

where, b_{yx} and b_{xy} are coefficient of regression of y on x and x on y, and r is coefficient of correlation

Hence, the arithmetic mean of b_{yx} and b_{xy} is

$$\left[\frac{b_{xy} + b_{yx}}{2}\right] > r$$

if,

$$\left[\frac{b_{xy} + b_{yx}}{2}\right] = \frac{r}{2}\left[\frac{\sigma_x}{\sigma_y} + \frac{\sigma_y}{\sigma_x}\right] > 0$$

i.e. $[\sigma_{xy}^2 + \sigma_{yx}^2] > 2r[\sigma_x . \sigma_y]$

i.e. $(\sigma_x - \sigma_y)^2 \geq 0$, which is true. Hence equation (1) is true.

3 The coefficient of regression is independent of change of origin but not of scale.

4 The coefficient of regressions cannot simultaneously greater than unity and neither can their product. But, one of them can be greater than the other, individually.

5 If ? is the acute angle between the two regression lines of two variables x and y, then

$$\text{tana} = \left[\frac{1 - r^2}{r}\right] . \left[\frac{\sigma_x . \sigma_y}{\sigma^2_x . \sigma^2_y}\right]$$

13.6.1 Difference Between Correlation And Regression

- Correlation coefficient is a symmetrical function between x and y but the regression coefficient are not symmetrical functions between x and y.
- Correlation coefficient is independent of change scale and origin, whereas, the regression coefficients are independent of the origin but not of scale.
- Correlation coefficient is a relative measure of relationship, whereas, regression coefficient is an absolute measure of relationship.

13.7 Summary

Correlation and Regression are important statistical tools for measuring the degree of association between series of pairs of observations of two or more variables. These tools are used for studying and measuring the extent of relationship between two or more variables. Correlation is the possible relationship between two series of observations such that the changes in the values of one series are accompanied by changes in the values of other series. Two variables are said to be positively correlated if an increase (or decrease) in the value of one variable is accompanied by an increase (or decrease) in the value of other variable. Two variables are said to be negatively correlated if the increase (or decrease) in the value of one variable is accompanied by decrease (or increase) in the value of other variable. Coefficient of Correlation shows the strength (magnitude) and direction (direct or indirect) of the relationship between two variables. It stands for relative measure of relationship between two variables.

Regression is used to measure the average relationship between the correlated variables. It uses the known value of one variable to predict or estimate the unknown value of the other variable. Regression stands for absolute measure of relationship. In Linear Regression, only two variables are involved in a relationship. The Method of Least Squares is used to find the regression lines, which gives best fit for the set of observations. By the help of regression lines, we are able to estimate the value of x by the value of y and vice versa.

13.8 Key Words

- **Bivariate Distribution -** a distribution in which each individual of the set is made up of two values.
- **Scatter Diagram -** a diagram of plotted points that show relationship between two sets of data.
- **Linear Correlation-** Correlation between two variables in which relationship is in the form of polynomial of one degree.
- **Coefficient of Correlation -** measure the intensity of relationship between two variables.
- **Coefficient of Regression -** indicate rate of average of one variable with respect to the other variable.

13.9 Self Assessment Test

1. Calculate the coefficient of correlation for the following observation:

X	5	7	8	4	9	3	2	5	4	3
Y	2	4	5	5	6	5	4	4	3	2

(Ans: 0.47)

2. Calculate the coefficient of correlation between the values of x and y from the following:

X	1	3	5	7	8	10
Y	8	12	15	17	18	20

(Ans: 0.98)

3. Eight student obtained the following marks (out of 100) in Maths and Chemistry.

Marks in Maths	15	13	27	45	20	60	20	75
Marks in Chemistry	50	30	55	25	30	10	30	70

Calculate the rank correlation coefficient.

(Ans: 0.02)

4. Heights of fathers and sons (in inches) are given in the following table:

Heights of father	65	66	67	67	68	69	71	73
Heights of Son 67	68	64	68	72	70	69	70	

Obtain the two lines of regression and calculate the expected average height of the son when the height of father is 67.5 inches.

(Ans: $y = 0.421x + 39.29$: $x = 0.524y + 39.29$; height of son is 68.19)

13.10 Refrences

- 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.

Unit - 14 : Computer Applcations in Research Methodology

Unit Structure:

14.0 Objectives
14.1 Introduction
14.2 Computers in Organisations
14.3 Computers and Research
14.4 Computers in Research Methodology
14.5 Online Research
14.6 Limitations of Computers in Research
14.7 Summary
14.8 Key Words
14.9 Self Assessment Test
14.10 References

14.0 Objectives

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

- Role of computers in business organizations.
- Applications of computers in research methodology.
- What is online research.
- Limitations of using computers in research.

14.1 Introduction

Computers are contributing to researches in every discipline, ranging from biology to astrophysics, discovering new patterns and providing novel insights. They help us in researches related to business field too. Perhaps the earliest use of technology in business research was when researchers first used tape recorders in their field to record interview sessions. Having a recording and a transcript meant that new ways of thinking about how the analysis developed out of the data and how the analysis was supported by the data became possible. Second, it allowed different kinds of analysis that could only be undertaken if accurate records of the speech were kept. It also opened up the possibilities of much larger scale studies and the use of multiple researchers and analysts.

The dual impact of new technology both on what kinds of data can be collected and recorded and on what kinds of analysis it makes possible has continued to the present day. In 21st century, new ways of recording and collecting data, and new ways of undertaking the analysis using computers has become indispensable for researches.

The development of information technology and particularly the growth of the Internet has created not only new ways in which researchers can analyse their data, but also created whole new areas from which data can be collected and ways in which it can be collected. The former include discussion lists, text forums, personal web pages and video-conferences. When the work in neural network based artificial intelligence advances and computers are granted with the ability to learn and think for them, future advances in technology and research will be even more rapid.

14.2 Computers in Organisations

Computers have tremendously improved the way businesses are done world over. Technology has advanced so remarkably that those who are not using computers in their business are at a major disadvantage against their competitors. In particular, there are several important advantages that computers can provide to any business.

In general computers allow the application of different types of software that can help businesses keep track of their files, documents, schedules and deadlines. Computers also allow businesses to organize all of their information in a very accessible manner. The ability to store large amounts of data on a computer is convenient and inexpensive, and saves space. A computer's ability to allow a company to organize its files efficiently leads to better time management and productivity.

Computers have made staff and companies more self-sufficient by allowing them to do tasks that previously had to be outsourced. For example, a company can now use office software to create their own training material. Desktop publishing software can be used to create marketing materials. Online tax and accounting programs allow companies to prepare their own taxes. This allows the dominant operations of a company to remain in-house and empowers the company to become more independent and less susceptible to errors committed by outside parties.

Emerging technology makes new tools and services more affordable and allows companies to save on their staff payroll and office equipment. Because computers allow work to be done faster and more efficiently, it is possible for a company to hire fewer staff. In addition, with networked and relatively inexpensive computers, companies can store data more easily, saving on the cost of outside file storage, and can avoid having to purchase as many copiers, fax machines, typewriters, and other such items that were used before computers became popular. Correspondingly, potentially profitable businesses can be started with a smaller overhead cost. Email capabilities decrease postage costs; software applications reduce the need for large accounting departments, while videoconferencing reduces the need for travel. All resources saved trickle down to the consumers, who are then provided with much more affordable products and service. Computers help speed up other business operations. The collecting of consumer feedback, ordering of raw materials, and inspection of products is made quicker through the use of computers, allowing companies to operate much faster and to produce better quality results.

Research and development costs also decrease with the help of computers. Scientific research can now be done using the Internet and computer software applications designed to develop and produce new products and services. For example, instead of a company having to do in-person focus groups on a potential new product or to determine their target market, the company can conduct a widespread online survey for a far lower cost. In addition, new models of a product can be created online using virtual pictures and drawings instead of having to be hand-drawn. These interactive models created using software programs can help bring the product and its features to life for a far lower cost than creating an actual physical model of the given product. Consumer Generated Media (CGM) – Information left by individuals online in the form of blogs or websites are also very useful.

Computers can help generate higher sales and profits for businesses via a company website. Many businesses now operate online and around the clock to allow customers from around the world to shop for their products and services.

In an office fully equipped with a computer network communication can be made easy via the company's intranet as well as the Internet. An intranet is extremely valuable. Sensitive or confidential information can be handled within the confines of your company's network without concern that unauthorized users will access your organization's private network. Computers are also necessary for organizations with employees who are telecommuters or stationed remotely. In these circumstances, the computers issued to remote employees remain company property and are returned when the employee leaves the company. Companies sometimes implement policies that prohibit the use of company-issued computers for personal use, to safeguard company information. Business computers are used for producing correspondence, processing accounting tasks and performing research functions. A vast number of software applications are available for an organization or business process of virtually any size. The advantages of computers for conducting general business transactions are numerous, one of which is ensuring higher levels of accuracy for functions previously performed by staff.

As technology continues to advance, increasing numbers of tasks will be done by computer. Like in the area of research. Using computers and the Internet, it is possible to access a wider library of information at a faster rate. Computer research also has its flaws, but by learning how to do it properly, you will be able to get information efficiently.

14.3 Computers and Research

Computers have a very important role to play in research activities. It has become an essential tool for research, whether it is for academic purpose or for commercial purpose. You can find all kinds of information on the Internet and you can even discuss research problems with people around the world. Computers have led the way to a globalised information portal that is the World Wide Web. By using www we can conduct primary as well as secondary research on a massive scale. Various computer programs and applications have eased our way into compiling our research data. Many statistical and other tools help us to organize and handle quantitative as well as qualitative data. Inference and analysis is also easier to make by using a computer.

It is truly impossible to capture the whole range of computer applications in scientific research. However, three important applications can be described:

a) **Data Storage and Analysis** – Data are the life blood of research. Every research generates a lot of data that needs to be stored and analyzed to derive important conclusions and to validate or disprove hypothesis. Computers also help in recording data and subject it to analysis through specially designed software. Data from different sources can be stored and accessed via computer networks.

   Audio recording is an analogue technology, as are film and traditional video. There is a long history of their use in many areas of social and psychological research. Recent changes in this type of technology have taken several forms. It has become cheaper and more widespread. Not only new technology is to researchers, but also that the people being researched are more used to being recorded and are even familiar with using it themselves. The spread of video and photographic technology means that images can be used both as sources of data and as tools for data collection. The data in digital form, audio and video, now makes possible new ways of creating, storing, and processing. Automated Speech Recognition (ASR) tool help in recording human voice in digital voice and then convert the same in written words.

   Analyzing tons of statistical data is made possible using specially designed algorithms that are implemented by computers. This makes the extremely time-consuming job of data analysis to be a matter of a few minutes. Word processors and databases are being frequently used in the analysis of data. Various statistical tools are also available to anaylse and interpret data.

Some programs have functions that go well beyond manipulating, searching and reporting on coded text. They assist with analytic procedures by providing a variety of facilities to help the analyst examine features and relationships in the texts. Such programs are often referred to as theory builders or model builders, not because on their own they can build theory, but because they contain various tools that assist researchers to develop theoretical ideas and test hypotheses. Another advantage of using software is that the analysis is structured and its progress can be recorded as it develops.

Already some computer software like HyperRESEARCH, ATLAS.ti, CI-SAID, MAXqda2, QSRN6 and the Qualitative Analyser allow researchers to code images, digitise speech and video. Some, like NVivo and ATLAS.ti allow video segments to be hyperlinked in a limited way. Such programs are close to providing the richness and fine details available to text coding for the coding of sound and video. These software allow the import of RTF, AIFF, WAV, PIC, GIF and MPEG files and Tatoe and ATLAS.ti are using XML and HTML as a medium for exporting text data files.

Software like ATLAS.ti and NVivo include facilities for visual and hierarchical modelling of concepts and codes. Others take a much more numerical and logical approach to modelling, often built around a hypothesis testing or case-based approach (as opposed to a code-based approach). Examples include HyperRESEARCH, Ethno and AQUAD Five.

b) **Scientific Simulations -** One of the prime uses of computers in research specially causal research is the running of simulations. A simulation is a mathematical modeling of a problem and a virtual study of its possible solutions. Problems which do not yield themselves to experimentation can be studied through simulations carried out on computers.

c) **Knowledge Sharing -** In the form of Internet, computers have provided an entirely new way to share knowledge. Today, any one can access the latest research papers that are made available for free on websites. Sharing of knowledge and collaboration through the Internet has made international cooperation on research projects possible.

14.4 Computers in Research Methodology

Research process consists of series of steps necessary to effectively carry out research. Some of these steps overlap continuously rather than following a strictly prescribed sequence. These steps are not mutually exclusive and are not separate and distinct. However, the order discussed below provides a useful procedural guideline. Computers help us in almost every step. The steps of research process are:

(a) **Defining and formulating the research problem** – Problem definition is the most critical part of the research process. Unless the problem is properly defined, the information produced by the research process is unlikely to have any value. It was Albert Einstein who noted that, "the formulation of a problem is often more essential than its solution". This is good advice for managers. Too often managers concentrate on finding the right answer rather than asking the right question. Many don't realise that defining a prob1em may be more difficult than solving it. In business research, if the data are collected before the nature of the business problem is carefully thought out, the data probably will not help solve the problem.

The adage "a problem well defined is half solved" is worth remembering. This adage emphasizes that an orderly definition of the research problem gives a sense of direction to the investigation. Careful attention to problem definition allows the researcher to set the proper research objectives.

If the purpose of the research is clear, the chances of collecting the necessary and relevant information without collecting surplus information will be much greater.

Research problem definition involves four interrelated steps: (i) management problem / opportunity clarification; (ii) situation analysis; (iii) model development; and (iv) specification of information requirements. Computers help us in all these phases.

- **Management problem/opportunity clarification –** the basic goal of problem clarification is to ensure that the decision 'makers' initial description of the management decision is accurate and reflects the appropriate area of concern for research. If the wrong management problem is translated into a research problem, the probability of providing management with useful information is low. Company' internal computer databases and MIS system help in giving more clarity to the problems.

- **Situation analysis –** The management problem can be understood only within the context of the decision situation. The situation analysis focuses on the variables that have produced the stated management problem or opportunity. It involves giving careful attention to company records; appropriate secondary sources such as census data, industry sales figures, economic indicators, and so on; and interviews with experts both external and internal. Computers are needed to scan all these.

- **Model development –** Once the researcher has a sound understanding of the decision situation, it is necessary to get as clear an understanding as possible of the situation model of the problem. A situation model is a description of the outcomes that are desired, the relevant variables, and the relationships of the variables to the outcomes. Computer software help us in building various kinds of models.

- **Specification of information requirements –** Research can't provide solutions. Solutions require executive judgment. Research provides information relevant to the decisions by the executive. The output of the problem definition process is a clear statement of the information required to assist the decision-maker. A common temptation is to try to collect data on all possible variables. Unfortunately, this is generally impractical and always costly. The best approach for ensuring that any data collected is indeed relevant is to ask questions concerning the probable use of the data. Specifically, the researcher should list the findings that seem possible. Computers help us in generating timely, accurate and reliable information.

(b) **Estimating the value of the information –** The principle involved in deciding whether to do research is that the research should be conducted only when it is expected that the value of the information to be obtained will be greater than the cost of obtaining it. Cost and value of information also helps the research department in determining which research projects to conduct, which research design to use, amid whether to gather more information after the initial results are in. Two approaches can be taken to arrive at an assessment of whether the expected value of the information in a proposed research project is greater than at estimated cost; the intuitive and the expected value approaches to the problem.

- **The intuitive approach –** It relies entirely on the private judgment of the person making the assessment.

- **The expected value approach –** It uses application of Bayesian statistics that allow judgmental probabilities to be used.

The use of computers has become essential specially for expected value approach. Computers help us in conducting cost benefit analysis.

(c) Selecting the data collection approach – There are three basic data collection approaches in research: (i) secondary data (ii) survey data, and (iii) experimental data. Secondary data are collected for some purpose other than helping to solve the current problem but may be used for research purpose, whereas primary data are collected expressly to help solve the problem at hand. Survey and experimental data are, therefore, secondary data if they were collected earlier for another study; they are primary data if they were collected for the present one. Secondary data are virtually always collected first because of their time and cost advantage. Companies generally rely on both primary as well as secondary data and computers help us in collecting both.

When the data collection approach is finalised, researcher decides about the contact methods that will be using for gathering data. The following contact methods may be used:

- **Mail Questionnaire** – The mail questionnaire is the best way to reach individuals who will not give personal interviews or whose responses might be biased or distorted by the interviewers or who are scattered at distant places. E-mail is extensively used to send the designed questionnaire to the respondents. It saves on postal and printing charges.
- **Telephone Interviewing** – It is the method for gathering information quickly. The interviewer is also able to clarify questions if they are not understood. The response rate is typically higher than in the case of mailed questionnaires. The two main drawbacks of this technique of data collection are that only people with telephones can be interviewed, and the interviews have to be short and not too personal. Computers linked with telephones aid any researcher in collecting data. CATI (Computer Assisted Telephone Interviewing) and CAMI (Computer Assisted Mobile Phone Interviewing) are some of the tools which are being used in researchers now-a-days.
- **Personal Interviewing** – It is the most versatile of the three methods. The interviewer can ask more questions and can record additional observations about the respondent such as dress and body language. Personal interviewing is the most expensive method and requires more administrative planning and supervision. It is also subject to interviewer bias and distortion. CAPI (Computer Assisted Personal Interviewing), CASI (Computer Assisted Self-completion Interviewing), CAWI (Computer Assisted Web Interviewing) are some of the tools which are being used in researchers now-a-days.

Personal interviewing takes two forms, arranged interviews and intercept interviews. In **arranged interviews**, respondents are randomly selected and are either telephoned or approached at homes or offices and asked to give an interview. **Intercept interviews** involve stopping people at a shopping centre or busy street corner and requesting an interview. Intercept interviews have the drawback of being non-probability samples, and the interviews must be quite short. Tape recorders and video cameras linked to a computer system is frequently used for this purpose.

(d) Selecting the measurement technique – There are five basic measurement techniques in a research: (i) questionnaire (ii) observation (iii) Interviews and projective techniques (iv) focus group research and (v) mechanical/laboratory instruments.

(e) Selecting the sample – Most studies including marketing studies involve a sample or subgroup of the total population relevant to the problem, rather than a census or the entire group. The population is generally specified as a part of problem definition process. Sampling involves the decisions about the following:

(i) Population – determines who (or what objects) can provide the required information.

(ii) **Sample frame –** develops a list of population members.

(iii) **Sampling unit –** determines the basis for drawing the sample - individuals, households, city blocks, etc.

(iv) **Sampling method –** choose between probability and non-probability methods. Probability sampling include simple random sampling, stratified random, sampling and cluster sampling whereas non-probability sampling include convenience sampling, judgment-sampling, and quota sampling.

(v) **Sample size –** determines how many population members are to be included in the sample. Usually, large samples give more reliable results than small samples.

(vi) **Sample plan –** develops a method for selecting and contacting the sample members.

(vii) **Execution –** carries out the sampling plan.

We can decide all of the above using suitable statistical software packages.

(f) **Select the method (s) of analysis –** Data are useful only after analysis. Data analysis involves converting a series of recorded observations into descriptive statements and/or inferences about relationships. The analysis of data requires a number of closely related operations such as establishment of categories, coding, tabulation etc. The unwieldy data should necessarily be condensed into a few manageable groups and tables for further analysis. Data dredging and data mining tools come handy in this stage.

Software packages such as SPSS, Micro stat etc. help us in the analysis of data. Many statistical tools and techniques such as Correlation and Regression, Tests of Significance, ANOVA, Factor analysis, Conjoint Analysis, Cluster Analysis, Randomised block designs, Latin Square Designs, Multi-dimensional scaling, Perceptual mapping etc. can be done with the help of these software. Even simple analysis like mean, averages etc can be done using MS-Excel etc. Various kinds of graphs and charts can be made using computers.

(g) **Dissemination of results/Report preparation –** Results of any research are conveyed with the help of various types of reports. These reports are first written using various word processors and spread sheets like MS-Word, MS-Excel, Adobe-PageMaker etc. Reports are then presented to decision makers either in printed form or using presentations made on MS-PowerPoint.

14.5 Online Research

Online research is possible if respondents are connected by computer networks or tele-communication systems. Computer Assisted Web Interviewing (CAWI) is the most widely used approach and is simply a questionnaire that appears as a set of web pages to be answered by internet users.

Online research is live, connected in real time, and there is little if any delay in transferring data from the point of collection to the point of analysis. It is usually self completion. Internet research, web research are other names of online research. Computer Assisted Web Interviewing (CAWI), Computer Assisted Mobile Interviewing (CAMI), Audience Response System (ARS) and netnography have further increased the domain of online research.

Many types of researches can be done using online mode. They can be used to test promotion's success or failure, consumer behaviour, consumer perception and satisfaction related to products and services etc.

Web based questionnaire and CAMI, CAWI etc. come handy for this purpose. These researches are quick and save costs in comparison to traditional researches. Large respondents can be approached in no time. However, traditional methods have more control especially in interviewing and sampling. Online questionnaires, however, are normally better than traditional counterparts as they have unique features which include filtering, routing and piping; progress indicators; radio buttons; check boxes; drtop down menus; free text input; slider bars and so on.

14.6 Limitations Computers in Research

There are so many advantages of using computers in research. They increase accuracy and saves time, effort and money. However, there are some limitations of using computers in research.

One of the limitation we can identify is a feeling of being distant from the data. Researchers using paper-based analysis felt they were closer to the words of their respondents or to their field notes than if they used computers. It was certainly true that some of the early software made it hard to track back from extracted text to the context in the original documents from which it came. But most programs now emphasise their facilities for the recontextualisation of data. However, the feeling still remains that we are too much dependent on these computers.

Second limitation is that computers are only a tool. Decision making rests on the interpretation of analysed data which is entirely the domain of decision maker. Computer use is limited in how far they can help with such an interpretation.

Another limitation is that the strength of the analysis depends to a large extent on the well-established strategies used in analysing research data. Choice of appropriate tool is in the hands of researcher and not computers. Computers will accurately do whatever they are asked to do. They can not point out the validity of selected tool. It is just a tool for analysis, and good analysis still relies on good analytic work by a careful human researcher, in the same way that good writing is not guaranteed by the use of a word processor.

14.7 Summary

Computers have a very important role to play in research activities. It has become an essential tool for research whether it is for academic purpose or for commercial purpose. You can find all kinds of information on the Internet and you can even discuss research problems with people around the world. Statistical tools help us to organize data and handle quantitative as well as qualitative data. Computers were initially used in business research to collect data and perform simple analysis. However, the introduction of specialized packages like SAS, Stata and SPSS has enabled automated data analysis and output generation. This has resulted in increased accuracy and savings in terms of time, effort and money.

14.8 Key Words

- **Automated Speech Recognition (ASR)** – A method of recording data by capturing human voice in a digital form and converting it to another medium such as written word.
- **CATI** - Computer Assisted Telephone Interviewing.
- **Consumer Generated Media (CGM)** – Information left by individuals online in the form of blogs or websites.
- **Database** – A Collection of data in organized form which can be processed to provide accurate, timely and relevant information.

- **Data Mining** – The procedure that selects and manipulates large amounts of data to uncover previously unknown relationships and patterns.
- **Observation** – A method of primary data collection that involves seeing, tracking, or sensing behaviour or actions in some way.
- **Search Engines** – Provide users with a way of locating and retrieving information from documents located on the internet.

14.9 Self Assessment Test

1. Discuss the uses and application of computers in business research
2. Discuss the role computers play in various stages of research methodology.
3. With the help of suitable examples, discuss the advantages and limitations of using computers in research.
4. 'Modern day business research is not possible without using computers.' Comment.

14.10 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", Pearson Education Publications.
- Naresh Malhotra, "Marketing Research: An applied Orientation", Prentice Hall International Edition.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- Nigel Bradley, "Marketing Research – Tools and Techniques", Oxford University Press, New York.

Unit - 15 : Report Writing

Unit Structure:

15.0 Objectives
15.1 Introduction
15.2 Types of Reports
15.3 Functions of a Research Report
15.4 Principles of Report Writing
15.5 Structure of Report
15.6 Model of Report
15.7 Formatting Guidelines
15.8 Organization of a Research Report
15.9 Summary
15.10 Key Words
15.11 Self Assessment Test
15.12 References

15.0 Objectives

After studying this unit you should be able to:

- Understand the meaning and importance of a research report
- Classify various types of reports
- Understand the process of report writing
- Understand the ways of presenting references, citations etc used in the research
- Understand the key features of report formatting

15.1 Introduction

The purpose of a research is to draw inferences and come out with suggestions. Therefore the outcome of the research should be well documented in the form of a research report for implementation and future use. The research task will remain incomplete till the report has been well documented, written and presented.

A research report is considered to be one of the most important components of the research study. In fact it would not be wrong to say that even the most brilliant of the research studies are of little value unless they are effectively presented and communicated to others. All this explains the significance of writing research report.

A report is a statement of the results of an investigation or of any matter on which definite information is required. A report is a clearly structured document in which the researcher identifies and examines issues, events, or findings of a research. It forms a major component of the research study. A research report is "a written document or oral presentation that communicates the purpose, scope, objective(s), hypotheses, methodology, findings, limitations and finally, recommendations of a research project to others."

The basic aim of the report is:

- To communicate the objectives of the work done.

- To give findings, analyses of findings, interpretations, conclusions and recommendations.
- To highlight important details regarding the research process
- It gives the readers certain insights about the research and its background.
- Improves interpretation skills.
- Written reports give scope for future references.

15.2 Types of Reports

The form and structure of the research report might change according to the purpose for which it has been designed. Based on the size of the report, it is possible to divide the report into the following types:

- **Brief Reports**

These kinds of reports are not formally structured and are generally short. The information provided is of a limited scope and is prepared either for immediate consumption or as a prelude to the formal structured report that would subsequently follow. These reports could be designed in several ways.

Working papers or basic reports are written for the purpose of collating the process carried out in terms of scope and framework of the study, the methodology followed and instrument designed. The results and findings are also recorded here. The study background might be missing, as the focus is more on the present study. These reports are significant because they serve as a reference point while preparing the final report.

Survey reports focus to communicate findings in easy-to-comprehend format that includes figures and tables. The reader can then study the patterns in findings to arrive at appropriate conclusions, beneficial for resolving the business dilemma. The advantage of these reports is that they are simple and easy to understand and present the findings in a clear and usable format.

- **Detailed Reports**

These are more formal and pedantic in their structure and are essentially academic, technical or business reports. Sometimes, the researcher may prepare both kinds-for an academic as well as for a business purpose. The language, presentation and format of the two kinds of reports would be different as they would need to be prepared for the understanding of the reader's capabilities and intentions.

- **Technical Reports**

These are major documents and would include all elements of the basic report, as well as the interpretations and conclusions, as related to the obtained results. This would have a complete problem background and any additional past data/records that are essential for comprehending and interpreting the study output. All sources of data, sampling plan, data collection instruments, data analysis outputs would be formally and sequentially documented.

15.3 Functions of Research Report

A well – written research report performs several functions:

1. It serves as a means for presenting the problems studied, methods and techniques used for collecting and analyzing data, the findings, conclusions and recommendations in a organized manner.
2. It serves as a basic reference material for future use in developing research proposals in the same or related area.

3. A report serves as a means for judging the quality of the completed research project.
4. It is a means for evaluating the researcher's ability and competence to do research.
5. It provides factual base for formulating policies and strategies relating to the subject matter studied.
6. It provides systematic knowledge on problems and issues analysed.

15.4 Principles of Report Writing

The writing of a research report is governed by certain principles of standard practices. These are described below:

- A research report requires clear organization. Each chapter may be divided into sections with appropriate headings and in each section margin headings and paragraph headings may be used to indicate subject shifts. Headings serve as pointer to where the discussion is going.
- Standard usage in paragraphing and giving headings to topics and subtopics should be followed. The general grouping is form "Centered section headings" to "marginal heading" to "paragraph heading" to "individual paragraphs".
- Physical presentation is another aspect of organization. A page should not be fully filled in from top to bottom. Wider margins should be provided on both sides and on top and bottom as well.
- A research report requires a style different form other academic writings like essay, fiction or poetry. In the latter, the richness of words usage is often desirable to achieve beautiful effects. But a research report, being a formal presentation of an objectives, unbiased investigation should be written at a formal level of standard English.

15.5 Structure of Report

The following common elements are found in nearly all types of reports:

- **Title page / Cover Page**

This should include the title of the report, the author's name, name of the guide if any, course, date (if required) and affiliations (Name of university or college).

- **Acknowledgements**

Every body that has helped in carrying out the project should be acknowledged in this segment. For example the help may come from librarians, technicians or computer centre staff persons etc.

- **Contents**

This should include all the main sections of the report in sequence with the page numbers they begin on. If there are charts, diagrams or tables included in the report, they should be listed separately under a title such as 'List of Illustrations / Diagrams' together with the page numbers on which they appear.

- **Abstract or Summary**

It should be a short paragraph summarizing the main contents of the report. It should include the main tasks, the methodology used, conclusions reached and any recommendations made or suggestions given. The abstract or summary should be concise, informative and independent of the report. This section should be written after finishing the detailed research report.

• **Introduction**

This should give the context and scope of the report and should include the terms of reference. It should state the objectives clearly, define the limits of the report, outline the method of enquiry, give a brief general background to the subject of the report and indicate the proposed development.

• **Methodology**

This section talks about the research methodology adopted for the research. In this section one should state how the enquiry was carried out. What form did the enquiry take? How the data collection was done? What kinds of measurement tools or scales were used? How the sample was selected? All such and other information related to the methodology should be presented and logically and concisely.

• **Results or Findings**

The results or findings should be presented in as simple a way as possible. The following ways can be used. There are a number of ways in which results can be presented like; Tables, Graphs, Pie charts, Bar charts, and Diagrams,

Following things should be ensured in case of illustrations or diagrams

- Are the diagrams / illustrations clearly labeled?
- Do they all have titles?
- Is the link between the text and the diagram clear?
- Are the headings precise?
- Are the axes of graphs clearly labeled?
- Can tables be easily interpreted?

• **Discussion**

This is the section where the researcher analyzes and interprets the results drawing from the information collected during the research. In the discussion important issues are identified and discussed. Findings are also explained during the discussion.

• **Conclusions and Recommendations**

This is the section of the report which draws together the main issues. It should be expressed clearly and in easy language. If required the recommendations can be listed separately.

• **References / Bibliography**

It is important that precise detail of all the work by other authors which has been referred to within the report is mentioned in this section. The references can be classified into journals, books, magazines, newspapers, research studies, etc. These are illustrated below:

• **Journal articles:**

The articles which are taken from journals should be included in the bibliography as per the following format.

Author(s), Title of the article, Name of the journal, Volume of the journal, Issue number of the volume, Year of publication, page numbers (from – to) of the article.

For example:

1. Panneerselvam, R. and C. Oudaya Sankar, 'New heuristics for assembly line bal-ancing problems', International Journal of Management and Systems, Vol. 9, No. I, 1993, pp. 25-36.
2. Shah, A.M., 'Environmental factors for strategy formulation', Productivity, Vol. 36, No.3, 1996, pp. 594-599.

If the number of authors is more than two, it is preferable to use et al. after the first author.

1. Panneerselvam, R. et al., 'Models for warehouse location problem', International Journal of Management and Systems, Vol. 6, 1990, pp. 1-8.

Books: If some of the items from a book are referred in a research, then following format is used

Author(s), Title of the book, Name of the publisher, Place of publication, Year of publication.

For example:

1. Gopalakrishnan, P. and M. Sundaresan, Materials Management: An Integrated Approach, Prentice-Hall of India, New Delhi, 1979.
2. Green, P.E. et al., Research for Marketing Decisions, Prentice-Hall of India, New Delhi, 1988.

Articles in a Book:

The format of an article in a book is as shown below:

Name of the author(s), Title of the article, Title of the book, Name of the Editor, Name of the publisher, Pages, Year of publication

For example:

1. Panneerselvam, R., 'Integrated business logistic-a tool for effective supply chain management', Supply Chain Management, Sahay, B.S. (Ed.), Macmillan India Limited, pp. 336-346, 1999

Government Publications

Ministry of Law, Government of India, The Copyright Act, 14 of 1957, Delhi, The Manager' of Publications, 1960, p. 10.

Conference Articles:

The format of an article in a conference proceeding is shown below:

Name of the author(s), Title of the article, Name of conference proceedings, Place of the conference, Dates-month-year of the conference, Pages if applicable

In some conferences, only abstracts of the articles will be published in which case pages are not applicable.

1. Pires, S.R.I. et al., 'Measuring supply chain performance', Proceedings of the Twelfth Annual Conference of the Production and Operations Management Society, POM-2001, Orlando, March 30-April 2, 2001.
2. Senthilkumar, P. et al., 'Algorithm for AGVS scheduling', Proceedings of 43rd National Convention of Indian Institution of Industrial Engineering, Chennai, September, 7-8, 2001, pp. 42-46. No. 298, p. 17, December 19, 2002

Articles in Newspapers

1. Preeti Mishra, 'Super computing and drug discovery', The Hindu, Vol. 125, No. 298, p. 17, December 19, 2002. .

2. Shankar, T.S., 'Aircraft maintenance: anticipate and manage', The Hindu, Vol. 125, No. 298, p. 17, December 19, 2002.

More than one item of author(s)

1. Panneerselvam, R., 'Efficient heuristics for total covering problem', Productivity, Vol. I 36, No.4, 1996,

   pp.649-657.

2. Production and Operations Management, Prentice-Hall of India, New Delhi, 1999.

Research Studies

1. Narayanan, S., 'Development of new efficient heuristic for deterministic assembly line balancing problem, Unpublished Ph. D. Thesis, School of Management, Pondicherry University, Pondicherry, November, 2000.

2. Panneerselvam, R., 'Algorithmic grouping of operation sequences as an aid to cellular production system', Unpublished Ph. D. Thesis, Industrial Engineering Division, Anna University, Chennai, August 1986.

References should be listed in alphabetical order of the authors' names.

Appendices

An appendix contains additional information related to the report but which might not be essential to the main findings. Appendices can be consulted if required but the report does not depend on this. This can include details of interview questions, statistical data, a glossary of terms, or other information which may be useful for the reader.

15.6 Model of Report

The main parts of a research report are:

• **The Preliminary Pages**

- The Title Page
- Researcher's Declaration
- Research Supervisor's Certificate
- Preface/Abstract/Summary
- Acknowledgements
- Contents
- List of Tables
- List of Figures
- List of Abbreviations

• **The Main Text**

Chapter 1: Introduction

1.1.1. Introduction of the Study
1.1.2. Theoretical/Conceptual Framework
1.1.3. Industry/Company Profile
1.1.4. Problem Formulation
1.1.5. Objectives of The Study

Chapter 2: Literature Reviews

This chapter includes the outcomes of the students in the present area of research

Chapter 3: Research Methodology
3.1.1 Scope of the Study
3.1.2 Research Title
3.1.3 Objectives of the Study
3.1.4 Research Design
3.1.5 Sampling Framework
3.1.6 Statistical Techniques for Data Analysis
3.1.7 Demographic Profile of the Respondents

Chapter 4: Data Analysis and Interpretations

Chapter 5: Findings of The Study

Chapter 6: Conclusion and Recommendations
6.1.1 Recommendations
6.1.2 Conclusion
6.1.3 Limitations of the Study

- **References**

Bibliography

Annexure

1 Questionnaire

2 List of Respondents

3 Other Supporting Materials

15.7 Formatting Guidelines

Research Project

1. **Page Size:** A -4
2. **Front Page-** As prescribed by the University
3. **Font:** Times New Roman
4. Same font should be used in one report/thesis/research paper.
5. **Font Size:**

    a. Main Headings: 16 pt- (Title Case)

    b. Sub Headings: 14 pt – (Title Case)

c. All Other Headings: 12 pt – (Title Case)

d. Main Text: 12 pt – (Normal/Sentence Case)

6. No Page Borders until specifically prescribed
7. **Text Color:** Automatic for text but diagrams may be colored.
8. **Paragraph Alignment:** Justified
9. **Line Spacing:** 1.5 line spacing throughout the cepart
10. **Margins:** 1.25" margin on the left side and 1" on all other sides viz. top, bottom and right.
11. Do not underline the text.
12. Text should be free from hyperlinks or unusual formatting, which usually comes when matter is copied from web pages
13. Do not use table format for the content page
14. Do not break table or diagram in two pages, if a table is too large follow:

    a. Break the table and write at the bottom of the table (where you break the table) – "continue to next page"

    b. Smaller the font size (This freedom is only in the case of tables and the font size should not be less than 9 pt in any case)

15. Divide the text in sections and present them with points. A illustration is presented below.

    2. Bank Selection Criteria

    2.1 Convenience

    2.1.1 New Age Technologies

    2.1.1.1 ATMs

    2.1.1.2 Internet Banking

16. Limit sections up to four points only.
17. Write table/exhibit title above the table/exhibit and figure title below the figure.

References for tables and figures are to be given below them only.

15.8 Organization of a Research Report

It does not require elegant word usage and allusion. It just need a plain discourse with accuracy, clarity, coherence, conciseness and readability.

(a) Accuracy is the first requirement of the report writing.

(b) Clarity is another requirement of presentation. It is achieved by using familiar terms and common words and unambiguous statement, explicitly defining concepts and unusual terms and by taking care of logical flow of ideas and sequence of sentences.

(c) Coherence is an essential part of clarity. Each sentence must be so linked with other sentences that the writer's thoughts move smoothly and naturally from one statement to the next.

(d) Guidelines for ensuring coherence :

(i) Arrange paragraphs in a logical sequence.

(ii) Place at the beginning of a paragraph the topic sentence to link it with the one before it and to prepare the reader for what is to follow.

(e) Clarity and readability call for avoiding the following types of unclear writings :

- Avoid the jargon, and pretentious and pompous
- Avoid offensive words
- Omit needless words that cause verbosity.
- Avoid superfluous phrases.
- Avoid words that exaggerate such as stupendous, immeasurable, gigantic, awfully, dreadfully.
- Avoid tautology or repetition.
- Avoid smothered verbs and instead use vigorous verbs.
- Avoid clichés or stereotyped phrases.
- Avoid using slang, i.e. language of a colloquial type.

First Draft :

The points to be kept in mind in writing the first draft are :

1. Keep in mind that you are writing for communicating with the target audience.
2. Keep the purpose of each chapter, section and paragraph in mind.
3. Expand the outline an put all ideas that occur on paper.
4. Do not hesitate to write in any order those sections of your total work that seem to have grown ripe in your mind.
5. Do not struggle of words and phrases.
6. When any idea, word or phrase refuses to come to mind, leave it blank. It will easily on revision.
7. Leave double space between lines to provide for corrections.
8. Do not copy the tables on the text. Just pin them in the appropriate places.
9. Give appropriate headings to chapters sections and paragraphs.
10. Ensure the correctness of facts and citations.

Revisions :

The following points may be kept in mind.

1. Keep in mind the requirements of a research report
2. Fill in the blanks with appropriate ideas, words or phrases.

3. Reorganize ideas wherever necessary.
4. Eliminate gaps in continuity and unclear statements.
5. Simplify sentences and improve their effectiveness.
6. Cut-off repetitions by giving cross references.
7. Improve the readability and clarity of the writing.
8. Correct the spelling and grammatical errors.
9. Make a critical evaluation of the draft.

Methodological Aspects :

Always keep the following points in the mind:

1. Does the report fulfill the aims of the Study?
2. Does the title reflect the core of the problem studied?
3. Are the hypothesis and operational or concepts specific and clear?
4. Is the chapter scheme relevant to the objectives of the study?
5. Are the hypothesis tested appropriately and adequately?
6. Are the findings and inferences clear and substantiated by data evidences?
7. Are the conclusions effective, logical and based on the findings?
8. Are the recommendations specific Practical, weighty and convicting?

Language and Style :

Keep the following in consideration:

1. Is the presentation accurate, clear, crisp, logical and complete?
2. Are the opening sentences of chapters, sections and paragraphs attractive?
3. Is the style smooth and coherent ?
4. Is the expression strong, vigorous and dignified and free from grammatical and spelling errors ?
5. Are the findings and arguments concise ?

15.9 Summary

A research report s a formal statement of the research process and its results. The purpose of a research report is to communicate to interested persons the methodology and the results of the study in such a manner as to enable them to understand the research process and to determine the validity of the conclusions. The aim of the report is not to convince the reader of the value of the but to convey to him what was done, why it was done, and what was its outcome. There are various types of reports like technical reports, interim reports, and summary reports etc. Proper planning is required in writing a report. Every report has well defined contents. Adequate care must be taken in writing various sections like introduction, research methodology, analysis and interpretation, bibliography, appendixes etc. of report. While writing the report, keep in mind that you are writing for communication with the target audience. Keep the purpose of each chapter, section and paragraph in mind. Do no struggle for words and phrases. Give appropriate headings to chapters, sections and paragraphs. Ensure the correctness of facts and citations.

15.10 Key Words

- **Research Report :** It is a formal statement of the research process and its results, It narrated the problem studied, methods used for studying it, and the findings and conclusions of the study.
- **Popular Report :** This type of report is designed for an audience of executives/ administrations and other non- technical users.
- **Interim Report :** It contains a narration of what has been done do far and what are its outcome. It presents a summary of the findings of that part of analysis which has been completed.
- **Summary Report :** A summary report is generally prepared for the lay-audiences, i.e., the general public.
- **Research Abstract :** This is a short summary of the technical report.

15.11 Self Assessment Test

1. What is a research report? Discuss it importance and significances.
2. What are the types of reports? Explain them in brief.
3. What are the items in a research report? Explain them in brief.
4. With the helps of suitable examples discuss the guidelines for preparing bibliography. Also give illustration.

15.12 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.
- William G. Zikmund, "Business Research Methods", Thomson, South-Western Publication, Singapore.
- C.R. Kothari, "Research Methodology – Methods and Techniques", Wilely Eastern Limited, Delhi.