Syllabus

T.Y.B.A. Paper VII

RESEARCH METHODOLOGY

Preamble: Today research is of immense importance in every field of life. Hence students need sound initiation in the world of research. Thus this syllabus is prepared to equip students with basics of research methodology and also provide them acquaintance with the main ingredients of major sources secondary data on Economics, some hands-on experience in conduct so survey including designing questionnaire and interview schedules, collection of data, analysis of data and preparation of report.

SECTION - I

Module 1 : Introduction

- Research meaning and significance
- Characteristics of scientific Research
- Type of research : pure, applied, analytical, exploratory, descriptive, surveys,
- Case-study
- Conceptual or theoretical models
- Research process
- Limitations of Social science research
- Role of computer technology in research

Module 2 : Data : Types, Measurement, Sources and Sampling methods

- Data, information and statistics
- Data types Qualitative and Quantitative; Cross and Time series
- Scales of measurement :nominal, ordinal, interval, ratio
- Sources of data: Primary and secondary
- Census and sample survey-criterion of good sample, choice of sample, probability and non-probability sampling methods, sampling and non-sampling errors

Module 3 : Data collection methods

- Primary data : Methods of collecting primary data : Observation, interview, schedules and questionnaires, case-study
- Framing questionnaire and interview schedule for socio-economic exploratory surveys

- Conducting case study of sick firm or a successful organization, or entrepreneur or an NGO, or a govt. department or a school or a hospital etc. and reporting in format such as objectives, performance, problems and future plans.
- Secondary data
- Sources : Published statistics

  Brief Overview of contents of publications such as

  Economic survey, RBI Bulletin, Budget Documents, Statistical outline of India, Stock Exchanges etc .

  - Newspapers and periodicals providing information on current socio-economic problems
  - Unpublished statistics : records relating to internal activities of institutions such as cost records, profit and loss statement, balance sheet, progress reports, performance records, etc.
- Acquaintance with the Internet websites of important Central government ministries, RBI, W B, IMF, ADB etc,
- Use of search engines, and search methods .

Module 4 : Presentation and preliminary analysis of data

- Classification and tabulation
- Graphical presentation of frequency and cumulative frequency distributions, and of socio-economic data
- Measures of central tendency, Graphical location of locational averages
- Measures of Dispersion : absolute and relative
- Karl Pearson and Bowley's measures of skewness

SECTION - II

Module 5 : Advance analysis of data

- Correlation : Scatter diagram, Pearson's and Spearman's
- Two variable linear regression analysis : Principle of Least Squares, Coefficient of determination
- Relationship between r, byx and bxy
- Time Series Analysis : Components, Estimation of Trend : Moving average, Linear trend

Module 6 : Index Numbers

- Simple Indices Index number : definition, types, uses, problems in its construction, concept of WPI.
- Weighted indices : Laspeyers, paasche, Fisher
- Deflator
- Base shifting and splicing
- Cost of Living Index Number

Module 7 : Hypothesis : Nature and Role in Research

- Definition of a Hypothesis
- Role of Hypothesis
- Types of Hypothesis
- Criteria of Good Hypothesis
- Null and Alternative Hypothesis, parameter and statistic, Type I and type ii errors, Level of significance, Critical region

Module 8 : Report Writing

- Meaning and significance of a Research Report
- Types of Research Report : Technical, Popular, Interim, Summary, Article
- Format of a Research Report : Title to Bibliography
- Principles of writing the Research Report : Organization and Style
- Writing the report on the survey conducted by the student

Reference:

1. Krishnaswamy, O.R. Methodology of Research In Social Sciences, Himalya publishing House, 1993.

2. Wilkinson and Bhandarkar Methodology and Techniques of Social Research, Himalaya Publishing House.

3. Kothari R.C. Research Methodology, Methods and Techniques, New Age International Publishers, IInd revised edition, reprint 2008.

4. Les Oakshott Essential Quantitative Methods for Business Management and Finance, Palgrave .

5. Cooper D. and Schindler P. Business Research Methods, Tata McGraw Hill. Sultan Chand & Sons.

7. Don E. Ehridge research Methodology in Applied Economics : Organizing Planning and Conducting Economics Research, John Wiley and sons, April 2004

8. Gopal M.H. An Introduction to Research Procedure in Social Sciences, Asia

9. Young P.V. scientific Social Survey and Research, Prentice Hall of India Ltd, N.Delhi,1984

1

Module 1
INTRODUCTION

Unit Structure:

1.0 OBJECTIVES

- To know the meaning and objectives of research.
- To understand the limitations of research in social science.
- To understand that the research in social science is differ from the research in physical science.
- To familiarize the students with the significance of economic research.
- To know the various types of research.

1.1. MEANING OF RESEARCH

Research in general refers to a search for knowledge. It means research is an attempt to discover intellectual and practical answers to the various problems through the application of scientific methods to the knowable universe. In fact, research is an art of scientific investigation.

Some Definitions:

(1) The advanced Learner's Dictionary: Research means "a careful investigation or inquiry ".

(2) Webster's Dictionary: research means "a careful critical inquiry or examination in seeking facts or principles, diligent investigation in order to ascertain something."

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

(4) Redman and Mory define research as a "systematized effort to gain new knowledge".

In general, we can say research is an activity or a course of action to go from known to the unknown.

Check your progress:

1) Do you think that research is an art of scientific investigation? Yes / No.
2) Prepare a general definition for research.

_____

_____

_____

_____

1.2. OBJECTIVES OF RESEARCH

The purpose of research is to discover answers to questions through the application of scientific procedures. The main aim of research is to find out the truth which is hidden and which has not been discovered as yet. Thus each research study has its own specific purpose, we may think of research objectives as following:

1) To study purposive, systematic and critical investigation into a phenomena.
2) To aim at describing, interpreting and explaining a phenomenon.
3) To adopt scientific methods of investigation and involve observable or empirical facts.
4) Research directs towards finding answers to relevant questions and solutions to problems.
5) Research emphasizes the development of generalization, principles or theories.

6) To enable researchers to establish generalized laws and to enable them for reliable predictions for future events.

7) To develop new tools, concepts and theories for a better study of unknown phenomena.

It is observed that social sciences are normative to a great extent. Therefore, here we are concerned with research methodology to be used in the science of Economics.

Check your progress:

1) What are the purposes of research?

2) Underline the key words in the above listed objectives of the research.

_____

_____

_____

_____

_____

1.3. LIMITATIONS OF SOCIAL SCIENCE RESEARCH

Limitations of social science research can easily be understood by comparing it with physical science research. For this purpose, we see the difference between social science research and physical science research.

1.3.1. Difference between Social Science Research and Physical Science Research

Social sciences are not exact like Physical sciences i.e. physics, chemistry, etc. They are to some extent normative. The approach in Social Science studies is evaluative. This is because these sciences deal with human beings. For example, economics is defined as a study of mankind in ordinary business of life. However, the environment in which this ordinary business of life is conducted is complex. It is, therefore, more difficult to comprehend and predict human behaviour than the physical phenomena. Not only no two individuals behave in uniform fashion under given circumstances, but the same person can react differently to a given situation at different points of time. It is said that human behaviour is more complex and more mysterious than that of nuclear forces.

Goode and Hatt have noted the following aspects of human behaviour-

1. Human behaviour changes too much from one period to the next to permit scientific exact predictions.
2. Human behaviour is too elusive, subtle and complex to yield to the rigid categorization and artificial instruments of science.
3. Human behaviour can be studied only by other human observers and these always distort, fundamentally, the facts being observed, so that there can be no objective procedures for achieving the truth.
4. Human beings are the subject of predictions and have the ability to upset predictions made by the researchers.

These characteristics constitute a weak foundation to research in social sciences. Hence the study of truth becomes infinitely variable, unique in each case and often non-measurable, whereas in physical sciences, it is repetitive, simplified and observable.

Work in physical sciences on the other hand, lends itself to orderly and close analysis in a laboratory. A physicist can exclude 'outside' disturbances like wars, floods and famines, political instability from the system he is studying, whereas an economist cannot exclude the effect of such disturbances on the subject matter under investigation. In fact, a social scientist cannot exclude anything that affects the society and the economy. His system is 'open' to all the socio-economic-political-technological-psychological changes. Even the change in climate or a timely or untimely arrival of the monsoon affects his predictions.

The work of a physical scientist is confined within the four walls of a laboratory. The laboratory of a social scientist encompasses the entire society. Hence, the scope of work is too vast. It is difficult to obtain precise knowledge of facts to verify the theories. A scientist can conduct controlled experiment repeatedly; a social scientist cannot do it, because society is not amenable directly for experimentation.

The formal logic, which is obvious, is many times not applicable in social sciences. It is like two plus two equal to more than or less than four. For example, a given rise in price may not reduce demand for a commodity to the same extent every time or a rise in the price though reducing demand may increase total revenue in some cases and decrease it in some other cases. Further, it is very difficult to assess the cause and effect relationship between different forces in social sciences. Exact role

of different causes in a relationship changes under different circumstances. For example, in period of rising prices any increase in the price of a commodity, increases its demand instead of reducing it. Kenneth Boulding says that I have been gradually coming under the conviction that there is no such thing as economics- there is only social science applied to economic problems.

Amongst the social scientists, economists frequently disagree among themselves. There are endless among them on both analytical and policy issues. Such things do not happen in the case of physical sciences. However, we can compare these to researches on some basic points as under:-

Point	**Research in**	
	Social Sciences	**Physical Sciences**
Nature	Based on human behaviour.	Based on the Laws of the Nature.
Method of study	Deductive method is used in social science research.	Inductive method is used in physical science research.
Nature of Laws	Social sciences studies laws related human social life and human behaviour.	The physical sciences studies physical laws in natural phenomena.
Fundamentals	The fundamental elements of social sciences are psychologically related.	The basic elements of physical sciences have a physical relation.
Basic Elements	The basic elements of social sciences are man, his mental state and behaviour.	The basic elements of the physical sciences are the physical elements ruled by natural laws.
Measurability	Social sciences provide comparatively lesser scope for measurement of subject matter. There are no standard units of measurement.	There is greater possibility of measurement in study of physical phenomena. There are certain and standard micro units of measurement.

Accuracy	Being related to the study of society the social sciences have comparatively less accuracy.	Because they study physical elements the physical sciences possess greater accuracy.
Objectivity	Objectivity is achieved with difficulty in social sciences. There is more subjectivity.	Objectivity is attained easily in physical sciences whereas subjectivity is absent.
Scope of Laboratory	It is difficult to construct laboratories for social sciences. Society is their laboratory.	Physical sciences have their laboratories because they can easily be made for studying physical objects.
Limitations	The basic elements of the social sciences cannot be separated analytically.	The basic elements of the physical sciences can be separated by analysis.
Predictability	Because of their lesser accuracy, social sciences can make comparatively fewer predictions.	The physical sciences can make more predictions due to a higher degree of accuracy.

Check your progress:

1) Do you thin that research in social science is differ from research in physical science? Yes / No.

2) List down the points on which basis you say that research in social science is differ from research in physical science.

_____

_____

_____

_____

_____

1.4. SIGNIFICANCE OF ECONOMIC RESEARCH

Research leads to discovery of facts and the interpretation of the facts and helps in understanding economic reality.

The developing countries have urgent problems such as poverty, unemployment, economic imbalance, economic inequality, low productivity, etc. The nature and dimensions of such problems have to be diagnosed and analyzed. Economic research plays a significant role in this respect. An analysis of problems leads to an identification of appropriate remedial actions.

The facts discovered through research are systematized and lead to development of knowledge. For example, studies on poverty have lead to development of theoretical concepts such as poverty line, income inequality, measures of income inequality, theoretical discussions on problems of more equitable distribution of income, difference between economic growth and development and so on.

Economic research equips the Society with the first hand knowledge about the organization and working of the economy and economic institutions. This gives economists a greater power of control and insight over economic phenomena.

Research also aims at finding an order among facts. This affords a sound basis for prediction. Such predictions are useful in economic planning e.g. predictions about population growth and consequent requirements of this population growth in the number of educated unemployed, requirements of power or energy in different productive sectors of the economy.

Planning for socio-economic development calls for baseline data on the various aspects of our economy like resource requirement, resource availability, future needs, etc. Systematic research can give the required data base for planning schemes for development and for taking policy decisions in vital areas of development and for taking policy decisions in vital areas of development. Evaluation studies of various projects, feasibility studies, etc. help in gauging the impact of plans on the economic development and also suggest ideas for change or reformulation.

Economic research can suggest various areas where remedial measures are required for improving socio-economic welfare.

In short, the area of economic research is almost unlimited. There is a vast scope for useful economic research.

Check your progress

1) List down the points showing the significance of economic research.

_____

_____

_____

_____

_____

1.5. TYPES OF RESEARCH

Any classification of research into different types is inevitably arbitrary. However, researches have been classified differently depending on the approach, the purpose and the nature of research activity. Broadly speaking, research can be classified according to its aim, purpose and method as follows:

1) Pure research,
2) Applied research,
3) Exploratory research,
4) Descriptive research,
5) Diagnostic research,
6) Evaluation studies,
7) Action research,
8) Experimental research,
9) Analytical study,
10) Survey, and
11) Historical study

The above classification is just an approach to differentiate the various aspects of research. These different types do not indicate water-tight demarcations and are not clearly distinguishable from each other. However, the outcome and quality of a research project depends on the suitable choice of the type. However, each type needs a different research design. We now, discuss the salient features of these different types.

(1) Pure Research: It is also called fundamental or theoretical research. It is original or basic in character. Pure research is undertaken for the sake of knowledge without any intention to apply it in practice. It is undertaken out of intellectual curiosity. The findings of pure research form the basis of applied research.

Pure research may also involve improvements in the existing theory by relaxing some of its assumptions or re-interpretation or development of new theory on the basis of the existing one. For example, the Malthusian Theory of population gave rise to the Optimum theory of population. By questioning some of the assumptions of the Keynesian theory, Friedman came out with new interpretation of the monetary phenomenon. Theories of distribution have also been altered or re-interpreted to suit the changing economic conditions.

(2) Applied Research: It is carried out to find out solutions to real life problems which require an action or policy decision. It is thus policy-oriented or action-directed. This type of research is based on the application of known theories and models to the actual operational fields. The applied research is conducted to test the empirical content or even the validity of a theory under given condition. For example, a researcher may use Lewis model of growth in labour-surplus economies and examine whether the real wage rate remains constant till all the surplus labour is completely used. The results of his research may stimulate further studies. Alterative strategies are developed, if a model does not work under specific conditions. In short, applied research contributes to social science by providing the required convincing evidence about usefulness of a theory to society. It also develops and utilizes techniques useful for further basic research. Applied research also provides data and ideas which may help in speeding up the process of generalization.

Applied research seeks immediate and practical results. For example, marketing research for developing new markets for a product indicates the marketability and hence the viability of a new product venture. There is a vast scope for applied research in the fields of technology, management, commerce, economics, etc. Incidentally, it may contribute to the development of theoretical knowledge and may lead to the discovery of new facts.

Applied research has practical utility in the developing countries. These countries can apply the theories developed by the developed counties since pure research is very expensive and the developing countries cannot afford to spend on pure research. Applied research often takes the form of field investigation and aims at collecting data. The accuracy and adequacy of data has considerable effect on the way in which a model is tested.

We, now briefly describe the purpose of a few other types of research studies.

(3) Exploratory research: It is also called formulative study. Such a study is conducted to gain familiarity with a phenomenon as it is a

preliminary study of an unfamiliar problem. It is similar to a doctor's initial investigation of a patient, for getting some clues for identifying his ailment. Exploratory research is ill structured and is usually in the form of a pilot study. It helps the researcher to formulate a more precise research problem or to develop a hypothesis, or concentrate on discovery of new ideals and insights.

In short, the purpose of an exploratory study may be:

i. To gather information for clarifying concepts,

ii. To develop familiarity with the problem,

iii. To generate new ideas, and

iv. To determine the feasibility of the study.

Exploratory research is necessary in social sciences as these sciences are relatively of recent origin and there are only a few researches in them.

(4) Descriptive Research: It is a fact finding mission. The aim of these studies is to accurately lay down the characteristics of any group, situation or individual. It is the simplest type of research, but is more specific than exploratory study.

Descriptive study is possible in the case of:

i. Problems which are describable and not arguable. For example, philosophical or other controversial issues are not suitable for descriptive research.

ii. The date should be amenable to an accurate, objective and if possible quantitative analysis for reliability and significance.

iii. It should be possible to develop valid standards of comparison.

iv. It should lend itself to verifiable procedure of collection and analysis of data.

However, the study by itself does not deal with the testing of propositions or hypothesis.

(5) Diagnostic Research: It is similar to the descriptive study. But it is focused towards discovering the frequency with which something occurs and to find out why does it occur. It aims at identifying the causes of a problem or association of occurrence of a problem with something else and tries to obtain a possible solution. Thus, it requires:

i. Prior knowledge of the problem.

ii. Its thorough formulation.

iii. Clear cut definition of the population.

iv. Adequate method for collecting correct information.

v. Precise measurement of variables.

vi. Statistical analysis.

A diagnostic research involves large amount of work. It is guided by hypothesis and it is not possible in areas where knowledge is not advanced enough to make a diagnosis of the given situation.

(6) Evaluation Studies: It is a type of applied research. It is made for assessing the effectiveness of various planned or implemented social or economic programmes (like family planning, adult literacy, mid-day meal for students) or to assess the impact of developmental projects on the development of the area (like irrigation projects, rural development projects, slum clearance).

Thus, its aim is:

i. To appraise the effect of any activity for its qualitative and quantitative performance,

ii. To determine the conditions required for success of a programme or activity,

iii. To assess the changes required over time, and

iv. To find the means to bring about these changes.

(7) Experimental Research: It tries to assess the effect of one particular variable or a set of variables on a phenomenon. It aims at determining whether and in what manner variables are related to each other, i.e. it tests a hypothesis of causal relationship between variables. There are various types of designs available for an experimental study. These designs aim at using procedures which can reduce bias and increase the reliability of results.

Discussion about type of research, based on method of study mainly concentrates on analytical, historical and survey research. Analytical study is called a study involving use of statistical methods and (***we shall discuss these methods in the later part of this Book.)***

(8) Survey: Survey is a fact finding study. It is a method of research which involves collection of data directly from the individuals (either from the entire group i.e. a sample). It requires planning by experts in the field of surveys. It also needs careful analysis and then rational interpretation of the findings.

The surveys may be conducted for various purposes:

i. Fact-finding surveys may have a descriptive purpose. They provide required information to the Government, business houses, private bodies etc. on the various subjects like expenditure patterns, market demand, etc.

ii. Many inquires are conducted to explain certain phenomenon. Such surveys test hypothesis, to explain relationships between different variables, e.g. consumer behaviour or responses to certain stimuli, effect of changes in tax structure on expenditure, income earning capacity, labour productivity etc. Such surveys are useful for making predictions and for taking policy decisions.

iii. Surveys are also concluded to make spatial and/or temporal comparisons of behavioural groups, economic status, demographic patterns etc.

The above mentioned purposes of the survey clearly indicate that a survey is a field study and it is conducted in a natural setting. It seeks direct responses from individuals. It can cover a large population and tackle various problems. The study may be intensive or extensive. It covers a definite geographical area.

Surveys can be conducted on all aspects of human behaviour. However, we can broadly divide them into two categories namely, (1) Social, (2) Economic.

Survey method of conducting research is the most versatile of all methods of research. It is also the only practical way to obtain detailed first hand information on different aspects of a given topic. It enables a researcher to draw generalizations about large population on the basis of a sample study. Survey is also a useful instrument for verifying theories.

However, success of a survey depends on the willingness and co-operation of the respondents. Results obtained from a survey are, therefore, subject to these response errors to consider while judging the reliability or accuracy of the results obtained from a survey.

(9) Historical Research: It is a study based on past records. It is based on the belief that the past contains the key to the present and then present influences the future. This method is more often used in sociological research. Historical study helps in tracing the evolution, growth and transformation of society.

The major limitation of this method is that it is difficult to judge the accuracy and reliability of past data. If the data relates to very distant past period, it is difficult to perceive the significance of such data. Apart from these limitations, verification of results is not possible in historical studies.

Check your progress:

(1) What are the types of research?

(2) What types of researches are generally found in social sciences?

1.6 SUMMARY

1. Research is an attempt to discover intellectual and practical answers to the various problems through the application of scientific methods to the knowable universe.
2. The main aim of research is to find out the truth which is hidden and which has not been discovered yet.
3. There is a weak foundation to research in social sciences. Hence study of truth becomes infinitely variable, unique in each case and often non measurable, whereas in physical sciences, it is repetitive, simplified and observable .
4. Research leads to discovery of facts and the interpretation of the facts and helps in understanding economic reality. Economic research can suggest various areas where remedial measures are required for improving socio-economic welfare.
5. Researchers have been classified differently depending on the approach, the purpose and the nature of research activity. However, the outcome and quality of a research project depends on the suitable choice of the type.

1.7 QUESTIONS

1) What are the objectives of research?
2) Distinguish between research in social sciences and research in physical sciences.
3) Explain the term research and outline the significance of economic research.
4) Write a note on significance of Economic Research.
5) Explain various types of research.

2

RESEARCH PROCESS

Unit Structure:

2.0 OBJECTIVES

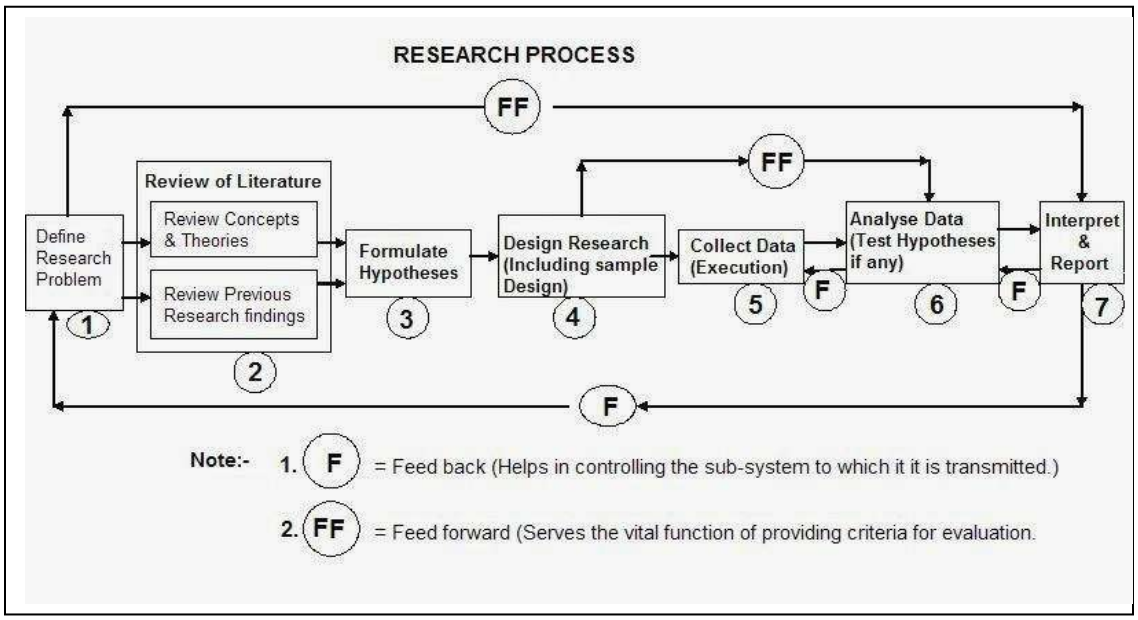
- To know the research process.
- To know the characteristics and qualities of a good research.
- To provide the essential of a good research to the students.
- To make aware of the general problems of the researchers in India.
- To acquaint the students with the use of computer technology.

2.1. RESEARCH PROCESS

Before going to the details of research methodology and techniques, it is appropriate to present a brief overview of the research process. A cursory glance at different published researches suggests that research comprises of a definite sequence of processes. The process involves a number of clearly distinct steps. The number of steps, their names and their sequences may vary in different cases. However, it is necessary to note that the research process involves a number of interrelated activities. These activities overlap continuously and do not always follow a prescribed sequence. The operations involved in the research process are so interdependent that the earlier steps may determine the later steps. Thus, the researcher has to be

constantly anticipating at each step, the requirements of the subsequent steps. Even then, it is worth while to determine the steps involved in planning a research project.

Thus, research process consists of series of actions or steps necessary to effectively carry out research and the desired sequencing of these steps. The chart given below illustrates a research process:



The chart indicates that the research process consists of a number of closely related activities, as shown through 1 to 7. But such activities overlap continuously rather than following a strictly prescribed sequence. At times, the first step determines the nature of the last step to be undertaken. If subsequent procedures have not been taken into account in the early stages, serious difficulties may arise which may even prevent the completion of the study. One should remember that the various steps involved in a research process are either not mutually exclusive nor they are separate and distinct. They do not necessarily follow each other in any specific order and the researcher has to be constantly anticipating at each step in the research process the requirements of the subsequent steps. However, the following order concerning various steps provides a useful procedural guideline regarding the research process: (1) formulating the research problem, (2) extensive literature survey, (3) developing the hypotheses, (4) preparing the research design, (5) determining sample design, (6) collecting the data, (7) execution of the project, (8) analysis of data, (9) hypothesis testing, (10) generalizations and interpretation and (11) preparation of the report or presentation of the results i.e. formal write-up of conclusions reached.

A brief description of the above steps will be helpful:

1. **Formulating the Research Problem:** The formulation of a general topic into a specific research problem 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.

2. **Extensive Literature Survey:** Once the problem is formulated, a brief summary of it should be written down. At step the researcher should undertake extensive literature survey connected with the problem. For this purpose, the abstracting and indexing journals and published or unpublished bibliographies are the first place to go to. Academic journals, conference proceedings, government reports, books, etc., must be tapped depending on the nature of the problem. In this process, it is to be remembered that one source will lead to another. The earlier studies, if any, which are similar to the study in hand, should be carefully studied. A good library will be a great help to the researcher at this stage.

3. **Development of Working Hypotheses:** After extensive literature survey, researcher should state in clear terms the working hypothesis or hypotheses. Working hypothesis is a tentative assumption made in order to draw out and test its logical or empirical consequences. As such the manner in which research hypotheses are developed is particularly important since they provide the focal point for research.

4. **Preparing the Research Design:** The research problem having been formulated in clear cut terms, the researcher will be required to prepare a research design, i.e. he will have to state the conceptual structure within which research would be conducted. The preparation of such a design facilitates research to be as efficient as possible yielding maximal information. In other words, the function of research design is to provide for the collection of relevant evidence with minimal expenditure of effort, time and money. But it depends mainly on the research purpose.

5. **Determining Sample Design:** All the items under consideration in any field of inquiry constitute a 'universe' or 'population'. A complete enumeration of all the items in the 'population' is known as a census inquiry. It can be presumed that in such an inquiry when all the items are covered no element of chance is left and highest accuracy is obtained. But in practice this may not be true. But in practice this may not be true because of

bias. Besides, this type of inquiry involves a great deal of time, money and energy. Also, census inquiry is not possible in practice under many circumstances. For instance, blood testing is done only on sample basis. Hence, quite often we select only a few items from the universe for our study purposes. The items so selected constitute what is technically called a sample. A sample design is a definite plan determined before any data are actually collected for obtaining a sample from a given population. Samples can be either probability samples or non-probability samples. With probability samples each element has a known probability of being included in the sample but the non-probability samples do not allow the researcher to determine this probability. Probability samples are those based on simple random sampling, systematic sampling, stratified sampling, cluster/ area sampling whereas non-probability samples are those based on convenience sampling, judgment sampling and quota sampling techniques.

6. **Collecting the Data:** In dealing with any real life problem it is often found that data at hand are inadequate and hence, it becomes necessary to collect data that are appropriate. 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. Primary data can be collected either through experiment or through survey. If the researcher conducts an experiment, he observes some quantitative measurements, or the data, with the help of which he examines the truth contained in his hypothesis. But in the case of a survey, data can be collected by any one or more of the ways like, by observation, through personal interviews, through telephone interviews, by mailing of questionnaires, through schedules. The researcher can select one or more of these methods taking into consideration the nature of investigation, objective and scope of the inquiry, financial resources, available time and the desired degree of accuracy.

7. **Execution of the project:** It is a very important step in the research process. If the execution of the project proceeds on correct lines, the data to be collected would be adequate and dependable. The researcher should see that the project is executed in a systematic manner and in time. If the survey is to be conducted by means of structured questionnaires, data can be readily machine-processed. In such a situation, questions as well as the possible answers may be coded. If the data are to be collected through interviewers, arrangements should be made for proper selection and training of the interviewers. The training may be given with the help of instruction manuals which explain clearly the job of the interviewers at each step.

8. **Analysis of Data:** After the data have been collected, the researcher turns to the task of analyzing them. The analysis of data requires a number of closely related operations such as establishment of categories, the application of these categories to raw data through coding, tabulation and then drawing statistical inferences. The data difficult to handle should necessarily be made easy by a few manageable groups and tables for further analysis. Thus, researcher should classify the raw data into some purposeful and usable categories. Coding operation is usually done at this stage through which the categories of data are transformed into symbols that may be tabulated and counted. Editing is the procedure that improves the quality of the data for coding. With coding the stage is ready for tabulation. Tabulation is a part of the technical procedure wherein the classified data are put in the form of tables. The mechanical devices can be made use of at this juncture. A great deal of data, especially in large inquiries, is tabulated by computers. Computers not only save time but also make it possible to study large number or variables affecting a problem simultaneously.

   Analysis work after tabulation is generally based on the computation of various percentages, coefficients, etc., by applying various well defined statistical formulae. In the process of analysis, relationships or differences supporting or conflicting with original or new hypotheses should be subjected to rests of significance to determine with what validity data can be said to indicate any conclusion(s). For instance, if there are two samples of weekly wages, each sample being drawn from factories in different parts of the same city, giving two different mean values, then our problem may be whether the two mean values are significantly different or the difference is just a matter of chance. Through the use of statistical tests we can establish whether such a difference is a real one or is the result of random fluctuations. If the difference happens to be real, the inference will be that the two samples come from different universes and if the difference is due to chance, the conclusion would be that the two samples belong to the same universe. Similarly, the technique of analysis of variance can help us in analyzing whether three of more varieties of seeds grown on certain fields yield significantly different results or not. In brief, the researcher can analyze the collected data with the help of various statistical measures.

9. **Hypothesis Testing:** After analyzing the data as stated above, the researcher is in a position to test the hypotheses, if any, he had formulated earlier. Do the facts support the hypotheses or they happen to be contrary? This is the usual question which should be answered while testing hypotheses. Various test,

such as *Chi*-square test, *t*-test, *F*-test, have been developed by statisticians for the purpose. The hypotheses may be tested through the use of one or more of such test, depending upon the nature and object of research inquiry. Hypothesis-testing will result in either accepting the hypothesis or in rejecting it. If the researcher had no hypotheses to start with, generalizations established on the basis of data may be stated as hypotheses to be tested by subsequent researches in times to come.

10. **Generalizations and Interpretation:** If a hypothesis is tested and upheld several times, it may be possible for the researcher to arrive at generalization, i.e., to build a theory. As a matter of fact, the real value of research lies in its ability to arrive at certain generalizations. If the researcher had no hypothesis to start with, he might seek to explain his findings on the basis of some theory. It is known as interpretation. The process of interpretation may quite often causes to new questions which in turn may lead to further researches.

11. **Preparation of the Report or the Thesis:** Finally, the researcher has to prepare the report of what has been done by him. Writing of report must be done with great care keeping in view the structure of a research. **(*Structure of a Research Report is explained in next topic separately*).**

Check your progress:

1. Prepare flow chart of Research Process.
2. List down sequentially various steps involved in Research Process.

_____

_____

_____

_____

_____

2.2. CHARACTERISTICS OF SCIENTIFIC RESEARCH

It is expected on general or common ground that scientific research has to satisfy the following criteria**:**

1. The purpose of the research should be clearly defined and common concepts be used.

2. The research procedure used should be described in sufficient detail to permit another researcher to repeat the research for further advancement, keeping the continuity of what has already been attained.

3. The procedural design of the research should be carefully planned to yield results that are as objective as possible.

4. The researcher should report with complete frankness, flaws in procedural design and estimate their effects upon the findings.

5. The analysis of data should be sufficiently adequate to reveal its significance and the methods of analysis used should be appropriate. The validity and reliability of the data should be checked carefully.

6. Conclusions should be confined to those justified by the data of the research and limited to those for which the data provide an adequate basis.

7. Greater confidence in research is warranted if the researcher is experienced, has a good reputation in research and is a person of integrity.

2.3. QUALITIES OF A GOOD RESEARCH

On the basis of the above criteria we can easily state some qualities of a good research as under:

1. **Good research is Systematic:** It means that research is structured with specified steps to be taken in a specified sequence in accordance with the well defined set of rules. Systematic characteristic of the research does not rule out creative thinking but it certainly does reject the use of guessing in arriving at conclusions.

2. **Good Research is Logical:** This implies that research is guided by the rules of logical reasoning and the logical process of induction and deduction are of great value in carrying out research. Induction is the process of reasoning from a part to the whole whereas deduction is the process of reasoning from the whole to a part. In fact, logical reasoning makes research more meaningful in the context of decision making.

3. **Good Research is Empirical:** It implies that research is related basically to one or more aspects of a real situation and deals with concrete data that provides a basis for external validity to research results.

4. **Good Research is Replicable:** This characteristic allows research results to be verified by replicating the study and thereby building a sound basis for decisions.

2.4. EVALUATION OF GOOD RESEARCH

A good scientific research has to satisfy several requisites. For example the Advisory Committee on Economics and Social Research in Agriculture of the Social Science Research Council summarized the essentials of good scientific research as follows:

i. Careful logical analysis of the problem, isolating it from other problems and separating its elements. This means, in some cases the formulation of a hypothesis, in the proper meaning of this expression- a trial hypothesis that will point the investigation.

ii. Unequivocal definition of terms and concepts and statistical units and measures, so that others will understand exactly, and be able to repeat the analysis and test the generalizations.

iii. Collection of cases and data pertinent to the subject on hand.

iv. Classification of cases and phenomena and data.

v. Expression of factors in quantitative terms whenever possible.

vi. Rigorous and exacting experimental or statistical procedure in summarizing the data and in isolating the attributes or variables and measuring their relationships and inter-effects.

vii. Statement in confusing terms of the exact conclusion that warranted for the cases examined.

viii. Sound logical reasoning as to the bearing of these conclusions on the trial hypothesis and in the formulation of generalizations.

ix. Statement of conclusions or generalizations definitely and clearly so that others will be able to check them.

x. Complete elimination of the personal equation.

xi. Complete and careful reporting of the data and the methods of analysis so that others can check the analysis, or test the generalizations with new sets of data.

Check your progress:

1) Which common criteria are to be satisfied by a scientific research?
2) What are the qualities of a good research?
3) How do you evaluate a research?

_____

_____

_____

_____

_____

2.5. PROBLEMS OF RESEARCHERS IN INDIA

Researchers in India, particularly those engaged in empirical research, are facing several problems regarding following areas:

1. The lack of a scientific training in the methodology of research is a great impediment for researchers in out country. For this, efforts should be made to provide short-duration intensive courses for meeting this requirement.

2. There is insufficient interaction between the university research departments on one side and business establishments, government departments and research institutions on the other side. Efforts should be made to develop satisfactory communication and co-operation among all concerned for better and realistic researches.

3. Most of the business units in our country do not have the confidence that the material supplied by them to researchers will not be misused and as such they are often unwilling to supply the needed information to researchers. There is need for generating the confidence that the information or data obtained from a business unit will not be misused.

4. Research studies overlapping one another are undertaken quite often for want of adequate information. This results in duplication and waste of money, time and resources. This problem can be solved by proper revision of a list of subjects on which and the places where the research is going on at regular intervals.

5. There does not a code of conduct for researchers and inter-university and inter-departmental rivalries are also quite

common. Hence, there is need for developing a code of conduct for researchers which, if adhered sincerely, can win over this problem.

6. Many researchers in our country also face the difficulty of adequate and timely secretarial assistance, including computation assistance. This causes unnecessary delays in the completion of research studies. All the possible efforts are made in this direction so that efficient secretarial assistance is made available to researchers and that too well in time.

7. Library management and functioning is not satisfactory at many places and much of the time and energy of researchers are spent in tracing out the books, journals, reports, etc., rather than in tracing out relevant material from them.

8. There is also the problem that many of our libraries are not able to get copies of old and new Acts or Rules, reports and other government publications in time. Efforts should be made for the regular and speedy supply of all these publications to reach our libraries.

9. There is also the difficulty of timely availability of published data from various government and other agencies doing this job in our country.

10. There may, at times, take place the problem of conceptualization and also problems relating to the process of data collection and related things.

Check your progress:

1) What are problems that have to face the Indian researchers?
2) How would they be solved?

_____

_____

_____

_____

_____

2.6. ROLE OF COMPUTER TECHNOLOGY IN RESEARCH

Computer technology in research, no doubt, plays vital role. This technology is very user-friendly and well familiar among the

various study groups and researchers. The computer is changing the technology of economic and social research. The ability of the computer to process large data makes possible.

We find at every step in the research process computer technology is an essential. We can observe its importance through research process. Following points, elaborates the role of computer technology in research:

1) Because of computer technology, various study groups and researchers are more curious and widen their knowledge which arises various research problems in their minds.

2) Once researchers dealing with the problems, computer technology helps to provide easy axis in thorough wide spread literature review.

3) With the help of computer technology, researcher is able to design various types of documents and data collection tools more neatly.

4) After data collection, now a day, it is very easy job to present the data more effectively with various pictures, diagrams and graphs.

5) Now with the help of the computer one can easily analyze and compute the data by using statistical tools and various soft wares, which require short training.

6) Because of all these, computer technology saves time, money and resources.

7) Researchers are now more confident and work out the jobs more accurately.

8) These technologies double-up the research work so that researchers can easily avoid unnecessary delays in the process.

9) Computer technology provides variety of effectiveness in printing jobs.

10) Computer technology makes researchers independent and isolate to study and getting well work done.

Check your progress:

1) Do you think that computer is changing the technology of economic and social research? How?

2) Explain the role of computer technology in research.

_____

_____

_____

_____

_____

2.7 SUMMARY

1. Research process consist of series of actions or steps necessary to effectively carry out research and the desired sequencing of these steps.

2. The scientific research should satisfy the criteria like the purpose should be clearly defined, the research procedure used should be described in detail, the procedural design should be carefully planned, the researcher should report with complete frankness, the analysis of data should be sufficiently adequate, greater confidence in research is warranted.

3. Some qualities of a good research are : good research is systematic, logical, empirical, and replicable.

4. A good scientific research has to satisfy several requisitions.

5. Researchers in India are facing several problems such as lack of a scientific training, insufficient interaction, lack of confidence, overlapping of research studies, lack of code of conduct for researchers, difficulty of adequate and timely secretarial assistance, difficulty of timely availability of published data, problem of conceptualization etc.

6. The computer is changing the technology of economic and social research. This technology is very user friendly and well familiar among the various study groups and researchers.

2.8 QUESTIONS

1) Explain / Describe the process of research work.
2) Explain various steps involved in research work.
3) What are the characteristics of a scientific research?
4) What are the qualities of a good research?
5) How do you evaluate a research work?
6) What are the difficulties faced by a researcher in India?
7) Explain the role of computer technology in social science research.

3

Module 2
DATA SOURCE AND MEASUREMENT

Unit Structure:

3.0 OBJECTIVES

- To understand the meaning of Data.
- To study the processing of data and approaches to data management.
- To study the meaning, types, importance, limitations of Statistical data.
- To understand the meaning of Measurement and levels of measurement.
- To study the various components of measurement.
- To study the different types and classification of data.

3.1 INTRODUCTION

Data is the raw material from which useful information is derived. The word Data is the plural of Datum. Data is commonly used in both singular and plural forms. It is defined as raw facts or observations, typically about physical phenomenon or business transactions. Example: A sale of a machine tool or an automobile

would generate a lot of data describing those events. Data are objective measurements of attributes (the characteristics) of entities (such as people, place, things and events). These measurements are usually represented by symbols such as numbers, words, and codes, composed of a mixture of numerical, alphabetical and other characters. It takes a variety of forms, including numeric data, text, voice and images.

Data is the collection of facts, which is unorganized but can be organized into useful information. The term data and information come across in our daily life and are often interchanged. Example: Dates, weights, prices, costs, number of items sold, employees' names, product names etc.

3.2 DATA PROCESSING

The conversion of facts into meaningful information is known as data processing. It is also called in general as information processing. It is the processing of data to make it more usable meaningful, thus transforming it into information.

Database

- Shared collection – can be used simultaneously by many departments and users.
- Logically related - comprise the important objects and the relationships between these objects.
- Description of the data – the system catalog (meta-data) provides description of data to enable data independence.

3.2.1 Approaches to Data Management

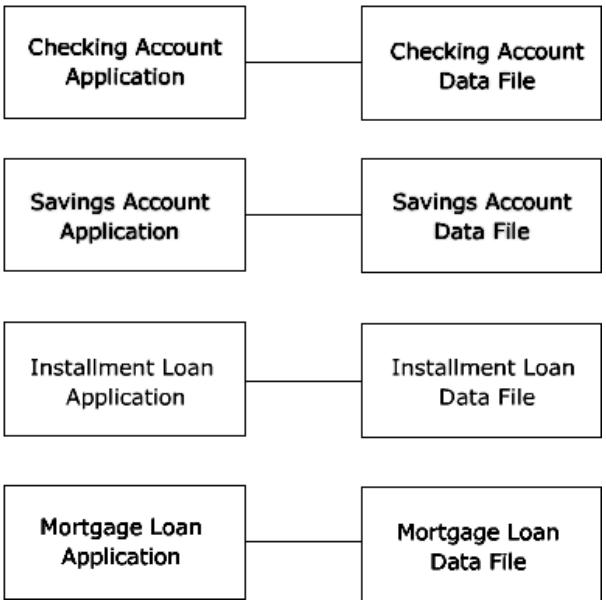
A) File-Based Systems

Conventionally, before the Database systems evolved, data in software systems was stored in and represented using flat files.

B) Database Systems

Database Systems evolved in the late 1960s to address common issues in applications handling large volumes of data which are also data intensive. Some of these issues could be traced back to the following disadvantages of File-based systems.

Drawbacks of File-Based Systems



File Based System

As shown in the figure, in a file-based system, different programs in the same application may be interacting with different private data files. There is no system enforcing any standardized control on the organization and structure of these data files.

- **Data Redundancy and Inconsistency**

Since data resides in different private data files, there are chances of redundancy and resulting inconsistency. For example, in the above example shown, the same customer can have a savings account as well as a mortgage loan. Here the customer details may be duplicated since the programs for the two functions store their corresponding data in two different data files. This gives rise to redundancy in the customer's data. Since the same data is stored in two files, inconsistency arises if a change made in the data in one file is not reflected in the other.

- **Unanticipated Queries**

In a file-based system, handling sudden/ad-hoc queries can be difficult, since it requires changes in the existing programs.

- **Data Isolation**

Though data used by different programs in the application may be related, they reside in isolated data files.

- **Concurrent Access Anomali**es

In large multi-user systems the same file or record may need to be accessed by multiple users simultaneously. Handling this in a file-based systems is difficult.

- **Security Problems**

In data-intensive applications, security of data is a major concern. Users should be given access only to required data and not the whole database. In a file-based system, this can be handled only by additional programming in each application.

- **Integrity Problems**

In any application, there will be certain data integrity rules which need to be maintained. These could be in the form of certain conditions/constraints on the elements of the data records. In the savings bank application, one such integrity rule could be “Customer ID, which is the unique identifier for a customer record, should be non-empty”. There can be several such integrity rules. In a file-based system, all these rules need to be explicitly programmed in the application program.

It may be noted that, we are not trying to say that handling the above issues like concurrent access, security, integrity problems, etc., is not possible in a file-based system. The real issue was that, though all these are common issues of concern to any data-intensive application, each application had to handle all these problems on its own. The application programmer needs to bother not only about implementing the application business rules but also about handling these common issues.

3.3 INFORMATION

The English word was apparently derived from the Latin accusative form (informationem) of the nominative (informatio): this noun is in its turn derived from the verb "informare" (to inform) in the sense of "to give form to the mind", "to discipline", "instruct", "teach": "Men so wise should go and inform their kings." (1330) *Inform* itself comes (via French) from the Latin verb *informare*, to give form to, to form an idea of. Furthermore, Latin itself already contained the word *informatio* meaning concept or idea, but the extent to which this may have influenced the development of the word *information* in English is not clear.

Check Your Progress:

1. What do you understand by data?
2. What is data processing?
3. State the two approaches to data management.

3.4 STATISTICAL DATA

Data literary means facts. The word data is used to denote information. Raw data means the information collected through censuses and surveys or in a routine manner from other sources. The first step involved in statistics is the collection of statistics data. It is a basis or foundation of statistical investigations.

Statistics data is of two types –

1) Primary data and
2) Secondary data

Primary data is collected by a particular person or organization for its own use from the primary source.

Statistic:

Statistics and its methodology have gained recognition as an important tool for analysis of interpretation of data in natural biological, agricultural methodology taking business decision and engineering science. This subject found relevance in various fields. The economics commerce, psychology, geography, geology, Engineering, Mathematics, Business management. Since it is applicable to various fields as explained above, is necessary to study this subject. This may be found before in banking to take investment decision and intervene for collective and predicting customer's behaviors etc.

Statistic is a word derived from a Latin word "STATUS" which means a Political state. Statistics is a tool for human being to translate complex facts into simple and understandable statement of facts.

3.4.1 Definition of Statistics

1. "Statistics is commonly understood nowadays as a mathematical disciple concerned with the study of masses of quantities data of any kind. " (Encyclopedic of Britannia (1996))

2. "As a name of a field of study, Statistics refers to the science and art of obtaining and analyzing quantitative data with a

view to make a sound interference in the face of certainty (encyclopedia American (1998)).

3. "Statistics deals with the inferential process in particular, with the planning and analysis of experiments and surveys, with the nature of observational error and sources of variability that observe underlying patterns with efficient summarizing of sets of data." (International encyclopedia of social science (1968))

Hence statistics deals with phenomena in which occurrence of events under study cannot be predicted with certainty.

3.4.2 Importance of statistics

Statistics involves

1. Collection of data.
2. Analysis of data.
3. Interpretation of facts therefrom. To do theses statistical methodology is to be used.

Collection of data is done by sampling method or experimental design. Here statistics is concerned with methodology used for summarizing data and to obtain salient information from the data. This way of doing this is called descriptive statistics. Inference that is drawn and conclusions arrived at in final stage and this is termed statistical inference.

Some of the areas where statistics forward its importance they are

1. **Consumption:** Statistical data of consumption enable us to find out various ways in which people in different strata of society spend their income.

2. **Production:** Statistical data of production gives total productivity in the country, which enables us to compare ourselves with other countries of the world.

3. **Exchange:** In this field, an economist study market, law of prices that are determined by the forces of demand and supply, production cost, competition banking etc. A detailed and systematic study of all these can be made only with the help of statistics

4. **Econometrics:** This is a combination of economics, statistics and mathematics. With the help of econometrics economics has become exact science.

5. **Public finance:** Study of revenue and expenditure of country is called public finance. Budget (A statistical data document) fiscal policy of the government, deficit financing etc. which based on statistics, is the concept of economics.

6. **Input-Output analysis:** Relationship of input-output is based on statistical data explains input-output analysis. Sampling Time series, Index numbers, probability correlation and regression are some of the important applications where statistics finds its place.

As per Prof. Alfred Marshall “statistics are the straws out of which I like every other economies, have to make bricks”

Further statistics found important application in the following area

a) Economics
b) Commerce
c) Auditing and accounting
d) Economics planning
e) Astronomy and meteorology
f) Biology
g) Mathematics
h) Natural science
i) Education
j) Business
k) Government
l) Common man
m) In army
n) Banking
o) Insurance companies
p) Politicians
q) Inversions
r) Public utility services etc.

Limitations of statistic:

i) Statistics deals with aggregate or group of items and not individual item.

ii) Statistics deals with the quantitative data. If the phenomena understudy yields only qualitative data. For ex. a situation like poverty is qualitative and as such not available to statistical analysis. This is possible only by assigning suitable quantitative measures to such a situation.

iii) Statistical data holds good for averages or average individual. It may not be true for particular individual. As per W. J. King, statistics deals with averages and these may be made up of individual item radically different from each other.

iv) Statistics does not reveal entire story of a problem. Since most of the problems are affected by such a factor, which are incapable of statistical analysis, it is not possible to examine problem in all its manifestation only by statistical approach.

v) Statistics is liable to be misused and misrepresented. The most significant limitation of statistics is that it must be used by experts. According to Bowley “Statistics only furnishes tool though imperfect, which is dangerous in the hands of those who do not know its use and deficiencies. According to W. J. King “Statistics is clay of which you can make a God or Devil as you please”. He remarks “Science of statistics is the useful servant, but only greater value only to those who understand the proper use.”

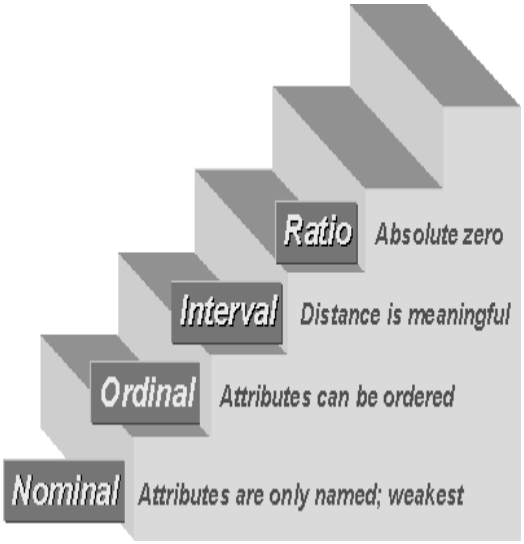
vi) Statistics data should be uniform and homogeneous; one of the important characters of statistical data is comparison. Uniform and homogeneous data can be compared. Unequal and incomparable data leads to wrong results. For e.g. by questionnaires, newspaper collected a data that 60 % of the population favors legalized abortion. But social organization shows that this information is incorrect. In fact more than 70% of people were against it. This error was based on the options of educated people, which constitutes only a small minority and the newspaper published this erroneous conclusions.

3.5 MEASUREMENT

Measurement can be defined as a standardized process of assigning numbers or other symbols to certain characteristics of the objects of interest, according to some pre-specified rules. Measurement often deals with numbers, because mathematical and statistical analysis can be performed only on numbers, and they can be communicated throughout the world in the same form without any translation problems. For a measurement process to be standardized process assignment, two characteristics are necessary. First, there must be one-to-one correspondence between the symbol and the characteristic in the object that is being measured. Second, the rules for assignment must be invariant over time and the objects being measured.

3.5.1 Levels of Measurement:

The level of measurement refers to the relationship among the values that are assigned to the attributes for a variable. What does that mean? Begin with the idea of the variable, in this example "party affiliation." That variable has a number of attributes. Let's assume that in this particular election context the only relevant attributes are "republican", "democrat", and "independent". For purposes of analyzing the results of this variable, we arbitrarily assign the values 1, 2 and 3 to the three attributes. The **level of measurement** describes the relationship among these three values. In this case, we simply are using the numbers as shorter placeholders for the lengthier text terms. We don't assume that higher values mean "more" of something and lower numbers signify "less". We don't assume the value of 2 means that democrats are twice something that republicans are. We don't assume that republicans are in first place or have the highest priority just because they have the value of 1. In this case, we only use the values as a shorter name for the attribute. Here, we would describe the level of Measurement as "nominal".



3.5.2 Why is Level of Measurement Important?

First, knowing the level of measurement helps you decide how to interpret the data from that variable. When you know that a measure is nominal (like the one just described), then you know that the numerical values are just short codes for the longer names. Second, knowing the level of measurement helps you decide what statistical analysis is appropriate on the values that were assigned. If a measure is nominal, then you know that you would never average the data values or do a t-test on the data.

There are typically four levels of measurement that are defined:

i) Nominal
ii) Ordinal
iii) Interval
iv) Ratio

In **nominal** measurement the numerical values just “name” the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number 30 is not more of anything than a player with number 15, and is certainly not twice whatever number 15 is.

In **ordinal** measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0=less than H.S.; 1=some H.S.; 2=H.S. degree; 3=some college; 4=college degree; 5=post college. In this measure, higher numbers mean *more* education. But is distance from 0 to 1 same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure.

In **interval** measurement the distance between attributes *does* have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).

Finally, in **ratio** measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most “count” variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that “...we had twice as many clients in the past six months as we did in the previous six months.”

It's important to recognize that there is a hierarchy implied in the level of measurement idea. At lower levels of measurement, assumptions tend to be less restrictive and data analyses tend to be less sensitive. At each level up the hierarchy, the current level includes all of the qualities of the one below it and adds something new. In general, it is desirable to have a higher level of measurement (e.g., interval or ratio) rather than a lower one (nominal or ordinal).

3.5.3 Scales of Measurement

Scaling is the process of creating a continuum on which objects are located according to the amount of the measured characteristics they possess. An illustration of a scale that is often used in research is the dichotomous scale for sex. The object with male (or female) characteristics is assigned the number 1 and the object with the opposite characteristics is assigned the number 0. This scale meets the requirements of the measurement process in that the assignment is one to one and it is invariant with respect to top time and object. Measurement and scaling are basic tools used in the scientific method and are used in almost every marketing research situation.

3.5.4 Components of Measurement

The assignment of numbers is made according to rules that should correspond to the properties of whatever is being measured. The rule may be very simple, as when a bus route is given a number to distinguish it from other routes. Here, the only property is identity, and any comparisons of numbers are meaningless. This is a nominal scale. At other extreme is the Ratio scale, which has very rigorous properties. In between the extremes are ordinal scales and interval scales.

i) Nominal Scale:

In a nominal scale, objects are assigned to mutually exclusive, labelled categories, but there are no necessary relationships among the categories; that is , no ordering or spacing is implied. If one entity is assigned the same number as another, they are identical with respect to a nominal variable. Otherwise, they are just different. Sex, geographic location, and marital status are nominally scaled variables. The only arithmetic operation that can be performed on such a scale is a count of each category. Thus we can count the number of automobiles dealers in the state of Karnataka or the number of buses seen on a given route in the past hour.

ii) Ordinal Scale:

An ordinal scale is obtained by ranking objects or by arranging them in order with regard to some common variable. The question is simply whether each object has more or less of this variable than some other object. The scale provides information as to how much difference there is between the objects. Because we do not know the amount of difference between objects, the permissible arithmetic operations are limited to statistics such as the median or mode but not median). For example, suppose a

sample of 1,000 consumers ranked five brands of frozen mixed vegetables according to quality. The results for Birds-eye brand were as follows:

Quality	Number of respondents giving Rankings to Bird-Eye brand
Highest	150
Second	300
Third	250
Fourth	200
Lowest	100
Total	1,000

The second quality category is mode; the third category is the median; however it is not possible to compute a mean ranking, because the differences between ordinal scaled values are not necessarily the same. The finishing order in a horse race per class standing illustrates this type of scale. Similarly, brands of frozen vegetables can be ranked according to quality, from highest to lowest.

iii) Interval Scale:

In an interval scale the numbers used to rank the objects also represent equal increments of the attribute being measured. This means that differences can be compared. The difference between 1 and 2 is the same as between 2 and 3, but is only half the difference between 2 and 4. The location of the zero point is not fixed, since zero does not denote the absence of the attribute. Fahrenheit and Celsius temperatures are measured with different interval scales and have different zero points.

Interval scales have very desirable properties, because virtually the entire range of statistical operations can be employed to analyze the resulting number, including addition and subtraction. Consequently, it is possible to compute an arithmetic mean from interval-scale measures.

iv) Ratio Scale:

A ratio scale is a special kind of interval scale that has a meaningful zero point. With such a scale – of weight, market share, or dollars in savings accounts, for example, it is possible to say how many times greater or smaller one object is than another. This is the only type of scale that permits us to make comparisons of absolute magnitude. For example, we can say that an annual income of Rs.80, 000/- is two times as large as

an income of Rs.40, 000/-.

There have been some contemporary efforts to adapt ratio scales to the measurement of social opinion. Some researchers have attempted to use magnitude estimation scales to overcome the loss of information that results when categories arbitrarily constrain the range of opinion. Magnitude scaling of attitudes has been calibrated through numeric estimation.

Check Your Progress:

1. Define Statistics.
2. How data is collected in Statistics?
3. What are the areas where statistics is important?
4. Define measurement.
5. State the four levels of measurement.
6. What is Scales of measurement?

_____

_____

_____

_____

_____

3.6 SOURCES OF DATA COLLECTION

A) The primary data can be collected by the following methods-

1) Direct Personal Observation
2) Indirect oral investigation
3) Estimates from the local sources and correspondence
4) Data through Questionnaires
5) Investigations through enumerator's secondary data.

It is the data collected by some other person or organization for their own use but the investigator also gets it for its own use.

B) Sources of collecting secondary Data. Secondary data can be obtained by

a) Published sources

b) Unpublished Sources

The secondary data can be collected by the following methods.

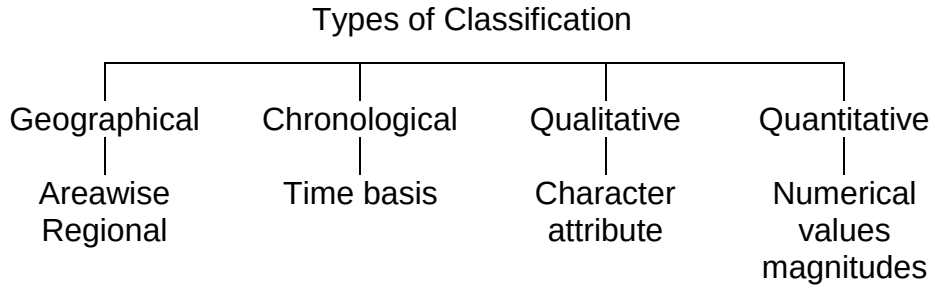
1) Information collected through newspapers and periodicals
2) Information obtained from the publication of trade associations
3) Information obtained from the research papers published by university departments or research bureaus or U.G.C.
4) Information obtained from the official publications of central, state and the local governments dealing with crop, industrial, trade and transport statistics.
5) Information obtained from the official publications of the foreign governments for international orgnisations.

3.7 TYPES AND CLASSIFICATION OF DATA

3.7.1 Types of classification

Classification of data depends on the objectives and the purpose of the enquiry.

There are four types of classification of data:



1) Geographical or Spatial Classification

Under this method the data is classified on the basis of geographical or location series which are arranged on the basis of place are called spatial series.

Examples – classification of data on the basis of states, cites, regions, zones etc.

Literacy figures are collected stat wise

2) Chronological Classification

When the data is classified on the basis of time it is known as chronological classification

For example -population recorded from 1997 to 2006. It is usually recorded with the earliest period.

3) Qualitative Classification

When the data is classified on the basis of different characteristics or attributes is called Qualitative classification

For example – rich-poor, Hindu-Muslim Married-Un married etc.

4) Qualitative Classification

When the data is classified on the basis of phenomenon which is capable of Quantitative measurement is called Quantitative classification.

For example age, height, price, sales, profits.

3.7.2 Classification of Data:

When data is collected through primary methods, it is in the form of unarranged facts and figures, which give no information. It must be sorted out, arranged and properly classified in such a manner which suits the purpose most. The process of arranging things in groups or classes according to their common characteristics or attribute is called the classification of data.

Secrist defined classification of data 'is the process of arranging data into sequences and groups according to their common characteristics or separating them into different but related parts.'

a) Characteristics of a Good Classification

1) It should be exhaustive
2) It should not be ambiguous
3) It should be mutually exclusive
4) It should not be unstable
5) It should be flexible.
6) It should be homogeneous
7) It should reveal classification.

b) Objects of Classification

The object s of classifying data are

1) To ignore unnecessary details.
2) To facilitate comparison between data
3) To study the relationship between several characteristics

4) To prepare tabulation

5) To pinpoints the points of similarity and dissimilarity

c) Advantages of Classification of Data

1) It condenses the data eliminating unnecessary details.

2) It is useful for comparison of data

3) It studies the relationship between several characteristics

4) It facilitates further statistical treatments.

d) Rules for Classification of Data

1) Classes should be clearly defined

2) Classes should be exhaustive

3) Classes should be of equal width

4) Class number should neither be too large nor too small.

3.8 SUMMARY

1. Data is defined as raw facts or observations, typically about physical phenomenon or business transactions. Data are objective measurements of attributes of entities.

2. The conversion of facts into meaningful information is known as data processing.

3. File based systems and Database systems are the two approaches to data management.

4. Statistics is a tool for human being to translate complex facts into simple and understandable statement of facts. Statistical data is of two types – Primary and Secondary data.

5. Statistics involves collection of data, analysis of data and interpretation of facts therefrom.

6. Measurement can be defined as a standardized process of assigning numbers or other symbols to certain characteristics of the objects of interest, according to some pre-specified rules.

7. The level of measurement refers to the relationship among the values that are assigning to the attributes for a variable.

8. There are four levels of measurement Nominal, Ordinal, Interval and Ratio.

9. Scaling is the process of creating a continuum on which objects are located according to the amount of the measured characteristics they possess.

10. There are four types of classification of data : Geographical, Chronological, Qualitative and Quantitative.

11. The process of arranging things in groups or classes according to their common characteristics or attribute is called the classification of data.

3.9 QUESTIONS

1. Discuss the different approaches to data management.
2. Define Statistics. Explain Statistics as a tool in collection of data.
3. Discuss the limitations of Statistics as a tool in collection of data.
4. What do you mean by levels of measurement? And why it is important?
5. What are the different components of Measurement?
6. Discuss the different types of classification of data.
7. Write a note on Classification of data.

4

DATA SAMPLING METHOD

Unit Structure:

4.0 OBJECTIVES

- To understand the meaning of a sample
- To study the criterion of a good sample
- To study the various sampling methods
- To study the different types of Probability or Random sampling
- To study the various types of Non Probability or Non Random sampling

4.1 INTRODUCTION

The sample method consisting of the selecting for study, a portion of the 'universe' with a view to draw conclusions about the 'universe' or 'population' is known as sampling. A statistical sample ideally purports to be a miniature model or replica of the collectively or the population constituted of all the items that the study should principally encompass, that is, the items which potentially hold promise of affording information relevant to the purpose of a given research. Sampling helps in time and cost saving. It also helps in checking their accuracy. But on the other hand it demands exercise of great care caution; otherwise the results obtained may be incorrect or misleading.

4.2 CRITERION OF GOOD SAMPLE

The characteristics of a good sample are described below:

1. **Representative character:** A sample must be representative of the population. Probability sampling technique yield representative sample.

2. **No bias and prejudices :** The selection of the sample should be objective . Sample should be free from bias and prejudices. Then only dependable result can be achieved . Investigator has to be very cautious in this task

3. **Conformity with the subject – matter and meant:** In sample methods the representative units selected, should be as per the subject matter and means.

4. **Accuracy:** Accuracy is defined as the degree to which bias is absent from the sample. An accurate sample is the one which exactly represents the population.

5. **Precision:** The sample must yield precise estimate. Precision is measured by standard error.

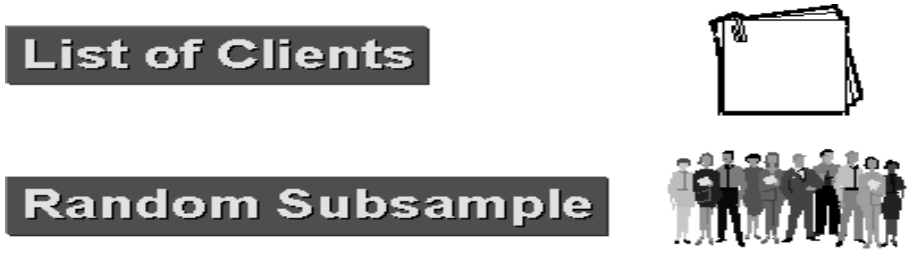
6. **Size:** A good sample must be adequate in size in order to be reliable.

4.3 METHODS OF SAMPLING

Sampling techniques or methods may be classified into two generic types

A] Probability or Random Sampling

Probability sampling method is any method of sampling that utilizes some form of random selection. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen. Humans have long practiced various forms of random selection, such as picking a name out of a hat, or choosing the short straw. These days, we tend to use computers as the mechanism for generating random numbers as the basis for random selection.



Some Definitions:
Before explaining the various probability methods have to define some basic terms. These are:

- **N** = the number of cases in the sampling frame
- **n** = the number of cases in the sample
- ${}_{N}C_{n}$ = the number of combinations (subsets) of n from N
- **f** = n/N = the sampling fraction

That's it. With those terms defined we can begin to define the different probability sampling methods.

Simple Random Sampling:
The simplest form of random sampling is called **simple random sampling**. Pretty tricky, huh? Here's the quick description of simple random sampling:

- **Objective**: To select *n* units out of *N* such that each ${}_{N}C_{n}$ has an equal chance of being selected.

- **Procedure**: Use a table of random numbers, a computer random number generator, or a mechanical device to select the sample.

A somewhat stilted, if accurate, definition. Let's see if we can make it a little more real. How do we select a simple random sample? Let's assume that we are doing some research with a small service agency that wishes to assess client's views of quality of service over the past year. First, we have to get the sampling frame organized. To accomplish this, we'll go through agency records to identify every client over the past 12 months. If we're lucky, the agency has good accurate computerized records and can quickly produce such a list. Then, we have to actually draw the sample. Decide on the number of clients you would like to have in the final sample. For the sake of the example, let's say you want to select 100 clients to survey and that there were 1000 clients over the past 12 months. Then, the sampling fraction is f = n/N = 100/1000 = .10 or 10%. Now, to actually draw the sample, you have several options. You could print off the list of 1000 clients, tear then into separate strips, put the strips in a hat, mix them up real good, close your eyes and pull out the first 100. But this mechanical procedure would be tedious and the quality of the sample would depend on how thoroughly you mixed them up and how randomly

you reached in. Perhaps a better procedure would be to use the kind of ball machine that is popular with many of the state lotteries. You would need three sets of balls numbered 0 to 9, one set for each of the digits from 000 to 999 (if we select 000 we'll call that 1000). Number the list of names from 1 to 1000 and then use the ball machine to select the three digits that selects each person. The obvious disadvantage here is that you need to get the ball machines. (Where do they make those things, anyway? Is there a ball machine industry?).

Neither of these mechanical procedures is very feasible and, with the development of inexpensive computers there is a much easier way. Here's a simple procedure that's especially useful if you have the names of the clients already on the computer. Many computer programs can generate a series of random numbers. Let's assume you can copy and paste the list of client names into a column in an EXCEL spreadsheet. Then, in the column right next to it paste the function =RAND() which is EXCEL's way of putting a random number between 0 and 1 in the cells. Then, sort both columns -- the list of names and the random number -- by the random numbers. This rearranges the list in random order from the lowest to the highest random number. Then, all you have to do is take the first hundred names in this sorted list. pretty simple. You could probably accomplish the whole thing within a minute.

Simple random sampling is simple to accomplish and is easy to explain to others. Because simple random sampling is a fair way to select a sample, it is reasonable to generalize the results from the sample back to the population. Simple random sampling is not the most statistically efficient method of sampling and you may, just because of the luck of the draw, not get good representation of subgroups in a population. To deal with these issues, we have to turn to other sampling methods.

This sampling technique gives each element an equal and independent chance of being selected. An equal chance means equal probability of selection. An independent chance means that the draw of one element will not affect the chances of other elements being selected. The procedure of drawing a simple random sample consists of enumeration of all elements in the population.

1. Preparation of a List of all elements, giving them numbers in serial order 1, 2, 3 and so on, and

2. Drawing sample numbers by using (a) lottery method, (b) a table of random numbers or (c) a computer.

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

Advantages: The advantage of this is that it is one of the easiest methods, all the elements in the population have an equal chance of being selected, simple to understand, does not require prior knowledge of the true composition of the population.

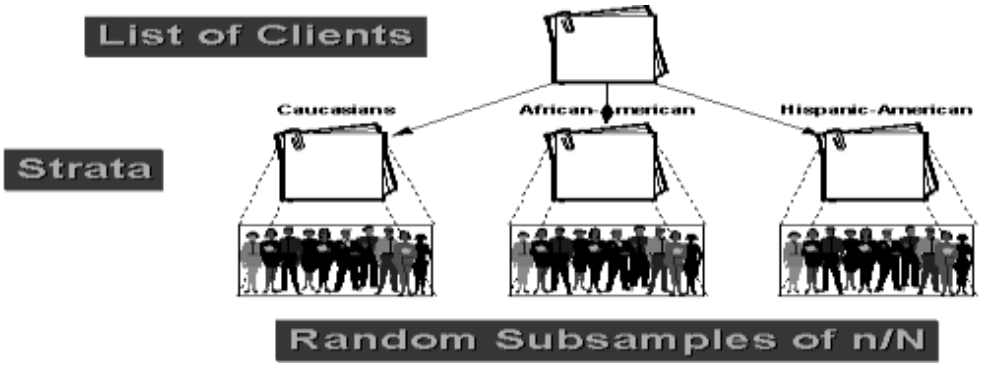
Disadvantages: It is often impractical because of non-availability of population list or of difficulty in enumerating the population, does not ensure proportionate representation and it may be expensive in time and money. The amount of sampling error associated with any sample drawn can easily be computed. But it is greater than that in other probability samples of the same size, because it is less precise than other methods.

Stratified Random Sampling:

Stratified Random Sampling, also sometimes called *proportional* or *quota* random sampling, involves dividing your population into homogeneous subgroups and then taking a simple random sample in each subgroup. In more formal terms:

Objective: Divide the population into non-overlapping groups (i.e., *strata*) N_1, N_2, N_3, ... N_i, such that $N_1 + N_2 + N_3 + ... + N_i = N$. Then do a simple random sample of f = n/N in each strata.

There are several major reasons why you might prefer stratified sampling over simple random sampling. First, it assures that you will be able to represent not only the overall population, but also key subgroups of the population, especially small minority groups. If you want to be able to talk about subgroups, this may be the only way to effectively assure you'll be able to. If the subgroup is extremely small, you can use different sampling fractions (f) within the different strata to randomly over-sample the small group (although you'll then have to weight the within-group estimates using the sampling fraction whenever you want overall population estimates). When we use the same sampling fraction within strata we are conducting *proportionate* stratified random sampling. When we use different sampling fractions in the strata, we call this *disproportionate* stratified random sampling. Second, stratified random sampling will generally have more statistical precision than simple random sampling. This will only be true if the strata or groups are homogeneous. If they are, we expect that the variability within-groups is lower than the variability for the population as a whole. Stratified sampling capitalizes on that fact.



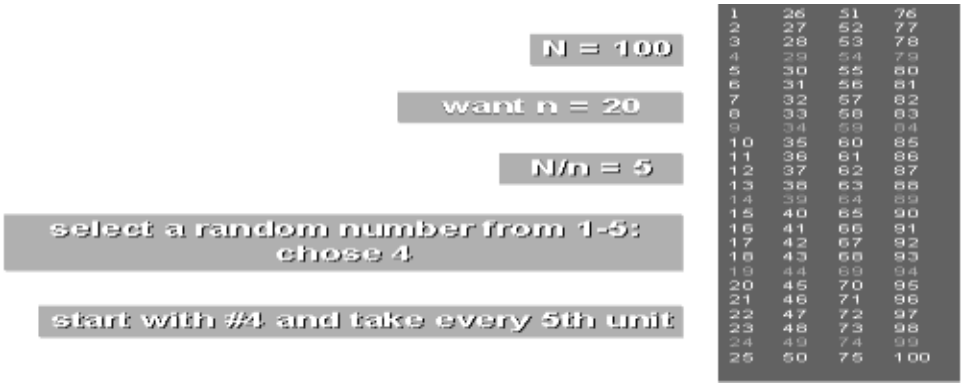
For example, let's say that the population of clients for our agency can be divided into three groups: Caucasian, African-American and Hispanic-American. Furthermore, let's assume that both the African-Americans and Hispanic-Americans are relatively small minorities of the clientele (10% and 5% respectively). If we just did a simple random sample of n=100 with a sampling fraction of 10%, we would expect by chance alone that we would only get 10 and 5 persons from each of our two smaller groups. And, by chance, we could get fewer than that! If we stratify, we can do better. First, let's determine how many people we want to have in each group. Let's say we still want to take a sample of 100 from the population of 1000 clients over the past year. But we think that in order to say anything about subgroups we will need at least 25 cases in each group. So, let's sample 50 Caucasians, 25 African-Americans, and 25 Hispanic-Americans. We know that 10% of the population, or 100 clients, are African-American. If we randomly sample 25 of these, we have a within-stratum sampling fraction of 25/100 = 25%. Similarly, we know that 5% or 50 clients are Hispanic-American. So our within-stratum sampling fraction will be 25/50 = 50%. Finally, by subtraction we know that there are 850 Caucasian clients. Our within-stratum sampling fraction for them is 50/850 = about 5.88%. Because the groups are more homogeneous within-group than across the population as a whole, we can expect greater statistical precision (less variance). And, because we stratified, we know we will have enough cases from each group to make meaningful subgroup inferences.

This is an improved type of random or probability sampling. In this method, the population is sub-divided into homogenous groups or strata, and from each stratum, random sample is drawn. E.g., university students may be divided on the basis of discipline, and each discipline group may again be divided into juniors and seniors. Stratification is necessary for increasing a sample's statistical efficiency, providing adequate data for analyzing the various sub-populations and applying different methods to different strata. The stratified random sampling is appropriate for a large heterogeneous population. Stratification process involves three major decisions. They are stratification base or bases, number of strata and strata sample sizes.

1. Systematic Random Sampling

Here are the steps you need to follow in order to achieve a **systematic random sample**:

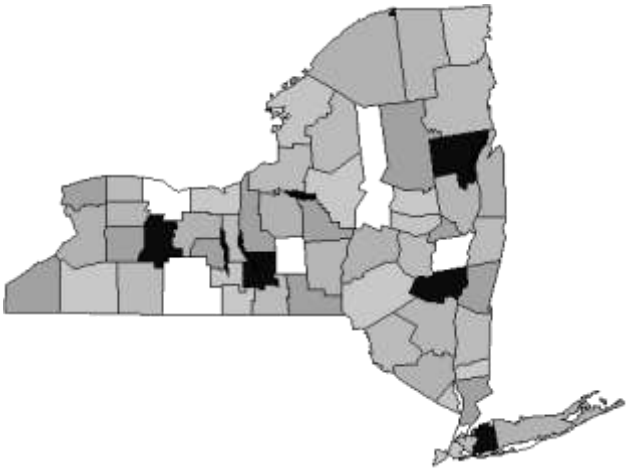
- number the units in the population from 1 to N
- decide on the n (sample size) that you want or need
- k = N/n = the interval size
- randomly select an integer between 1 to k
- then take every kth unit



All of this will be much clearer with an example. Let's assume that we have a population that only has N=100 people in it and that you want to take a sample of n=20. To use systematic sampling, the population must be listed in a random order. The sampling fraction would be f = 20/100 = 20%. in this case, the interval size, k, is equal to N/n = 100/20 = 5. Now, select a random integer from 1 to 5. In our example, imagine that you chose 4. Now, to select the sample, start with the 4th unit in the list and take every k-th unit (every 5th, because k=5). You would be sampling units 4, 9, 14, 19, and so on to 100 and you would wind up with 20 units in your sample.

For this to work, it is essential that the units in the population are randomly ordered, at least with respect to the characteristics you are measuring. Why would you ever want to use systematic random sampling? For one thing, it is fairly easy to do. You only have to select a single random number to start things off. It may also be more precise than simple random sampling. Finally, in some situations there is simply no easier way to do random sampling. For instance, I once had to do a study that involved sampling from all the books in a library. Once selected, I would have to go to the shelf, locate the book, and record when it last circulated. I knew that I had a fairly good sampling frame in the form of the shelf list (which is a card catalog where the entries are arranged in the order they occur on the shelf). To do a simple random sample, I could have estimated the total number of books and generated random numbers to draw the sample; but how would

I find book #74,329 easily if that is the number I selected? I couldn't very well count the cards until I came to 74,329! Stratifying wouldn't solve that problem either. For instance, I could have stratified by card catalog drawer and drawn a simple random sample within each drawer. But I'd still be stuck counting cards. Instead, I did a systematic random sample. I estimated the number of books in the entire collection. Let's imagine it was 100,000. I decided that I wanted to take a sample of 1000 for a sampling fraction of 1000/100,000 = 1%. To get the sampling interval k, I divided N/n = 100,000/1000 = 100. Then I selected a random integer between 1 and 100. Let's say I got 57. Next I did a little side study to determine how thick a thousand cards are in the card catalog (taking into account the varying ages of the cards). Let's say that on average I found that two cards that were separated by 100 cards were about .75 inches apart in the catalog drawer. That information gave me everything I needed to draw the sample. I counted to the 57th by hand and recorded the book information. Then, I took a compass. (Remember those from your high-school math class? They're the funny little metal instruments with a sharp pin on one end and a pencil on the other that you used to draw circles in geometry class.) Then I set the compass at .75", stuck the pin end in at the 57th card and pointed with the pencil end to the next card (approximately 100 books away). In this way, I approximated selecting the 157th, 257th, 357th, and so on. I was able to accomplish the entire selection procedure in very little time using this systematic random sampling approach. I'd probably still be there counting cards if I'd tried another random sampling method. (Okay, so I have no life. I got compensated nicely, I don't mind saying, for coming up with this scheme.)



a) Proportionate stratified sampling: This sampling involves drawing a sample from each stratum in proportion to the latter's share in the total population. It gives proper representation to each stratum and its statistical efficiency is generally higher. This method is therefore very popular.

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

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

b) Disproportionate stratified random sampling: This method does not give proportionate representation to strata. It necessarily involves giving over-representation to some strata and under-representation to others. The desirability of disproportionate sampling is usually determined by three factors, viz, (a) the sizes of strata, (b) internal variances among strata, and (c) sampling costs.

Suitability: This method is used when the population contains some small but important subgroups, when certain groups are quite heterogeneous, while others are homogeneous and when it is expected that there will be appreciable differences in the response rates of the subgroups in the population.

Advantages: The advantages of this type is it is less time consuming and facilitates giving appropriate weighing to particular groups which are small but more important.

Disadvantages: The disadvantage is that it does not give each stratum proportionate representation, requires prior knowledge of composition of the population, is subject to classification errors and its practical feasibility is doubtful.

2. Cluster Sampling:

Cluster (Area) Random Sampling

The problem with random sampling methods when we have to sample a population that's disbursed across a wide geographic region is that you will have to cover a lot of ground geographically in order to get to each of the units you sampled. Imagine taking a simple random sample of all the residents of New York State in order to conduct personal interviews. By the luck of the draw you will wind up with respondents who come from all over the state. Your interviewers are going to have a lot of traveling to do. It is for precisely this problem that **cluster or area random sampling** was invented.

In cluster sampling, we follow these steps:

- divide population into clusters (usually along geographic boundaries)
- randomly sample clusters
- measure all units within sampled clusters

For instance, in the figure we see a map of the counties in New York State. Let's say that we have to do a survey of town governments that will require us going to the towns personally. If we do a simple random sample state-wide we'll have to cover the entire state geographically. Instead, we decide to do a cluster sampling of five counties (marked in red in the figure). Once these are selected, we go to *every* town government in the five areas. Clearly this strategy will help us to economize on our mileage. Cluster or area sampling, then, is useful in situations like this, and is done primarily for efficiency of administration. Note also, that we probably don't have to worry about using this approach if we are conducting a mail or telephone survey because it doesn't matter as much (or cost more or raise inefficiency) where we call or send letters to.

It means random selection of sampling units consisting of population elements. Each such sampling unit is a cluster of population elements. Then from each selected sampling unit, a sample of population elements is drawn by either simple random selection or stratified random selection. Where the population elements are scattered over a wide area and a list of population elements is not readily available, the use of simple or stratified random sampling method would be too expensive and time-consuming. In such cases cluster sampling is usually adopted.

The cluster sampling process involves: identify clusters, examine the nature of clusters, and determine the stages.

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

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

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

Area Sampling

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

3. Multi-Stage and Sub-Sampling

The four methods we've covered so far -- simple, stratified, systematic and cluster -- are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this multi-stage sampling**.**

For example, consider the idea of sampling New York State residents for face-to-face interviews. Clearly we would want to do some type of cluster sampling as the first stage of the process. We might sample townships or census tracts throughout the state. But in cluster sampling we would then go on to measure everyone in the clusters we select. Even if we are sampling census tracts we may not be able to measure *everyone* who is in the census tract. So, we might set up a stratified sampling process within the clusters. In this case, we would have a two-stage sampling process with stratified samples within cluster samples. Or, consider the problem of sampling students in grade schools. We might begin with a national sample of school districts stratified by economics and educational level. Within selected districts, we might do a simple random sample of schools. Within schools, we might do a simple random sample of classes or grades. And, within classes, we might even do a simple random sample of students. In this case, we have three or four stages in the sampling process and we use both stratified and simple random sampling. By combining different sampling methods we are able to achieve a rich variety of probabilistic sampling methods that can be used in a wide range of social research contexts.

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

In multi-stage sampling method, sampling is carried out in two or more stages. The population is regarded as being composed of a number of second stage units and so forth. That is, at each stage, a sampling unit is a cluster of the sampling units of the subsequent stage. First, a sample of the first stage sampling units is drawn, then from each of the selected first stage sampling unit, a sample of the second stage sampling units is drawn. The procedure continues down to the final sampling units or population elements. Appropriate random sampling method is adopted at each stage. It is appropriate where the population is scattered over a wider geographical area and no frame or list is available for sampling. It is also useful when a survey has to be made within a limited time and cost budget. The major disadvantage is that the procedure of estimating sampling error and cost advantage is complicated.

Sub-sampling is a part of multi-stage sampling process. In a multi-stage sampling, the sampling in second and subsequent stage frames is called sub-sampling. Sub-sampling balances the two conflicting effects of clustering i.e., cost and sampling errors.

Check Your Progress:

4. What do you mean by sampling?
5. Discuss the characteristics of a good sample.
6. State the two methods of sampling.
7. Define Probability sampling.
8. Distinguish between Proportionate Stratified sampling and Disproportionate stratified random sampling.
9. What do you understand by Area sampling?

_____

_____

_____

_____

B] Non-probability or non-random sampling

Non-probability sampling or non-random sampling is not based on the theory of probability. This sampling does not provide a chance of selection to each population element. The difference between non-probability and probability sampling is that non-probability sampling does not involve *random* selection and probability sampling does. Does that mean that non-probability samples aren't representative of the population? Not necessarily. But it does mean that non-probability samples cannot depend upon the rationale of probability theory. At least with a probabilistic sample, we know the odds or probability that we have represented the population well. We are able to estimate confidence intervals for the statistic. With non-probability samples, we may or may not represent the population well, and it will often be hard for us to know how well we've done so. In general, researchers prefer probabilistic or random sampling methods over non-probabilistic ones, and consider them to be more accurate and rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Here, we consider a wide range of non-probabilistic alternatives.

We can divide non-probability sampling methods into two broad types: *accidental* or *purposive*. Most sampling methods are purposive in nature because we usually approach the sampling problem with a specific plan in mind. The most important distinctions among these types of sampling methods are the ones between the different types of purposive sampling approaches.

1. Accidental, Haphazard or Convenience Sampling:

One of the most common methods of sampling goes under the various titles listed here. I would include in this category the traditional “man on the street” (of course, now it's probably the “person on the street”) interviews conducted frequently by television news programs to get a quick (although non representative) reading of public opinion. I would also argue that the typical use of college students in much psychological research is primarily a matter of convenience. (You don't really believe that psychologists use college students because they believe they're representative of the population at large, do you?). In clinical practice, we might use clients who are available to us as our sample. In many research contexts, we sample simply by asking for volunteers. Clearly, the problem with all of these types of samples is that we have no evidence that they are representative of the populations we're interested in generalizing to -- and in many cases we would clearly suspect that they are not.

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

Disadvantages: The demerits are it does not ensure a selection chance to each population unit. The selection probability sample may not be a representative one. The selection probability is unknown. It suffers from sampling bias which will distort results. The reasons for usage of this sampling are when there is no other feasible alternative due to non-availability of a list of population, when the study does not aim at generalizing the findings to the population, when the costs required for probability sampling may be too large, when probability sampling required more time, but the time constraints and the time limit for completing the study do not permit it. It may be classified into:

Convenience or Accidental Sampling:

It means selecting sample units in a just 'hit and miss' fashion E.g., interviewing people whom we happen to meet. This sampling also means selecting whatever sampling units are conveniently available, e.g., a teacher may select students in his class. This method is also known as accidental sampling because the respondents whom the researcher meets accidentally are included in the sample.

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

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

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

2. Purposive (or Judgement) Sampling:

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

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

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

Disadvantage: It is less efficient for generalizing, does not ensure the representativeness, requires more prior extensive information and does not lend itself for using inferential statistics.

a. Quota Sampling:

In quota sampling, you select people non-randomly according to some fixed quota.

There are two types of quota sampling: *proportional* and *non proportional*.

i) Proportional quota sampling:

Proportional quota sampling you want to represent the major characteristics of the population by sampling a proportional amount of each. For instance, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop. So, if you've already got the 40 women for your sample, but not the sixty men, you will continue to sample men but even if legitimate women respondents come along, you will not sample them because you have already "met your quota." The problem here (as in much purposive sampling) is that you have to decide the specific characteristics on which you will base the quota. Will it be by gender, age, education race, religion, etc.?

ii) Non-proportional quota sampling:

Non-proportional quota sampling is a bit less restrictive. In this method, you specify the minimum number of sampled units you want in each category. here, you're not concerned with having numbers that match the proportions in the population. Instead, you simply want to have enough to assure that you will be able to talk about even small groups in the population. This method is the non-probabilistic analogue of stratified random sampling in that it is typically used to assure that smaller groups are adequately represented in your sample. This is a form of convenient sampling involving selection of quota groups of accessible sampling units by traits such as sex, age, social class, etc. it is a method of stratified sampling in which the selection within strata is non-random. It is this Non-random element that constitutes its greatest weakness.

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

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

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

4.4 SAMPLING AND NON-SAMPLING ERRORS

A sample survey implies the study of small proportions of the total universe and drawing inference about the population. It is natural that there would be certain amount of inaccuracy or errors. Such errors are known as sampling errors or sampling fluctuations. If a census is taken, sampling errors could be expected to disappear.

(A) Sampling errors:

The errors which arise due to the use of sampling surveys are known as the sampling errors. Even when a sample is random one, it may be exactly representative of teh population from which it is chosen. This is because samples are seldom, if ever, perfect miniature of the populations. However these errors can be controlled.

Two types of sampling errors:

i. **Biased errors:** Those errors which arise as a result of any bias or prejudice of the person in selecting a particular sampling method e.g., purposive sampling method may be adopted in place of a simple random sampling method. As a result of such a selection, some errors are bound to arise, and they are known as biased sampling errors, or cumulative errors or non-compensating errors. As biased or prejudice forms a constant component of error that does not decrease in a large population as the number in the sample increases i.e., such errors are likely to increase with an increase in the size of the sample. These errors may arise due to:

1. Faulty process of selection: Faulty selection of the sample may give rise to bias in a number of ways, such as purposive sampling, selection of sample in a haphazard way, substitution of the selected item in the sample by another, incomplete investigation or response etc.

2. Faulty work during the collection of formation: During the process of collecting the actual information in a survey , certain inaccuracies may creep in. These may arise due to the improper formulation of the decision, problem wrongly defining the population, specifying wrong decision, securing an inadequate frame, poorely designed questionnaire, an ill-trained interviewer, failure of a respondent's memory, unorganised collection procedure, faulty editing or coding the response.

3. Faulty method of analysis: Faulty methods of analysis may also introduce bias. Such bias can be avoided by adopting the proper methods of analysis.

Suggestions to reduce biases and improving sampling designs are:

i. Manageable and specific problem selection

ii. Intensive study, verification and reporting of methodological biases

iii. Systematic documentation of related research

iv. Greater investment in enumeration

v. Effective pre-testing

vi. Use of complementary research methods

vii. Replication

ii. Unbiased errors:

These errors arise due to chance differences, between the members of the population included in the sample and those not included. It is known as random sampling error. The random sampling error decreases on an average as the size of the sample increases. Such error is, therefore, also known as non cumulative or compensating error.

(B) Non-sampling errors:

This type of error can occur in any survey, whether it be a complete enumeration or sampling. Non-sampling errors include biases and mistakes. Some of the factors responsible for the non-sampling errors are

1. vague definition of population

2. vague questionnaire

3. vague conception regarding the information desired

4. inappropriate statistical unit

5. inaccurate/inappropriate methods of interview

6. observation or measurement

7. errors in data processing operations such as coding, punching, verification, tabulation etc.

8. errors committed during presentation and printing of tabulated results.

Non-sampling errors tend to increase with sample size and require to be controlled and reduced to a minimum.

4.5 SUMMARY

1. The sample method consisting of the selecting for study, a portion of the Universe with a view to draw conclusions about the Universe or population is known as sampling.
2. Sampling techniques or methods can be classified into Probability or Random sampling and Non-Probability or Non-Random sampling.
3. Probability sampling method is any method of sampling that utilizes some form of random selection.
4. Simple Random sampling is the simplest form of random sampling. This sampling technique gives each element an equal and independence chance of being selected.
5. Stratified Random sampling sometimes called proportional or quota random sampling, involves dividing entire population into homogeneous subgroups and then taking a simple random sample in each subgroup. It is appropriate for a large heterogeneous population.
6. Where the population elements are scattered over a wide area and a list of population elements is not readily available, Cluster sampling is usually adopted because it is easier and more convenient.
7. In multi-stage sampling method, sampling is carried out in two or more stages. At each stage, a sampling unit is a cluster of the sampling units of the subsequent stage. Appropriate random sampling method is adopted at each stage. It is appropriate where the population is scattered over a wider geographical area and no frame or list is available for sampling. In a multi-stage sampling, the sampling in second and subsequent stage frames is called sub-sampling.
8. In applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Therefore Non probability or Non Random sampling is used. It is divided into two broad types: accidental or purposive.

9. Sampling errors are those which arise due to the use of sampling surveys. These are of two types: Biased and Unbiased errors. Non-sampling errors can occur in any survey, whether it be a complete enumeration or sampling. It include biases and mistakes.

4.6 QUESTIONS

1. What is the significance of Sampling in research?
2. What are the types of Probability or random sampling?
3. What is the Non-probability or non random sampling?
4. Discuss sampling and non-sampling errors.

5

Module 3
DATA COLLECTION METHODS

PRIMARY DATA COLLECTION

Unit Structure:

5.0 OBJECTIVES

- To study the meaning of primary data
- To understand the various sources of primary data collection
- To study the important methods of primary data collection
- To study the types of surveys
- To understand the meaning and types of questionnaire
- To study the sample of questionnaires
- To study the meaning of interviews
- To study the preparation for the interviews

5.1 INTRODUCTION

Primary data collection is necessary when a researcher cannot find the data needed in secondary sources. Market researchers are interested in primary data about demographic/ socioeconomic characteristics, attitudes / opinions / interests, awareness/knowledge, intentions, motivation, and behavior. Three

basic means of obtaining primary data are observation, surveys, and experiments. The choice will be influenced by the nature of the problem and by the availability of time and money.

5.2 PRIMARY SOURCES OF DATA

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

a. Observation

Observation means that the situation of interest is checked and a person or some mechanical device records the relevant facts, actions, or behaviors. Accurate data about what consumers do in certain situations is provided by observation. Observation does not tell why it happened.

b. Mechanical Approaches

Mechanical approaches are reliable data collection instruments because they provide objective measures. Data on the factors influencing product sales, such as competitor advertising and other promotional activities can be effectively assessed. Information can be obtained on a specific store or all the stores in a system, enabling rapid and effective comparisons at various local, regional and national levels. The information is available continuously and enables firms to plan down to the individual store level.

Scanner and bar coding technologies form the basis for capturing marketing information at the retail level. Scanners are electronic devices at retail checkouts that read the bar code for each item bought. They provide up-to-the-minute data on product purchases by item and also by household. Telecommunications can transmit the information directly to the manufacturer and shorten the communications cycle from weeks to minutes. With this information the manufacturer can develop a profile for each retailer and establish the optimum retail inventory for each location. The optimum inventory ensures stocking of merchandise that customers buy with a minimum amount of inventory investment for the retailer. Combining the retailer's information with the manufacturer's database yields local promotional mailings, fine-tuned shelf displays, and redesigned store layouts.

Other mechanical devices include video cameras, Nielsen People Meters, and single-source data systems that link consumers' exposure to television advertising, sales promotion, and other marketing efforts with their store purchases (Behavior Scan and Info Scan of Information Resources, Inc.). Furthermore, measurements might be taken of respondents' eye movements, pulse rates, or other physical reactions to advertisements.

c. Personal Approaches

Marketers can learn by personally observing or watching actions and situations. For example, when an organization is choosing a new location, it would observe the neighborhood conditions. Also, marketers of pet products and baby products are extremely interested in how respondents react to new products, but obviously cannot ask them to describe their opinions or to fill out surveys. They must depend on observational research.

d. Surveys

Surveys or questioning involve using a questionnaire (data collection instrument) to ask respondents questions to secure the desired information. Questionnaires may be administered by mail, over the telephone, by computer, or in person. Limitations of surveys include opportunities for error in construction and administering of a questionnaire, expense, and time needed to conduct a survey. Respondents may not respond, may be unable to respond, or may give misleading responses.

e. Mail

Mail interviews can be used to collect large amounts of data and have a low cost per respondent. Respondents can see a concept, read a description, and think about it at their leisure. There is no interviewer bias. However, the questionnaires are not flexible, cannot be adapted to individual respondents, and generally have low response rates. The researcher has no control over who completes the questionnaire.

f. Telephone

Telephone interviews are easy to administer and allow data to be collected quickly at a relatively low cost. The interviewer can clarify the questions. Response rates tend to be higher and telephone interviewing allows for greater sample control. However, it is more expensive than a mail questionnaire. The presence of an interviewer on the phone may bias responses since respondents may be unwilling to discuss personal information. Also, respondents can't see product. A major limitation is that they must be short.

g. Computer

Advances in computers and technology have led to sophisticated data collection methods. Computer and online interviewing allow rapid data collection from dispersed populations at a low cost.

h. Personal

Personal interviews may be conducted one-on-one or with a focus group. A personal interview is a direct, face-to-face interview between the interviewer and the respondent. In the past, personal interviews were conducted door-to-door. Today, most personal interviews (one-on-one) are conducted in malls and are referred to as mall intercept. Personal interviews are the most flexible since interviewers can clarify questions and probe for answers. Respondents can see a concept as well as read a description. More information can normally be obtained through observation of the respondent's surroundings. Personal interviewing is expensive, yet it offers a great deal of flexibility and allows for visual stimuli.

A focus group is a small group of people, carefully selected, who represent a specific target audience. They are used to generate concepts and hypotheses. The strength of focus groups is found in the group discussion and interaction. Focus group interviews are a popular way of gaining insight into consumer thoughts and feelings about a product. In the past, focus groups were regarded mainly as a simple and quick way of asking any group of respondents, usually in someone's home, to answer questions about a product. Today, focus groups are an important source of qualitative research. Advance preparation ensures that the facility, moderator, and respondents are of high quality. An example of a technique used in a focus group is a projective technique in which a People Board is used to obtain attitudes through photograph associations and forced relationships. Participants indicate which of several images in a category relate to the subject at hand. The findings from a focus group are useful for general information but do not suffice to give absolute quantifiable information.

A panel is a fixed sample of individuals from who repeated measurements are taken over time with respect to the same variables. An example is MRCA, which has a 12,000 household panel that is representative of the national census in terms of significant demographics. Surveys of the panel conducted frequently throughout the year provide a means to measure relatively small changes in household purchases and product usage. Another example is the consumer diary. Diaries are especially appropriate for answering questions on brand penetration and loyalty. This approach indicates which factors influence purchasing behaviors, such as price and advertising, and

where purchases are made -- supermarkets, discount stores, drugstores. Firms in packaged goods, apparel, home furnishings, financial services, travel, and entertainment use this method.

i. Experiments

In an experiment, a researcher selects matched groups, gives them different experimental treatments controlling for other related factors, and checks for differences in the responses of the experimental group and the control group. Experimental research attempts to explain cause-and-effect relationships. Data in an experiment may be collected through observation and surveys. An experiment can be done in either a laboratory or field setting. In a laboratory experiment, the researcher has complete control during the experiment. A field experiment is conducted under more realistic conditions. For example, if a charitable organization wanted to see whether inclusion of return-address labels affected donors' responses to a mail solicitation, it could select similar sets of donors and send them the donation solicitation with and without labels to see if one method is more effective than the other is.

Advantages of primary data:

1. It is original source of data
2. It is possible to capture the changes occurring in the course of time.
3. It is flexible to the advantage of researcher.
4. Extensive research study is based of primary data.

Disadvantages of primary data:

1. Primary data is expensive to obtain
2. It is time consuming
3. It requires extensive research personnel who are skilled.
4. It is difficult to administer.

5.3 METHODS OF COLLECTING PRIMARY DATA

Primary data are directly collected by the researcher 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 required data are not available from secondary sources and they have to be directly gathered from the primary sources. In such cases where the available data are inappropriate, inadequate or obsolete, primary data have to be gathered. They include: 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 managements studies, business management studies etc.

There are various methods of 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 instruments used for the method. For example, a schedule is used for interviewing. The important methods are

(a) observation, (b) interviewing, (c) mail survey, (d) experimentation, (e) simulation and (f) projective technique. Each of these methods is discussed in detail in the subsequent sections in the later chapters.

Check Your Progress:

1. What do you understand by primary data?
2. State the important methods of primary data collection.

5.4 TYPES OF SURVEYS

Surveys can be divided into two broad categories: the **questionnaire** and the **interview**. Questionnaires are usually paper-and-pencil instruments that the respondent completes. Interviews are completed by the interviewer based on the respondent says. Sometimes, it's hard to tell the difference between a questionnaire and an interview. For instance, some people think that questionnaires always ask short closed-ended questions while interviews always ask broad open-ended ones. But you will see questionnaires with open-ended questions (although they do tend to be shorter than in interviews) and there will often be a series of closed-ended questions asked in an interview.

Survey research has changed dramatically in the last ten years. We have automated telephone surveys that use random dialing methods. There are computerized kiosks in public places that allows people to ask for input. A whole new variation of group interview has evolved as focus group methodology. Increasingly, survey research is tightly integrated with the delivery of service. Your hotel room has a survey on the desk. Your waiter presents a short customer satisfaction survey with your check. You get a call for an interview several days after your last call to a computer company for technical assistance. You're asked to complete a short survey when you visit a web site. Here, I'll describe the major types of questionnaires and interviews, keeping in mind that technology is leading to rapid evolution of methods. We'll discuss the relative advantages and disadvantages of these different survey types in Advantages and Disadvantages of Survey Methods.

5.5 QUESTIONNAIRES

When most people think of questionnaires, they think of the **mail survey**. All of us have, at one time or another, received a questionnaire in the mail. There are many advantages to mail surveys. They are relatively inexpensive to administer. You can send the exact same instrument to a wide number of people. They allow the respondent to fill it out at their own convenience. But there are some disadvantages as well. Response rates from mail surveys are often very low. And, mail questionnaires are not the best vehicles for asking for detailed written responses.

A second type is the **group administered questionnaire**. A sample of respondents is brought together and asked to respond to a structured sequence of questions. Traditionally, questionnaires were administered in group settings for convenience. The researcher could give the questionnaire to those who were present and be fairly sure that there would be a high response rate. If the respondents were unclear about the meaning of a question they could ask for clarification. And, there were often organizational settings where it was relatively easy to assemble the group (in a company or business, for instance).

What's the difference between a group administered questionnaire and a group interview or focus group? In the group administered questionnaire, each respondent is *handed an instrument* and asked to complete it while in the room. Each respondent completes an instrument. In the group interview or focus group, the interviewer facilitates the session. People work as a group, listening to each other's comments and answering the

questions. Someone takes notes for the entire group -- people don't complete an interview individually.

A less familiar type of questionnaire is the **household drop-off** survey. In this approach, a researcher goes to the respondent's home or business and hand over the respondent the instrument. In some cases, the respondent is asked to mail it back or the interview returns to pick it up. This approach attempts to blend the advantages of the mail survey and the group administered questionnaire. Like the mail survey, the respondent can work on the instrument in private, when it's convenient. Like the group administered questionnaire, the interviewer makes personal contact with the respondent -- they don't just send an impersonal survey instrument. And, the respondent can ask questions about the study and get clarification on what is to be done. Generally, this would be expected to increase the percent of people who are willing to respond.

5.5.1 Question Issues

Sometimes the nature of what you want to ask respondents will determine the type of survey you select.

- **What types of questions can be asked?**

Are you going to be asking personal questions? Are you going to need to get lots of detail in the responses? Can you anticipate the most frequent or important types of responses and develop reasonable closed-ended questions?

- **How complex will the questions be?**

Sometimes you are dealing with a complex subject or topic. The questions you want to ask are going to have multiple parts. You may need to branch to sub-questions.

- **Will screening questions be needed?**

A screening question may be needed to determine whether the respondent is qualified to answer your question of interest. For instance, you wouldn't want to ask someone their opinions about a specific computer program without first "screening" them to find out whether they have any experience using the program. Sometimes you have to screen on several variables (e.g., age, gender, experience). The more complicated the screening, the less likely it is that you can rely on paper-and-pencil instruments without confusing the respondent.

- **Can question sequence be controlled?**

Is your survey one where you can construct in advance a reasonable sequence of questions? Or, are you doing an initial

exploratory study where you may need to ask lots of follow-up questions that you can't easily anticipate?

- **Will lengthy questions be asked?**

If your subject matter is complicated, you may need to give the respondent some detailed background for a question. Can you reasonably expect your respondent to sit still long enough in a phone interview to ask your question?

- **Will long response scales be used?**

If you are asking people about the different computer equipment they use, you may have to have a lengthy response list (CD-ROM drive, floppy drive, mouse, touch pad, modem, network connection, external speakers, etc.). Clearly, it may be difficult to ask about each of these in a short phone interview.

5.5.2 Sample of questionnaires:

1. Dichotomous Questions:

When a question has two possible responses, we consider it **dichotomous**. 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:

Do you believe that the death penalty is ever justified?

____Yes

____No

Please enter your gender:

☐ Male ☐ Female

2. Questions Based on Level Of Measurement

We can also classify questions in terms of their level of measurement. For instance, we might measure occupation using a **nominal** question. Here, the number next to each response has no meaning except as a placeholder for that response. The choice of a "2" for a lawyer and a "1" for a truck driver is arbitrary -- from the numbering system used we can't infer that a lawyer is "twice" something that a truck driver is.

Occupational Class:

1 = truck driver
2 = lawyer
3 = etc.

We might ask respondents to rank order their preferences for presidential candidates using an **ordinal** question:

Rank the candidates in order of preference from best to worst...

___ Bob Dole
___ Bill Clinton
___ Newt Gingrich
___ Al Gore

We want the respondent to put a 1, 2, 3 or 4 next to the candidate, where 1 is the respondent's first choice. Note that this could get confusing. We might want to state the prompt more explicitly so the respondent knows we want a number from one to 4 (the respondent might check their favorite candidate, or assign higher numbers to candidates they prefer more instead of understanding that we want rank ordering).

We can also construct survey questions that attempt to measure on an **interval** level. One of the most common of these types is the traditional 1-to-5 rating (or 1-to-7, or 1-to-9, etc.). This is sometimes referred to as a **Likert response scale** (see Likert Scaling). Here, we see how we might ask an opinion question on a 1-to-5 bipolar scale (it's called bipolar because there is a neutral point and the two ends of the scale are at opposite positions of the opinion):

The death penalty is justifiable under some circumstances.

1	2	3	4	5
strongly disagree	disagree	neutral	agree	strongly agree

Another interval question uses an approach called the **semantic differential**. Here, an object is assessed by the respondent on a set of bipolar adjective pairs (using 5-point rating scale):

Please state your opinions on national health insurance on the scale below

	very much	some-what	neither	some-what	very much	
interesting	☐	☐	☐	☐	☐	*boring*
simple	☐	☐	☐	☐	☐	*complex*
uncaring	☐	☐	☐	☐	☐	*caring*
useful	☐	☐	☐	☐	☐	*useless*

etc.

Finally, we can also get at interval measures by using what is called a **cumulative or Guttman scale** (see Guttman Scaling). Here, the respondent checks each item with which they agree. The items themselves are constructed so that they are cumulative -- if you agree to one, you probably agree to all of the ones above it in the list:

Please check each statement that you agree with:

___Are you willing to permit immigrants to live in your country?
___Are you willing to permit immigrants to live in your community?
___Are you willing to permit immigrants to live in your neighborhood?
___Would you be willing to have an immigrant live next door to you?
___Would you let your child marry an immigrant?

Check Your Progress:

1. State the two types of surveys.
2. What are the different types of questionnaire?
3. What do you mean by Dichotomous questions?

_____

_____

_____

_____

_____

_____

5.6 INTERVIEWS

Interviews are a far more personal form of research than questionnaires. In the **personal interview**, the interviewer works directly with the respondent. Unlike with mail surveys, the interviewer has the opportunity to probe or ask follow-up questions. And, interviews are generally easier for the respondent, especially if what is sought is opinions or impressions. Interviews can be very time consuming and they are resource intensive. The interviewer is considered a part of the measurement instrument and interviewers have to be well trained in how to respond to any contingency.

Almost everyone is familiar with the **telephone interview**. Telephone interviews enable a researcher to gather information rapidly. Most of the major public opinion polls that are reported were based on telephone interviews. Like personal interviews, they allow for some personal contact between the interviewer and the respondent. And, they allow the interviewer to ask follow-up questions. But they also have some major disadvantages. Many people don't have publicly-listed telephone numbers. Some don't have telephones. People often don't like the intrusion of a call to their homes. And, telephone interviews have to be relatively short or people will feel imposed upon.

Concept of Interviews

Interviews are among the most challenging and rewarding forms of measurement. They require a personal sensitivity and adaptability as well as the ability to stay within the bounds of the designed protocol. Here, I describe the preparation you need to do for an interview study and the process of conducting the interview itself.

5.6.1 Preparation:

The Role of the Interviewer

The interviewer is really the "jack-of-all-trades" in survey research. The interviewer's role is complex and multifaceted. It includes the following tasks:

- **Locate and enlist cooperation of respondents**

The interviewer has to find the respondent. In door-to-door surveys, this means being able to locate specific addresses. Often,

the interviewer has to work at the least desirable times (like immediately after dinner or on weekends) because that's when respondents are most readily available.

- **Motivate respondents to do good job**

If the interviewer does not take the work seriously, why would the respondent? The interviewer has to be motivated and has to be able to communicate that motivation to the respondent. Often, this means that the interviewer has to be convinced of the importance of the research.

- **Clarify any confusion/concerns**

Interviewers have to be able to think on their feet. Respondents may raise objections or concerns that were not anticipated. The interviewer has to be able to respond candidly and informatively.

- **Observe quality of responses**

Whether the interview is personal or over the phone, the interviewer is in the best position to judge the quality of the information that is being received. Even a verbatim transcript will not adequately convey how seriously the respondent took the task, or any gestures or body language that were evident.

- **Conduct a good interview**

Last, and certainly not least, the interviewer has to conduct a good interview! Every interview has a life of its own. Some respondents are motivated and attentive, others are distracted or disinterested. The interviewer also has good or bad days. Assuring a consistently high-quality interview is a challenge that requires constant effort.

5.6.2 Training the Interviewers:

One of the most important aspects of any interview study is the training of the interviewers themselves. In many ways the interviewers are your measures, and the quality of the results is totally in their hands. Even in small studies involving only a single researcher-interviewer, it is important to organize in detail and rehearse the interviewing process before beginning the formal study.

Here are some of the major topics that should be included in interviewer training:

- **Describe the entire study**

Interviewers need to know more than simply how to conduct the interview itself. They should learn about the background for the study, previous work that has been done, and why the study is important.

- **State who is sponsor of research**

Interviewers need to know who they are working for. They -- and their respondents -- have a right to know not just what agency or company is conducting the research, but also, who is paying for the research.

- **Teach enough about survey research**

While you seldom have the time to teach a full course on survey research methods, the interviewers need to know enough that they respect the survey method and are motivated. Sometimes it may not be apparent why a question or set of questions was asked in a particular way. The interviewers will need to understand the rationale for how the instrument was constructed.

- **Explain the sampling logic and process**

Naive interviewers may not understand why sampling is so important. They may wonder why you go through all the difficulties of selecting the sample so carefully. You will have to explain that sampling is the basis for the conclusions that will be reached and for the degree to which your study will be useful.

- **Explain interviewer bias**

Interviewers need to know the many ways that they can inadvertently bias the results. And, they need to understand why it is important that they not bias the study. This is especially a problem when you are investigating political or moral issues on which people have strongly held convictions. While the interviewer may think they are doing good for society by slanting results in favor of what they believe, they need to recognize that doing so could jeopardize the entire study in the eyes of others.

- **"Walk through" the interview**

When you first introduce the interview, it's a good idea to walk through the entire protocol so the interviewers can get an idea of the various parts or phases and how they interrelate.

- **Explain respondent selection procedures, including**

- **reading maps**

It's astonishing how many adults don't know how to follow directions on a map. In personal interviews, the interviewer may need to locate respondents who are spread over a wide geographic area. And, they often have to navigate by night (respondents tend to be most available in evening hours) in neighborhoods they're not familiar with. Teaching basic map reading skills and confirming that the interviewers can follow maps is essential.

- **identifying households**

In many studies it is impossible in advance to say whether every sample household meets the sampling requirements for the study. In your study, you may want to interview only people who live in single family homes. It may be impossible to distinguish townhouses and apartment buildings in your sampling frame. The interviewer must know how to identify the appropriate target household.

- **identify respondents**

Just as with households, many studies require respondents who meet specific criteria. For instance, your study may require that you speak with a male head-of-household between the ages of 30 and 40 who has children under 18 living in the same household. It may be impossible to obtain statistics in advance to target such respondents. The interviewer may have to ask a series of filtering questions before determining whether the respondent meets the sampling needs.

- **Rehearse interview**

You should probably have several rehearsal sessions with the interviewer team. You might even videotape rehearsal interviews to discuss how the trainees responded in difficult situations. The interviewers should be very familiar with the entire interview before ever facing a respondent.

- **Explain supervision**

In most interview studies, the interviewers will work under the direction of a supervisor. In some contexts, the supervisor may be a faculty advisor; in others, they may be the "boss." In order to assure the quality of the responses, the supervisor may have to observe a subsample of interviews, listen in on phone interviews, or conduct follow-up assessments of interviews with the respondents. This can be very threatening to the interviewers. You need to develop an atmosphere where everyone on the research team -- interviewers and supervisors -- feel like they're working together towards a common end.

- **Explain scheduling**

The interviewers have to understand the demands being made on their schedules and why these are important to the study. In some studies it will be imperative to conduct the entire set of interviews within a certain time period. In most studies, it's important to have the interviewers available when it's convenient for the respondents, not necessarily the interviewer.

5.6.3 The Interviewer's Kit

It's important that interviewers have all of the materials they need to do a professional job. Usually, you will want to assemble an interviewer kit that can be easily carried and includes all of the important materials such as:

- a "professional-looking" 3-ring notebook (this might even have the logo of the company or organization conducting the interviews)
- maps
- sufficient copies of the survey instrument
- official identification (preferable a picture ID)
- a cover letter from the Principal Investigator or Sponsor
- a phone number the respondent can call to verify the interviewer's authenticity

5.6.4 The Interview :

So all the preparation is complete, the training done, the interviewers ready to proceed, their "kits" in hand. It's finally time to do an actual interview. Each interview is unique, like a small work of art (and sometimes the art may not be very good). Each interview has its own ebb and flow -- its own pace. To the outsider, an interview looks like a fairly standard, simple, prosaic effort. But to the interviewer, it can be filled with special nuances and interpretations that aren't often immediately apparent. Every interview includes some common components. There's the opening, where the interviewer gains entry and establishes the rapport and tone for what follows. There's the middle game, the heart of the process, that consists of the protocol of questions and the improvisations of the probe. And finally, there's the endgame, the wrap-up, where the interviewer and respondent establish a sense of closure. Whether it's a two-minute phone interview or a personal interview that spans hours, the interview is a bit of theater, a mini-drama that involves real lives in real time.

5.6.5 Opening Remarks

In many ways, the interviewer has the same initial problem that a salesperson has. You have to get the respondent's attention initially for a long enough period that you can sell them on the idea of participating in the study. Many of the remarks here assume an interview that is being conducted at a respondent's residence. But the analogies to other interview contexts should be straightforward.

- **Gaining entry**

The first thing the interviewer must do is gain entry. Several factors can enhance the prospects. Probably the most important factor is your initial appearance. The interviewer needs to dress professionally and in a manner that will be comfortable to the respondent. In some contexts a business suit and briefcase may be appropriate. In others, it may intimidate. The way the interviewer appears initially to the respondent has to communicate some simple messages -- that you're trustworthy, honest, and non-threatening. Cultivating a manner of professional confidence, the sense that the respondent has nothing to worry about because you know what you're doing -- is a difficult skill to teach and an indispensable skill for achieving initial entry.

- **Doorstep technique**

You're standing on the doorstep and someone has opened the door, even if only halfway. You need to smile. You need to be brief. State why you are there and suggest what you would like the respondent to do. Don't ask -- suggest what you want. Instead of saying "May I come in to do an interview?", you might try a more imperative approach like " I'd like to take a few minutes of your time to interview you for a very important study."

- **Introduction**

If you've gotten this far without having the door slammed in your face, chances are you will be able to get an interview. Without waiting for the respondent to ask questions, you should move to introducing yourself. You should have this part of the process memorized so you can deliver the essential information in 20-30 seconds at most. State your name and the name of the organization you represent. Show your identification badge and the letter that introduces you. You want to have as legitimate an appearance as possible. If you have a three-ring binder or clipboard with the logo of your organization, you should have it out and visible. You should assume that the respondent will be interested in participating in your important study -- assume that you will be doing an interview here.

- **Explaining the study**

At this point, you've been invited to come in (After all, you're standing there in the cold, holding an assortment of materials, clearly displaying your credentials, and offering the respondent the chance to participate in an interview -- to many respondents, it's a rare and exciting event. They hardly ever get asked their views about anything, and yet they know that important decisions are made all the time based on input from others.). Or, the respondent has continued to listen long enough that you need to move onto explaining the study. There are three rules to this critical explanation: 1) Keep it short; 2) Keep it short; and 3) Keep it short! The respondent doesn't have to or want to know all of the neat nuances of this study, how it came about, how you convinced your thesis committee to buy into it, and so on. You should have a one or two sentence description of the study memorized. No big words. No jargon. No detail. There will be more than enough time for that later (and you should bring some written materials you can leave at the end for that purpose). This is the "25 words or less" description. What you ***should*** spend some time on is assuring the respondent that you are interviewing them confidentially, and that their participation is voluntary.

5.6.6 Asking the Questions

You've gotten in. The respondent has asked you to sit down and make yourself comfortable. It may be that the respondent was in the middle of doing something when you arrived and you may need to allow them a few minutes to finish the phone call or send the kids off to do homework. Now, you're ready to begin the interview itself.

- **Use questionnaire carefully, but informally**

The questionnaire is your friend. It was developed with a lot of care and thoughtfulness. While you have to be ready to adapt to the needs of the setting, your first instinct should always be to trust the instrument that was designed. But you also need to establish a rapport with the respondent. If you have your face in the instrument and you read the questions, you'll appear unprofessional and disinterested. Even though you may be nervous, you need to recognize that your respondent is most likely even more nervous. If you memorize the first few questions, you can refer to the instrument only occasionally, using eye contact and a confident manner to set the tone for the interview and help the respondent get comfortable.

- **Ask questions exactly as written**

Sometimes an interviewer will think that they could improve on the tone of a question by altering a few words to make it simpler or more "friendly." DON'T. You should ask the questions as they are on the instrument. If you had a problem with a question, the time to raise it was during the training and rehearsals, not during the actual interview. It is important that the interview be as standardized as possible across respondents (this is true except in certain types of exploratory or interpretivist research where the explicit goal is to avoid any standardizing). You may think the change you made was inconsequential when, in fact, it may change the entire meaning of the question or response.

- **Follow the order given**

Once you know an interview well, you may see a respondent bring up a topic that you know will come up later in the interview. You may be tempted to jump to that section of the interview while you're on the topic. DON'T. You are more likely to lose your place. You may omit questions that build a foundation for later questions.

- **Ask every question**

Sometimes you'll be tempted to omit a question because you thought you already heard what the respondent will say. Don't assume that. For example, let's say you were conducting an interview with college age women about the topic of date rape. In an earlier question, the respondent mentioned that she knew of a woman on her dormitory floor who had been raped on a date within the past year. A few questions later, you are supposed to ask "Do you know of anyone personally who was raped on a date?" You figure you already know that the answer is yes, so you decide to skip the question. Instead, you might say something like "I know you may have already mentioned this, but do you know of anyone personally who was raped on a date?" At this point, the respondent may say something like "Well, in addition to the woman who lived down the hall in my dorm, I know of a friend from high school who experienced date rape." If you hadn't asked the question, you would never have discovered this detail.

- **Don't finish sentences**

I don't know about you, but I'm one of those people who just hates to be left hanging. I like to keep a conversation moving. Once I know where a sentence seems to be heading, I'm aching to get to the next sentence. I finish people's sentences all the time. If you're like me, you should practice the art of patience (and silence) before doing any interviewing. As you'll see below, silence is one of the most effective devices for encouraging a respondent to talk. If you finish their sentence for them, you imply that what they had to say

is transparent or obvious, or that you don't want to give them the time to express themselves in their own language.

5.6.7 Obtaining Adequate Responses - The Probe

OK, you've asked a question. The respondent gives a brief, cursory answer. How do you elicit a more thoughtful, thorough response? You *probe*.

- **Silent probe**

The most effective way to encourage someone to elaborate is to do nothing at all - just pause and wait. This is referred to as the "silent" probe. It works (at least in certain cultures) because the respondent is uncomfortable with pauses or silence. It suggests to the respondent that you are waiting, listening for what they will say next.

- **Overt encouragement**

At times, you can encourage the respondent directly. Try to do so in a way that does not imply approval or disapproval of what they said (that could bias their subsequent results). Overt encouragement could be as simple as saying "Uh-huh" or "OK" after the respondent completes a thought.

- **Elaboration**

You can encourage more information by asking for elaboration. For instance, it is appropriate to ask questions like "Would you like to elaborate on that?" or "Is there anything else you would like to add?"

- **Ask for clarification**

Sometimes, you can elicit greater detail by asking the respondent to clarify something that was said earlier. You might say, "A minute ago you were talking about the experience you had in high school. Could you tell me more about that?"

- **Repetition**

This is the old psychotherapist trick. You say something without really saying anything new. For instance, the respondent just described a traumatic experience they had in childhood. You might say "What I'm hearing you say is that you found that experience very traumatic." Then, you should pause. The respondent is likely to say something like "Well, yes, and it affected the rest of my family as well. In fact, my younger sister..."

5.6.8 Recording the Response:

Although we have the capability to record a respondent in audio and/or video, most interview methodologists don't think it's a good idea. Respondents are often uncomfortable when they know their remarks will be recorded word-for-word. They may strain to only say things in a socially acceptable way. Although you would get a more detailed and accurate record, it is likely to be distorted by the very process of obtaining it. This may be more of a problem in some situations than in others. It is increasingly common to be told that your conversation may be recorded during a phone interview. And most focus group methodologies use unobtrusive recording equipment to capture what's being said. But, in general, personal interviews are still best when recorded by the interviewer using pen and paper. Here, I assume the paper-and-pencil approach.

- **Record responses immediately**

The interviewer should record responses as they are being stated. This conveys the idea that you are interested enough in what the respondent is saying to write it down. You don't have to write down every single word -- you're not taking stenography. But you may want to record certain key phrases or quotes verbatim. You need to develop a system for distinguishing what the respondent says verbatim from what you are characterizing (how about quotations, for instance!).

- **Include all probes**

You need to indicate every single probe that you use. Develop a shorthand for different standard probes. Use a clear form for writing them in (e.g., place probes in the left margin).

- **Use abbreviations where possible**

Abbreviations will help you to capture more of the discussion. Develop a standardized system (e.g., R=respondent; DK=don't know). If you create an abbreviation on the fly, have a way of indicating its origin. For instance, if you decide to abbreviate Spouse with a 'S', you might make a notation in the right margin saying "S=Spouse."

5.6.9 Concluding the Interview:

When you've gone through the entire interview, you need to bring the interview to closure. Some important things to remember:

- **Thank the respondent**

Don't forget to do this. Even if the respondent was troublesome or uninformative, it is important for you to be polite and thank them for their time.

- **Tell them when you expect to send results**

I hate it when people conduct interviews and then don't send results and summaries to the people who they get the information from. You owe it to your respondent to show them what you learned. Now, they may not want your entire 300-page dissertation. It's common practice to prepare a short, readable, jargon-free summary of interviews that you can send to the respondents.

- **Don't be brusque or hasty**

Allow for a few minutes of winding down conversation. The respondent may want to know a little bit about you or how much you like doing this kind of work. They may be interested in how the results will be used. Use these kinds of interests as a way to wrap up the conversation. As you're putting away your materials and packing up to go, engage the respondent. You don't want the respondent to feel as though you completed the interview and then rushed out on them -- they may wonder what they said that was wrong. On the other hand, you have to be careful here. Some respondents may want to keep on talking long after the interview is over. You have to find a way to politely cut off the conversation and make your exit.

- **Immediately after leaving -- write down any notes about how the interview went**

Sometimes you will have observations about the interview that you didn't want to write down while you were with the respondent. You may have noticed them get upset at a question, or you may have detected hostility in a response. Immediately after the interview you should go over your notes and make any other comments and observations -- but be sure to distinguish these from the notes made during the interview (you might use a different color pen, for instance).

5.7 SUMMARY

1. Primary sources are original sources from which the researcher directly collects data that have not been previously collected. Primary data are first hand information collected through various methods such as observation, interviewing, mailing etc.

2. Surveys can be divided into two broad categories: the

questionnaires and the interviews.

3. A questionnaire is either a mail questionnaire or group administered questionnaire. A less familiar type of questionnaire is the household drop-off survey.

4. Mail questionnaire can be sent to a wide number of people. It is inexpensive to administer. But it is not the best tool for asking detailed written responses.

5. In group administered questionnaire a sample of respondents is brought together and asked to respond to a structured sequence of questions.

6. In household drop-off survey , a researcher goes to the respondent's home or business and hand over the respondent instrument. In some cases, the respondent is asked to mail it back or the interviewer returns to pick it up. This approach attempts to blend the advantages of the mail survey and the group administered questionnaire.

7. Interview may be of personal interview or telephone interview. In the personal interview, the interviewer works directly with the respondents. It is time consuming. Telephone interviews enable a researcher to gather information rapidly.

8. Interviews are among the most challenging and rewarding forms of measurement. They require a personal sensitivity and adaptability as well as the ability to stay within the bounds of the designed protocol.

5.8 QUESTIONS

1. What are the primary sources of data?
2. Explain the Advantages and disadvantages of primary data
3. write short note on:
   a) Survey technique
   b) Interview technique
   c) Questionnaire
4. Write the method of collecting Primary Data. Write the detail of any one technique with example.

6

SECONDARY DATA COLLECTION

Unit Structure:

6.0 OBJECTIVES

- To know the meaning of secondary data
- To study The internal and external sources of secondary data
- To study the features and Uses of secondary data
- To study the advantages and disadvantages of secondary data
- To study the evaluation of secondary data

6.1 SECONDARY SOURCES OF DATA

These are sources containing data which have been collected and compiled for another purpose. The secondary sources consists of readily compendia and already compiled statistical statements and reports whose data may be used by researchers for their studies e.g., census reports , annual reports and financial statements of companies, Statistical statement, Reports of Government Departments, Annual reports of currency and finance published by the Reserve Bank of India, Statistical statements relating to Co-operatives and Regional Banks, published by the NABARD, Reports of the National sample survey Organization, Reports of trade associations, publications of international organizations such as UNO, IMF, World Bank, ILO, WHO, etc., Trade and Financial journals newspapers etc. Secondary sources consist of not only published records and reports, but also unpublished records. The latter category includes various records and registers maintained by the firms and

organizations, e.g., accounting and financial records, personnel records, register of members, minutes of meetings, inventory records etc. Secondary data is information gathered for purposes other than the completion of a research project. A variety of secondary information sources is available to the researcher gathering data on an industry, potential product applications and the market place. Secondary data is also used to gain initial insight into the research problem.

Secondary data is classified in terms of its source – either internal or external. Internal, or in-house data, is secondary information acquired within the organization where research is being carried out. External secondary data is obtained from outside sources. The two major advantages of using secondary data in market research are time and cost savings.

- The secondary research process can be completed rapidly – generally in 2 to 3 week. Substantial useful secondary data can be collected in a matter of days by a skillful analyst.
- When secondary data is available, the researcher need only locate the source of the data and extract the required information.
- Secondary research is generally less expensive than primary research. The bulk of secondary research data gathering does not require the use of expensive, specialized, highly trained personnel.
- Secondary research expenses are incurred by the originator of the information.

-There are also a number of disadvantages of using secondary data. These include:

- Secondary information pertinent to the research topic is either not available, or is only available in insufficient quantities.
- Some secondary data may be of questionable accuracy and reliability. Even government publications and trade magazines statistics can be misleading. For example, many trade magazines survey their members to derive estimates of market size, market growth rate and purchasing patterns, then average out these results. Often these statistics are merely average opinions based on less than 10% of their members.
- Data may be in a different format or units than is required by the researcher.
- Much secondary data is several years old and may not reflect the current market conditions. Trade journals and other publications often accept articles six months before appear in print. The research may have been done months or even years earlier.

As a general rule, a thorough research of the secondary data should be undertaken prior to conducting primary research. The secondary information will provide a useful background and will identify key questions and issues that will need to be addressed by the primary research.

6.1.1 Internal data sources:

Internal secondary data is usually an inexpensive information source for the company conducting research, and is the place to start for existing operations. Internally generated sales and pricing data can be used as a research source. The use of this data is to define the competitive position of the firm, an evaluation of a marketing strategy the firm has used in the past, or gaining a better understanding of the company's best customers.

There are three main sources of internal data. These are:

1. Sales and marketing reports:

These can include such things as:

- Type of product/service purchased
- Type of end-user/industry segment
- Method of payment
- Product or product line
- Sales territory
- Salesperson
- Date of purchase
- Amount of purchase
- Price
- Application by product
- Location of end-user

2. Accounting and financial records:

These are often an overlooked source of internal secondary information and can be invaluable in the identification, clarification and prediction of certain problems. Accounting records can be used to evaluate the success of various marketing strategies such as revenues from a direct marketing campaign.

There are several problems in using accounting and financial data. One is the timeliness factor – it is often several months before accounting statements are available. Another is the structure of the records themselves. Most firms do not adequately setup their accounts to provide the types of answers to research questions that they need. For example, the account systems should capture

project/product costs in order to identify the company's most profitable (and least profitable) activities.

Companies should also consider establishing performance indicators based on financial data. These can be industry standards or unique ones designed to measure key performance factors that will enable the firm to monitor its performance over a period of time and compare it to its competitors. Some example may be sales per employee, sales per square foot, expenses per employee (salesperson, etc.).

3. Miscellaneous reports:

These can include such things as inventory reports, service calls, number (qualifications and compensation) of staff, production and R&D reports. Also the company's business plan and customer calls (complaints) log can be useful sources of information.

6.1.2 External data sources:

There is a variety of statistical and research data available today. Some sources are:

- Federal government
- Provincial/state governments
- Statistics agencies
- Trade associations
- General business publications
- Magazine and newspaper articles
- Annual reports
- Academic publications
- Library sources
- Computerized bibliographies
- Syndicated services.

A good place to start your search is the local city, college or university library. Most reference librarians are very knowledgeable about what data is available, or where to look to find it. Also contact government libraries and departments for research reports/publications they may have done.

This includes all types of information sources that you may use, including:

- Books
- Articles (from print sources or from online article databases)

- Interviews
- E-mail or any other correspondence
- Web pages
- Government documents
- Non-print media (videotapes, audiotapes, pictures and images)
- Software or any digital formats

Check Your Progress:

1. What do you understand by secondary data?
2. State the main sources of secondary data.
3. Distinguish between Internal and External secondary sources of data.

_____

_____

_____

_____

_____

6.2 FEATURES AND USES OF SECONDARY DATA

6.2.1 Features:

Though secondary sources are diverse and consist of all sorts of materials, they have certain common characteristics. First, they are readymade and readily available, and do not require the trouble of constructing tools and administering them.

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

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

6.2.2 Uses of Secondary data:

The secondary data may be used in three ways by a researcher. First, some specific information from secondary

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

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

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

6.3 ADVANTAGES AND DISADVANTAGES OF SECONDARY DATA

6.3.1 Secondary sources have some advantages:

1. Secondary data, if available can be secured quickly and cheaply. Once their source of documents and reports are located, collection of data is just matter of desk work. Even the tediousness of copying the data from the source can now be avoided, thanks to Xeroxing facilities.

2. Wider geographical area and longer reference period may be covered without much cost. Thus, the use of secondary data extends the researcher's space and time reach.

3. The use of secondary data broadens the data base from which scientific generalizations can be made.

4. Environmental and cultural settings are required for the study.

5. The use of secondary data enables a researcher to verify the findings bases on primary data. It readily meets the need for additional empirical support. The researcher need not wait the time when additional primary data can be collected.

6.3.2 Disadvantages of secondary data:

The use of a secondary data has its own limitations.

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

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

3. The secondary data are not up-to-date and become obsolete when they appear in print, because of time lag in producing them. For example, population census data are published two or three years later after compilation, and no new figures will be available for another ten years.

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

Check Your Progress:

1. Explain how secondary data is useful to the researchers.
2. What are the limitations of secondary sources of data?

_____

_____

_____

_____

6.4 EVALUATION OF SECONDARY DATA

When a researcher wants to use secondary data for his research, he should evaluate them before deciding to use them.

1. **Data Pertinence:**

The first consideration in evaluation is to examine the pertinence of the available secondary data to the research problem under study. The following questions should be considered. What are the definitions and classifications employed? Are they

consistent ? What are the measurements of variables used? What is the degree to which they conform to the requirements of our research? What is the coverage of the secondary data in terms of topic and time? Does this coverage fit the needs of our research?

On the basis of above consideration, the pertinence of the secondary data to the research on hand should be determined, as a researcher who is imaginative and flexible may be able to redefine his research problem so as to make use of otherwise unusable available data.

2. **Data Quality:**

If the researcher is convinced about the available secondary data for his needs, the next step is to examine the quality of the data. The quality of data refers to their accuracy, reliability and completeness. The assurance and reliability of the available secondary data depends on the organization which collected them and the purpose for which they were collected. What is the authority and prestige of the organization? Is it well recognized? Is it noted for reliability? It is capable of collecting reliable data? Does it use trained and well qualified investigators? The answers to these questions determine the degree of confidence we can have in the data and their accuracy. It is important to go to the original source of the secondary data rather than to use an immediate source which has quoted from the original. Then only, there searcher can review the cautionary ands other comments that were made in the original source.

3. **Data Completeness:**

The completeness refers to the actual coverage of the published data. This depends on the methodology and sampling design adopted by the original organization. Is the methodology sound? Is the sample size small or large?

Is the sampling method appropriate? Answers to these questions may indicate the appropriateness and adequacy of the data for the problem under study. The question of possible bias should also be examined.

Whether the purpose for which the original organization collected the data had a particular orientation? Has the study been made to promote the organization's own interest? How the study was conducted? These are important clues. The researcher must be on guard when the source does not report the methodology and sampling design. Then it is not possible to determine the adequacy of the secondary data for the researcher's study.

6.5 SUMMARY

1. The secondary sources consists of readily compendia and already compiled statistical statements and reports whose data may be used by researchers for their studies. For e.g. census reports, annual reports etc.
2. Internal or In-house data, is secondary information acquired within the organization where research is being carried out.
3. External secondary data is obtained from outside sources.
4. The two major advantages of using secondary data in market research are time and cost savings.
5. Though secondary sources are diverse and consist of all sorts of materials, they have certain common characteristics.
6. Secondary data may be used in three ways by a researchers. Some specific information from secondary sources may be used for reference purposes. It may be used as bench marks against which the findings of research may be tested. It may be used as the sole source of information for a research project.
7. Secondary data have some advantages and disadvantages.
8. When a researcher wants to use secondary data for his research, he should evaluate them before deciding to use them on the basis of Data pertinence, Data quality and Data completeness.

6.6 QUESTIONS

4. Define and explain the secondary sources of data.
5. State and explain the secondary sources of data.
6. Discuss the characteristics of secondary data.
7. Describe the advantages and disadvantages of secondary data.
8. Discuss the evaluation of secondary data.

□ □ □ □

7

Module 4
PRESENTATION AND PRELIMINARY ANALYSIS OF DATA

Unit Structure:

7.0 OBJECTIVES

- To learn the process of classifying the data.
- To systematically arrange the data into statistical tables.
- To present the statistical results through graphs and diagrams.

7.1 INTRODUCTION

In the last few units, the process of data collection was discussed. The data collected by applying any of the methods of primary data collection, need to be classified and tabulated, to make them more meaningful. Too many figures and statistical tables may be confusing for a common man. So, graphical presentation of the statistical results, may be necessary. In the present unit, we will learn various techniques to classify, tabulate and diagrammatically represent the statistical results.

7.2 CLASSIFICATION

7.2.1 Meaning and types of classification:

Classification means grouping the data according to similar characteristics and putting them into various classes. The raw data is sorted according to some basis so that the data can be arranged

systematically. The necessary features of data may be highlighted and the unnecessary details may be removed through the process of classification.

Data may be classified according to following four categories.

7.2.1.1 Geographical classification:

Under this, data is grouped according to geographical location such as a country, state, city or area. The data related to state-wise wheat production or nation-wise per capita income or area wise rainfall are some of the examples of geographical classification.

Example 1

Country wise Manufacturing output (in Million tones)

Country	Output (in MT)
China	325
Japan	550
India	350
USSR	400
Kenya	75

Example 2

Area wise sales of product x (%)

Area	Percentage
North	15
South	25
Central	8
West	20
East	32

7.2.1.2 Chronological classification:

When the data are classified according to some element of time-say year, month or day. It is called Chronological classification. For example wheat production over last 10 years, sales per day. Profit per week etc.

Example 1

Day	Sales per day
Monday	203
Tuesday	200
Wednesday	150
Thursday	270
Friday	300
Saturday	325

7.2.1.3 Qualitative classification

In this type of classification, data are classified on the basis of some characteristics which are not measurable. Such as religion, language, gender, colour of hair, etc. Quantitative classification is made on the basis of whether a particular characteristic is present or not. No quantitative measurement is possible. For example, classifying a population into illiterate and literate, male and female, vegetarian and non-vegetarian, etc.

7.2.1.4 Quantitative classification

In this type of classification, data are classified on the basis of some characteristics which are quantifiable or measurable. For example height in centimeters, weights in kilograms, wages in Rs., etc. In statistics, such a classification is known as frequency distribution. Such a distribution may be discrete or continuous, on the basis of the nature of variable under study.

Discrete Frequency Distribution

No. of children	No. of families
0	10
1	35
2	20
3	8
4	1
Total	**74**

Continuous Frequency Distribution

Heights in inches	No. of Items
3 – 5	25
5 – 7	38
7 – 9	40
9 – 11	20
11 – 13	17
Total	**140**

7.2.2 Frequency Distribution:

The data collected by an investigator requires to be classified according to the values of variable under study. This is called a frequency distribution table. An unarranged data is a raw data. Such a data may not be useful for any statistical work. An arrangement of the data in ascending or descending order with the help of tally bars, makes an understanding of data in a better way.

Following are important steps in preparing the frequency distribution tables:

1) Define the classes according to which grouping is done (either discrete or continuous)
2) The number of classes should not be too less and too large. (It may be between 6 to 10)
3) Normally, the classes should be of equal length.
4) The column of tally bars should be prepared to facilitate counting of values of variable. (Have the blocks of five bars are prepared and then they are added)
5) By counting the bars, column of frequencies should be made.

7.2.2.1 Discrete frequency distribution:

Exercise 1

From the following data about 30 families in a society regarding the number of children per family, prepare a frequency distribution table.

2	3	2	2	1	1	0	0	0	1
2	3	2	1	1	1	0	1	0	2
4	1	2	2	1	0	1	2	1	2

No. of children per family

No. of children	Tally bars	No. of families
0	𝍸 I	6
1	𝍸 𝍸 I	11
2	𝍸 𝍸	10
3	II	2
4	I	1
	Total	**30**

7.2.2.2 Continuous Frequency Distribution

Some important concepts related to continuous frequency distribution are as follows:

a) Class limits: There are two class limits, upper and lower. In a class, say, 0-10, '0' is a lower and 10 is an upper class limit.

b) Class interval: A difference between upper and lower class limit is called class interval. For example in a class 0-10, Class interval is 10 (10-0) or in a class 100-200, class interval is 100 (200-100)

c) Class frequency: The number of observations corresponding to a particular class is known as class frequency. For example, if number of students getting marks between 10 to 20 are 7, then the frequency of class 10-20 is 7.

d) Mid point of class: It is found out with the help of following formula.

$$\text{Mid point of a class} = \frac{\text{Upper class limit} + \text{lower class limit}}{2}$$

For example, the mid point of a class 4-8 is $6\left(\frac{8+4}{2}\right)$.

e) Inclusive Method of classification: Under this method, both the lower and upper class-limits are included in the same class.

For example:

Weights (in kgs)	No. of students
25 – 29	10
30 – 34	12
35 – 39	15
40 – 44	7
45 – 49	3

In a class 30-34, we include all those students whose weight is 30,31,32,33 or 34 kgs. Since both the upper and lower class limits are included in the class, it is called 'inclusive' type of classification.

f) 'Exclusive' classification: Under this method, the upper limit of each class is excluded from the class. Following example will explain this.

Weights (in kgs)	No. of students
25 – 30	9
30 – 35	17
35 – 40	14
40 – 45	10
45 – 50	5

In a class 25-30, we include all those students whose weight is 25, 26, 27, 28 or 29 kgs. A student with weight 30 kg. is included in the next class. This type of classification is useful for those variables which can take fractional values.

g) Conversion of 'inclusive' classes into 'exclusive' classes. In order to ensure a continuity in the representation of data, it is necessary to convert the 'inclusive' classes into 'exclusive' ones. For this purpose, a correction factor is computed as follows.

$$\text{Correction Factor} = \frac{\text{Lower limit of } 2^{nd} \text{ class} - \text{Upper limit of the first class}}{2}$$

Following example will explain the procedure of conversion.

'inclusive' classes	Frequency	'Exclusive' classes
10 – 19	5	9.5 – 19.5
20 – 29	10	19.5 – 29.5
30 – 39	12	29.5 – 39.5
40 – 49	18	39.5 – 49.5

50 - 59	5	49.5 – 59.5

Adjustment is done as follows:

To adjust the class intervals, we take a difference between 20 (lower class limit of 2nd class) and 19 (upper class limit of first class), which is one . Dividing it by two we get ½ or 0.5. So 0.5 is a correction factor. Using this factor an adjustment is made which is shown in column 3 of the table.

After understanding various concepts related to frequency distribution, we will prepare frequency distribution for the continuous data.

Exercise 2

Prepare a frequency distribution table for the following data. Take class interval as 10 and we exclusive method.

57	44	80	75	00	18	45	14	04	64
72	51	69	34	22	83	70	20	57	28
96	56	50	47	10	34	61	66	80	46
22	10	84	50	47	73	42	33	48	65
10	34	66	53	75	90	58	46	39	69

The lowest value in the data is zero and the highest value is 96. We have to take class interval as 10 and use exclusive method so our classes will be 0 – 10, 10 – 20 ……..

Frequency Distribution

Marks	Tally Bars	Frequency
0 – 10	II	2
10 – 20	𝍸	5
20 – 30	IIII	4
30 – 40	𝍸	5
40 – 50	𝍸 III	8
50 – 60	𝍸 III	8
60 – 70	𝍸 II	7
70 – 80	𝍸	5
80 – 90	IIII	4
90 - 100	II	2
	$\sum f =$	**50**

Exercise 3

Marks obtained by 50 boys of a class are as under:

34	54	10	21	51	52	12	43	48	36
48	22	39	26	34	19	10	17	47	38
13	30	30	60	59	15	07	18	40	49
40	51	55	32	41	22	30	35	53	23
14	18	19	40	43	04	17	45	25	43

Construct frequency distribution table taking the class intervals as 0-9, 10-19, 20-29 etc.

Marks	Tally Bars	Frequency
0 – 9	II	2
10 – 19	𝍸 𝍸 II	12
20 – 29	𝍸 I	6
30 – 39	𝍸 𝍸	10
40 – 49	𝍸 𝍸 II	12
50 – 59	𝍸 II	07
60 – 69	I	01
Total		**50**

Check your progress

1. Construct frequency distribution table for the following data of number of letters in English words. (Discrete type)

3	4	2	3	5	7	6	5	4	3
4	2	4	3	5	6	7	2	3	5
5	3	8	4	2	3	7	4	2	3
6	4	7	8	6	2	6	5	3	7
7	5	4	3	5	6	5	8	2	3

2. Age at death of 50 persons in a town is given below. Arrange the data in frequency distribution by taking exclusive type of classes with class interval to be 10.

36	48	50	45	49	31	50	48	43	42
37	32	40	39	41	47	45	39	43	47
38	39	37	40	32	52	56	31	54	36
51	46	41	55	58	31	42	53	32	44
53	36	60	59	41	53	58	36	38	60

3. The weights of 50 persons are given below. Arrange the data by taking 5 as class interval. Use inclusive type of classes.

76	64	53	55	66	72	52	63	46	51
53	56	65	60	47	55	67	73	44	54
64	74	48	59	72	61	43	69	68	58
42	52	62	72	43	63	71	64	58	67
46	55	65	75	48	59	67	77	64	78

7.3 TABULATION

A table is a systematic arrangement of a data into rows and columns and a process of summarizing the data and putting it in a form of a table is called tabulation. A table is useful in presenting a huge data with a minimum space. Table saves a time of a reader as it presents a huge complex data in a simplistic way. Two tables can be easily compared. A table should be prepared and presented in a systematic manner to have a clear understanding about the data. So there are some features of a good table. Which should be kept in mind while tabulating the data.

7.3.1 Characteristics of a good table:

1) Size of a table should be in accordance with the size of the paper.
2) Table should be self explanatory and should have column headings (caption) and row-headings (stub).
3) Alphabetical, Chronological or Geographical order should be followed while arranging the items in a table.
4) Table should have a number and a title.
5) There should be a provision of source of table and foot notes for the table.
6) Explanation of signs, rounding and abbreviation of figures, etc. should be given in the footnotes.

7.3.2 Parts of a table:

Though the parts of a table will vary from case to case, following are main parts which each of the statistical tables must have.

1) Table number.
2) Title of the table.
3) Captions columns headings.

4) Stubs the row headings.
5) Body of a table where the data will be entered according to rows and columns.
6) Footnotes will be placed at the bottom of the table to explain any concept or information used in a table.
7) Source- It will be the name of a source form where the data is given.

Exercise 4

Prepare a blank table to represent the results of boys and girls students in a college. The results are represented by first class, second class, pass class and failed. Also show the totals.

Table 7.1
Gender wise results of students in college ABC

Students / Results	Boys	Girls	Total
First class Second class Pass class Failed			
Total			

Source: College ABC Magazine

Exercise 5

Prepare a table to represent the number of employees in the Bank of India, according to their age (Below 25, 25-35, 35-45, 45 and above), Sex (males and females) and Ranks (officers, Assistants and clerks) Also show totals.

Number of employees of Bank of India according to Age, Sex and Ranks

Age in years	Rank											
	Officers			Supervisors			Clerks			Total		
	M	F	T	M	F	T	M	F	T	M	F	T
Below 25												
25 – 35												
35 – 45												
45 & above												

Total												

Foot Note: M Stands for male
F Stands for females
T Stands for total

Source: Bank of India Annual Report

Check your progress

1. Prepare a blank table to represent the staff in a company showing following characteristics.
   i) Sex: Male and Females
   ii) Salary Grades: below 10,000, 10,000-15,000, 15,000 & above
   iii) Period – 2007 and 2010

2. Prepare a blank table to represent the coffee habits (coffee-drinkers, non-coffee drinkers) of males and females in two towns (town x and town y). Also show totals.

3. Prepare a blank table showing employment pattern of workers in a factory on the basis of their gender (males, females), marital status (married and unmarried) and residence (textile area and non-textile area)

7.4 GRAPHICAL REPRESENTATION OF DATA

The statistical data can also be represented in the form of graphs and diagrams. Graphical representation of data is an important step involved in statistics. Graphs and diagrams give a visual effect for the readers. These are useful for the following reasons.

7.4.1 Usefulness of Graphs and diagrams

1) Graphs and diagrams are simple and attractive way to represent complicated statistical data.
2) An information presented in diagrams can easily be understood.
3) Diagrams have memorizing effect. That means, once a diagram is seen, the trend represented through it has long run impact on the reader.
4) Through graphs and diagrams, data can be easily compared.

In this unit, we are going to learn following types of graph.

1) Histogram 2) Frequency Polygon 3) Frequency curve

2) 4) Ogives

7.4.2 Histogram

Histogram is the widely used method of graphical representation of frequency distribution. It consists of rectangular bars, the width of which represents classes and the height represents the frequency of the class. Histograms can be drawn for the continuous frequency distribution. We can locate mode with the help of histogram.

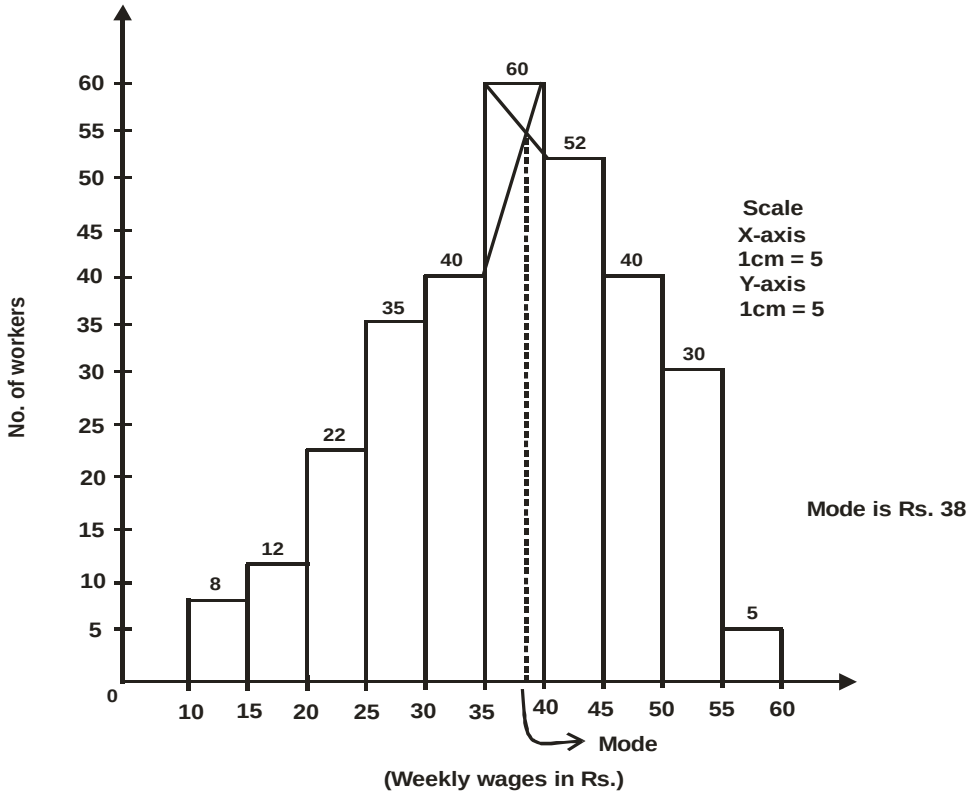
7.4.2.1 Histograms with equal class intervals:

Following is a method to draw histograms when the class intervals are equal. (That means width of all the classes in the given data is same)

Exercise 6

For the following data about the weekly wages of workers in a factory, draw histogram. Also locate mode graphically.

Weekly wages (Rs.)	No. of workers	
10 – 15	8	
15 – 20	12	These are the classes with equal class interval of 5.
20 – 25	22	
25 – 30	35	
30 – 35	40	
35 – 40	60	
40 – 45	52	
45 – 50	40	
50 – 55	30	
55 - 60	5	

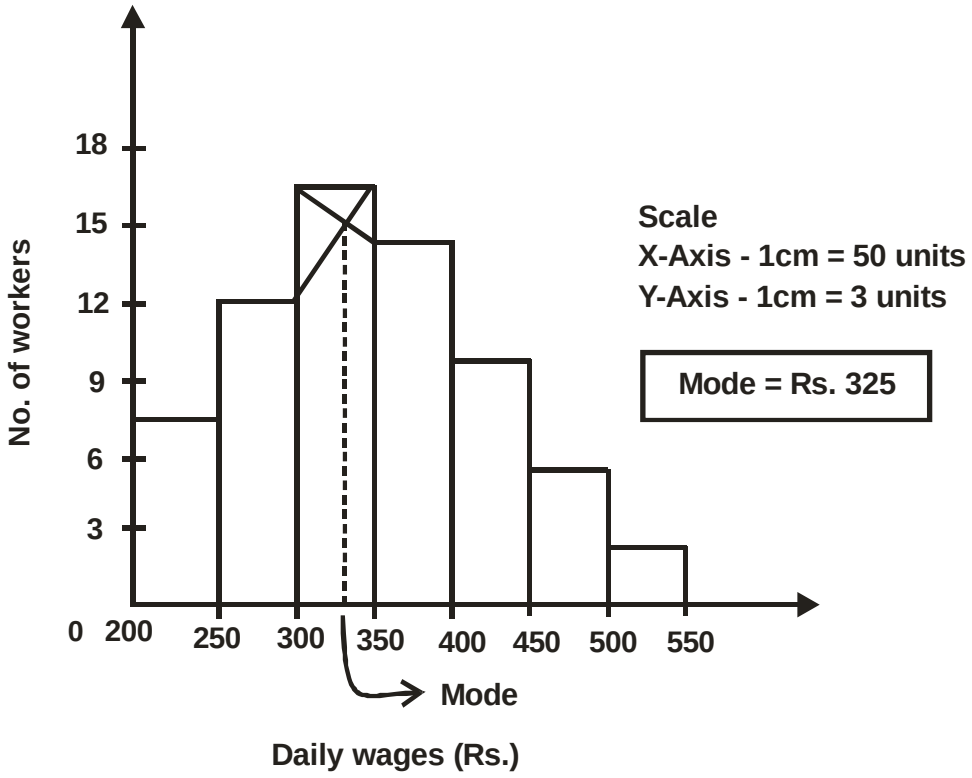


Exercise 7

For the following data, draw a histogram and locate mode graphically.

Daily wages (Rs.)	No. of workers
200 – 250	7
250 – 300	12
300 – 350	16
350 – 400	13
400 – 450	10
450 – 500	4
500 - 550	2

These are the classes with equal class interval of 50



7.4.2.2 Histograms with unequal class intervals:

If the class intervals are not uniform, some adjustment needs to be done in the frequencies before a histogram is drawn. Following steps are to be taken:

1) Take the class with the lowest frequency as 'normal' and adjust the other frequencies in relation to the normal class.
2) If the class interval is two times more than the normal, divide the frequency of that class by two.
3) If the class interval of the class is three times more than the normal, divide the frequency by three.

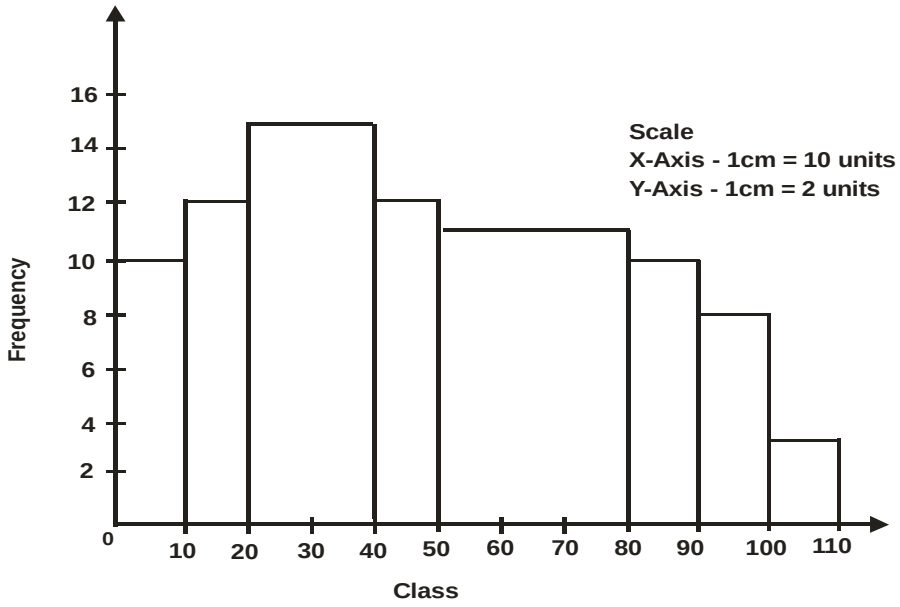
Exercise 8

Draw a histogram for the following set of data.

Class	Frequency
0 – 10	10
10 – 20	12
20 – 40	30/2 = 15
40 – 50	12
50 – 80	33/3 = 11
80 – 90	10
90 – 100	8
100 - 110	3

The class intervals of all the classes are not equal. For example class 0-10 has ten as class-interval. But class 20-40 has twenty as class interval.

Since all classes do not have equal class intervals. An adjustment needs to be done. Class 20-40 has 20 as class interval, so divide its frequencies by 2. Similarly, class 50-80 has 30 as class interval so divide its frequencies by 3.



7.4.3. Frequency Polygon and Frequency Curve:

Polygon is a graph with many angles and frequency polygon is a graph drawn with the help of frequencies. Frequency polygon may be constructed in two ways.

a) From histograms- by joining the mid-points of horizontal straight line of each bar of a histogram.

b) From mid-points of each class- by plotting the frequencies of each class against the mid-points of each class, a frequency polygon can be drawn.

Frequency curve is drawn using freehand through the mid-points of horizontal straight line of each bar of a histogram.

Exercise 9

Draw a frequency polygon and a frequency curve with the help of following data.

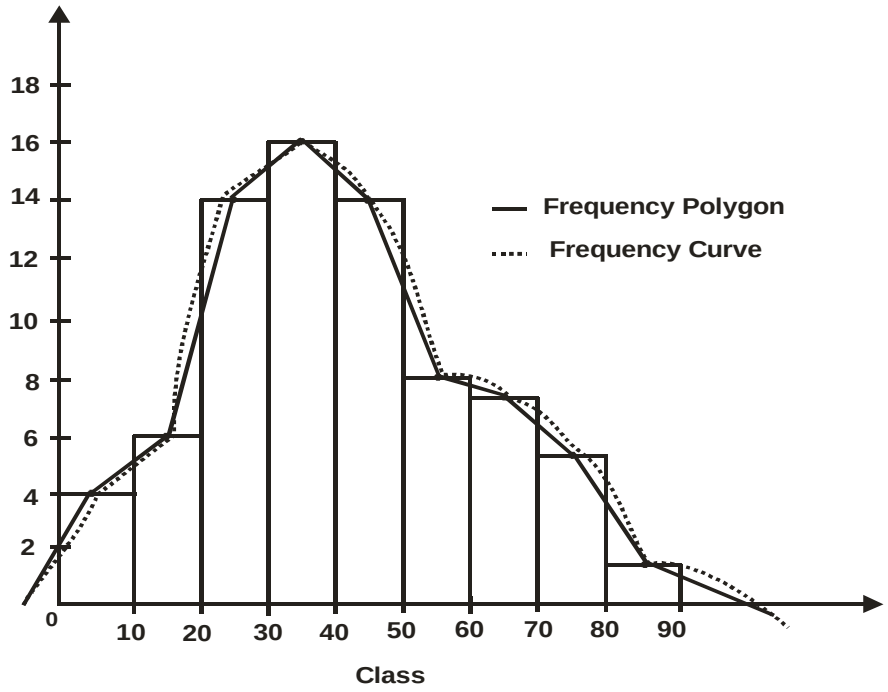
Marks of students	No. of students
0 – 10	4
10 – 20	6
20 – 30	14
30 – 40	16

Steps

1) Draw histogram
2) Take the mid-points of each horizontal line of bar.

40 – 50	14
50 – 60	8
60 – 70	7
70 – 80	5
80 - 90	2

3) Joint the mid-points with straight line to draw frequency polygon.
4) Join mid-points with free hand curve to draw frequency curve.



Check your progress

Draw histogram for the following and locate mode graphically. Also draw frequency polygon and frequency curve.

1)

Classes	200-300	300-400	400-500	500-600	600-700	700-800	800-900
Workers	300	500	800	1000	700	600	400

2)

Weights in (kg)	5-10	10-15	15-20	20-25	25-30	30-35
No. of articles	15	25	30	20	10	7

3)

Marks	0-10	10-20	20-40	40-50	50-60	60-80	80-90
Students	12	18	40	27	32	50	17

7.4..4 Cumulative Frequency curve or Ogives:

These are the curves drawn with the help of cumulative frequencies of “less than” or “more than” type. These curves are

very useful in answering some questions like “How many students have failed?” or “how many children have weights below 10 kg?”

There are two types of ogives:

1) “Less than” ogives- here we start with the upper limit of each class and go on adding the frequencies.
2) “More than” ogives – here we start with the lower limit of each class and go on subtracting the frequencies.

Exercise 10

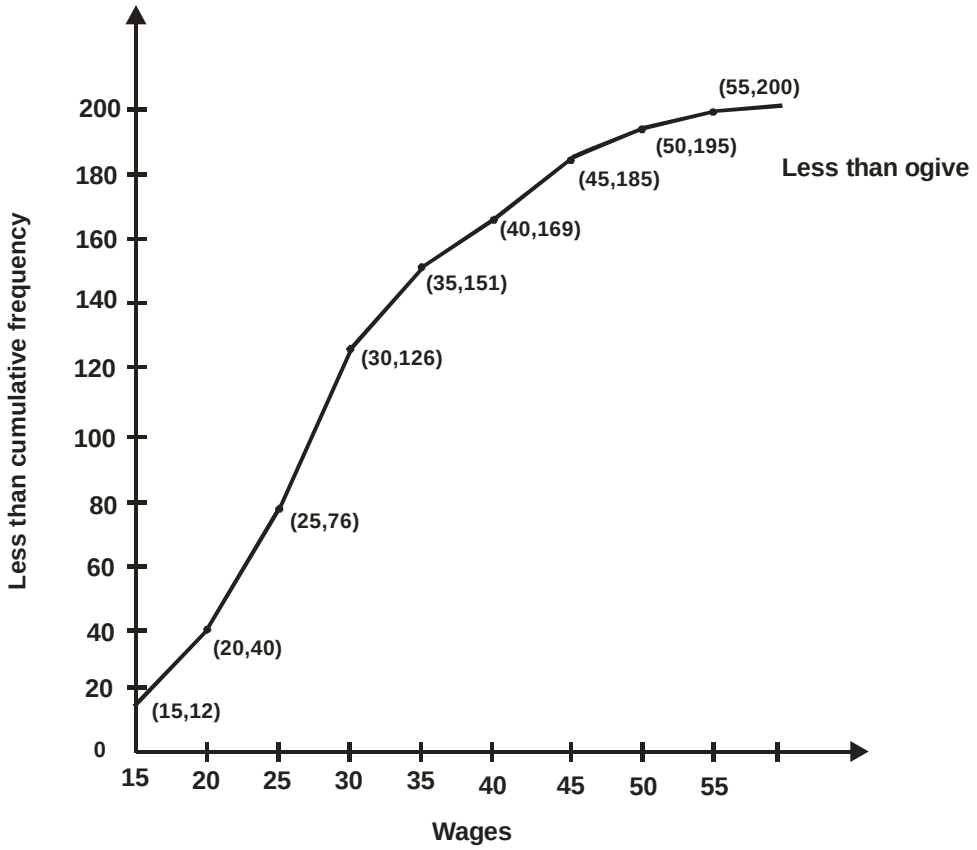
Draw ogives with the help of following data

Wages (Rs.)	No. of workers
10 – 15	12
15 – 20	28
20 – 25	36
25 – 30	50
30 – 35	25
35 – 40	18
40 – 45	16
45 – 50	10
50 - 55	05
N	**200**

First we need to find “less than” cumulative frequencies, for which we need to deal with upper-class limits and then go on adding the frequencies.

Wages less than	Cumulative frequencies
15	12
20	12+28 =40
25	40+36 =76
30	76+50 =126
35	126+25 =151
40	151+18 =169
45	169+16 =185
50	185+10 =195
55	195+15 =200

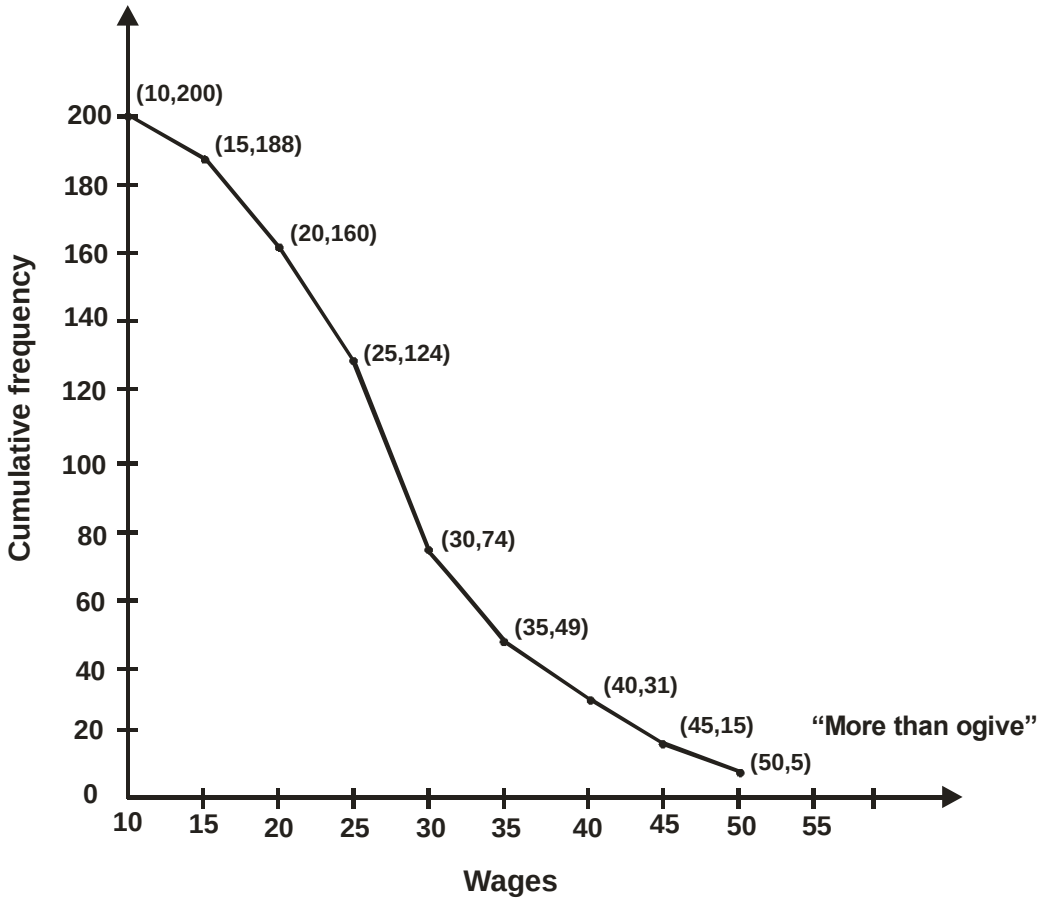
We can draw less than ‘ogive’ with the help of this table.



Now we will find “More than” cumulative frequencies to draw “More than ogive”. For this we need to deal with the lower limits of classes and subtract the frequencies from total frequencies.

Wages More than	Cumulative frequencies
10	200
15	200 -12 = 188
20	188 -28 =160
25	160 - 36 = 124
30	124-50 = 74
35	74 - 25 = 49
40	49 -18 = 31
45	31 -16 = 15
50	15 -10 = 5

We can draw “More than ogive” with the help of this table.



Exercise 11

Draw ogives for the following distribution.

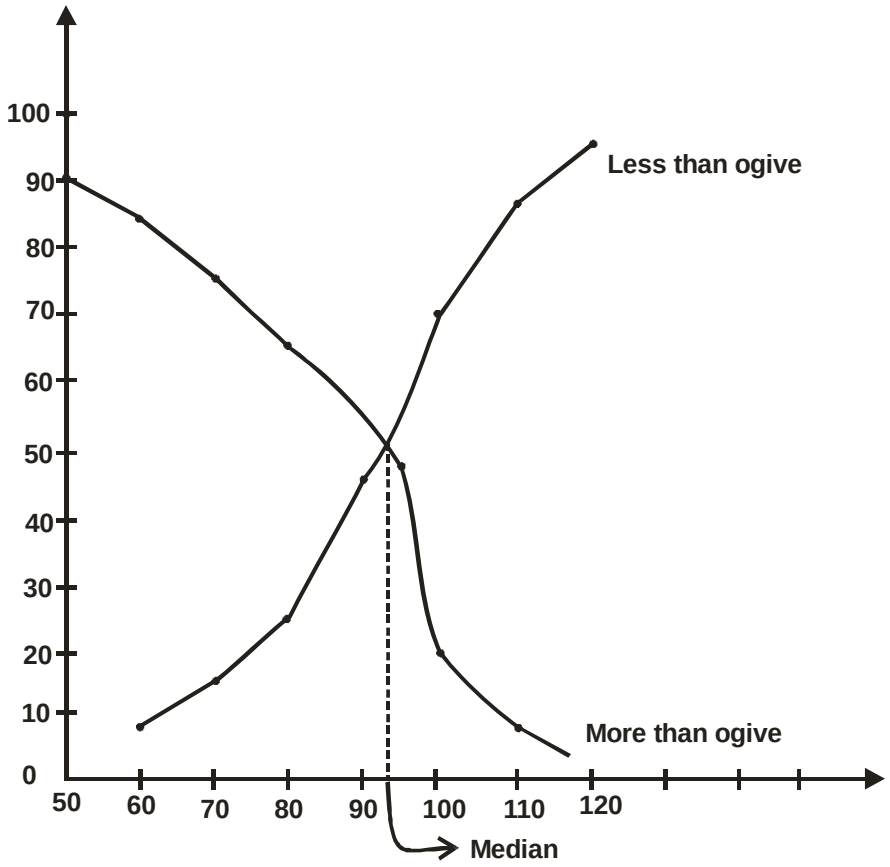
Wages	50-60	60-70	70-80	80-90	90-100	100-110	110-120
Frequency	6	8	12	18	25	16	5

N = 90

Wages less than	Frequency
60	6
70	14
80	26
90	44
100	69
110	85
120	90

Wages more than	Frequency
50	90
60	84
70	76
80	64
90	46
100	21
110	5

We can draw ogives with the help of these two tables.



Check your progress

Draw less than and more than ogive and locate median graphically.

1)

Daily wages (Rs.)	200-220	220-240	240-260	260-280	280-300
No. of workers	24	32	50	17	7

2)

Age in Yrs	0-10	10-20	20-30	30-40	40-50	50-60
No. of Persons	5	15	15	20	8	4

3)

Salaries (Rs.)	300-400	400-500	500-600	600-700	700-800	800-900
No. of employees	12	17	32	17	13	9

7.5 SUMMARY

In this unit we learnt to classify, tabulate and graphically represent the data. This is a very important step involved in research. Different graphs of frequency distribution such as,

a) Histogram
b) Frequency Polygon
c) Frequency Curve and
d) Ogives

are of a great help to the researcher.

7.6 QUESTIONS

1. Explain the meaning & types of classification of data.
2. What is frequency distribution? Discuss the types of frequency distribution.
3. Briefly explain tabulation.
4. Discuss the different types of graph.

8

MEASURES OF CENTRAL TENDENCY

Unit Structure:

8.0 OBJECTIVES

- To learn about different measures of central tendency.
- To determine these measures with the help of graphs.
- To understand their merits and demerits.

8.1 INTRODUCTION

An important task of statistical analysis is to get one single value that represents the important characteristic of the data. It is difficult to remember the entire set of data. In fact, a single representative figure describing the whole data will be more useful and meaningful for interpretation and comparison. For example, instead of stating marks of all subjects in the T.Y.B.A., it is more meaningful to give average marks in T.Y.B.A. says 62.3%. Thus, an important function of measures of central tendency, is to compute a single representative value that reveals the important features of a variable.

In the next section, we will discuss the meaning, types and merits and demerits of measures of central tendency.

8.2 MEASURES OF CENTRAL TENDENCY

The measures of central tendency or an average is a single value selected from a group of values to represent them in some

way (A.E. Waugh). In other words, an average or a measure of central tendency represent the entire data. The value of an average lies somewhere between the two extreme values (the smallest and the largest) in a data set. It is difficult to remember the incomes of millions of individuals. But it is easy to remember an average income of a country. Such a single figure, representing the entire population also is useful for comparisons. For example, average income of India can be easily compared with an average income of the USA, average marks of a student can be easily compared with the average marks of another student.

Following are important measures of central tendency or averages:

a) Arithmetic Mean b) Median c) Mode
d) Geometric Mean e) Harmonic Mean

Following is a brief overview of the first three measures of Central Tendency.

8.2.1 Arithmetic Mean (A.M.)

The most popular and most widely used measure of central tendency is arithmetic mean. Its value is obtained by adding the values of all observations in a given data and then dividing the sum by the number of observations in the data.

As noted earlier, there are three types of data:

i) Individual observations or ungrouped data where the frequencies are not given.

ii) Discrete data.

iii) Continuous data

Calculation of AM differs as per the nature of the data.

8.2.1.1

Calculation of AM: Individual observations. It can be done by two methods – direct and short-cut method.

I) <u>Direct Method :</u>

The direct method of calculating AM for ungrouped data is very simple. Here the AM is calculated by adding together the values of given variable and then by dividing the total by the number of observations.

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{N} \quad \text{i.e. } \bar{x} = \frac{\sum x}{N}$$

Where $\bar{x}$ (x bar) means AM

$x_1, x_2, x_3, \ldots x_n$ different values of a given variable

$\sum x$ (summation x) – total of values of the variables

N – number of observations

Example 1

Calculate AM for the following data of the monthly income of 10 families.

Families	1	2	3	4	5	6	7	8	9	10
Income (Rs.) thousand	18	20	35	55	40	52	90	95	27	63

Steps: 1) Find $\sum x$ i.e. find the total of all the values of given variable (income, in this case)

2) Divide the total by the number of observations (Families in this case).

Table of Calculation for Mean

Families	Income (Rs. ,000)
1	18
2	20
3	35
4	55
5	40
6	52
7	90
8	95
9	27
10	63
N = 10	$\sum x$ = 495

$$\bar{x} = \frac{\sum x}{N}$$

$$= \frac{495}{10}$$

$$= 49.5$$

Mean income of the families is Rs. 49.5 thousand

II) <u>Short Cut Method :</u>

Sometimes, the values of a variable are very large and hence computing AM by direct method becomes tedious. In such a case AM can be calculated by using an alternative method known as short-cut method. Here the deviations (differences) are taken from some assumed mean an then the actual mean is calculated. Accordingly,

$\bar{x} = A + \frac{\sum d}{N}$ Where $\bar{x} \rightarrow$ Arithmetic Mean

$A \rightarrow$ Assumed Mean
$d \rightarrow$ (X-A) or deviations from assumed Mean
$N \rightarrow$ No. of observations

Example 2

From the following data of daily sales of a shop for 7 days, calculated mean sales.

Day	Monday	Tuesday	Wed	Thu	Friday	Sat	Sun
Sales (in units)	128	157	107	155	219	358	410

Steps: 1) Take assumed mean (which can be any value of a given variable in the data)
2) Find deviations from the assumed mean (d = x – A).
3) Find $\sum d$ and apply the formula

Table of Calculation for Mean

Day	Sales	d = x - A
Mon	128	128-200 = -72
Tue	157	157-200 = -43
Wed	107	107-200 = -93
Thur	155	155-200 = -45
Fri	219	219-200 = 19
Sat	358	358-200 = 158
Sun	410	410-200 = 210
N = 7		$\sum d$ = 134

Let A = 200

$$\bar{x} = A + \frac{\sum d}{N}$$

$$= 200 + \frac{134}{7}$$

$$= 200 + 19.14$$

$$= 219.14$$

Mean Sales = 219.14 units

Check your progress

1. Compute AM by both direct and Short-cut Methods.

Roll No.	1	2	3	4	5	6	7	8	9	10
Marks in Stats	67	69	66	68	63	76	72	74	70	68

(Ans: 69)

2. Find the average height (in cm) for the following data.

Student	1	2	3	4	5	6	7	8
Height (cm)	121	135	127	130	141	132	130	140

(Ans: 132)

3. What is AM for the following data?

Observation	1	2	3	4	5	6	7	8	9	10
Value	79	82	36	38	51	72	68	70	64	63

(Ans: 62.3)

8.2.1.2 Calculation of AM: Discrete Data

I) <u>Direct Method</u>:

Discrete data is the data with frequencies the following formula is used to calculate AM for the such data.

$$\bar{x} = \frac{\sum fx}{\sum N}$$ Where $\bar{x} \rightarrow$ Arithmetic Mean

$f \rightarrow$ Frequency

$x \rightarrow$ value of a given variable

$N \rightarrow$ No. of observations (or total frequency)

Example 3

Following data gives marks of 50 students. Calculate average marks for the class.

Marks	5	6	7	8	9	10
No. of Students	8	10	12	9	7	4

Steps: 1) Multiply each value of a given variable (Marks in this case) with its respective frequency and get fx.

2) Find $\sum fx$

3) Apply the formula

Table of Calculation for Mean

Marks	No. of Students	fx
x	f	
5	8	40
6	10	60
7	12	84
8	9	72
9	7	63
10	4	40
	$\sum f = 50$	$\sum fx = 359$

$$\bar{x} = \frac{\sum fx}{N}$$

$$= \frac{359}{50}$$

$$= 7.18$$

Average Marks 7.18

II) <u>Short Cut Method :</u>

To simplify the calculations, short-cut method can be used in which the AM is calculated using the following formula:

$$\bar{x} = A + \frac{\sum fd}{N}$$ Where $\bar{x} \rightarrow$ Arithmetic Mean

$A \rightarrow$ Assumed Mean
$f \rightarrow$ frequency
$d \rightarrow$ deviations from assumed Mean (X-A)
$N \rightarrow$ total frequency

Example 4

From the following data of the marks of the students in a class of 60, compute AM using short-cut method.

Marks	10	20	30	40	50	60	70
Student	7	12	11	15	8	5	2

Steps: 1) Take an assumed mean (Any value from the x values)
2) Take deviations of x from the assumed mean (x – A) & denote it on 'd'.
3) Multiply 'd' by f and obtain $\sum fd$.
4) Apply the formula

Table of calculation for Mean

Marks (x)	No. of Students (f)	D = X – A X - 40	fd
10	7	10-40 = -30	-210
20	12	20-40 = -20	-240
30	11	30-40 = -10	-110
40	15	40-40 = 0	0
50	8	50-40 = 10	80
60	5	60-40 = 20	100
70	2	70-40 = 30	60
	$\sum f = 60$		$\sum fd$ = -320

Let A = 40 (Generally a middle X – value is taken as assumed mean)

$$\bar{x} = A + \frac{\sum fd}{\sum f}$$

$$= 40 + \frac{-320}{60} \qquad = 40 + (-5.33)$$

$$= 34.67$$

Mean Marks 34.67

Check your Progress

1. Calculate AM for the following by both direct and short-cut method.

X	30	40	50	60	70	80	100
f	4	15	20	25	18	12	6

(Ans: 60.4)

2. Find mean wage for the following

Wage (Rs.)	50	70	90	110	120	130
No. of workers	15	18	22	17	13	15

8.2.1.3 Calculation of AM: Continuous data

I) <u>Direct Method</u>:

Continuous data is the data with classes and frequencies. Using direct method, AM for such data is calculated using following formula.

$\bar{x} = \frac{\sum fm}{N}$ Where $\bar{x} \rightarrow$ Arithmetic Mean

$f \rightarrow$ frequency
$m \rightarrow$ mid point of a class
$N \rightarrow$ No. of observations

Example 5

Determine AM for the following data:

Classes	0-5	5-10	10-15	15-20	20-25	25-30
Frequencies	10	15	17	13	5	4

Steps: 1) Find the mid-points (m) of each class.

$$\left(\frac{\text{upper class limit} - \text{lower class limit}}{2}\right)$$

2) Multiply the mid-points (m) by the frequency of the corresponding class.
3) Find $\sum fm$ and apply the formula

Table of calculation for Mean

Classes (x)	Frequency (f)	Mid-points (m)	fm
0-5	10	2.5	25.0
5-10	15	7.5	112.5
10-15	17	12.5	212.5
15-20	13	17.5	227.5
20-25	5	23.5	117.5
25-30	4	27.5	110.0
	$\sum f = 64$		805.0

$\bar{x} = \frac{\sum fm}{N} = \frac{805}{64} = 12.58$ $\boxed{\bar{x} = 12.58}$

II) <u>Step Deviation Method :</u>

The arithmetic mean can also be found out by using the step deviation method. This method simplifies the calculations required in finding the mean value. The formula is as follows:

$\bar{x} = A + \frac{\sum fd'}{N} \times i$ Where $A \rightarrow$ Assumed mean

$f \rightarrow$ frequency

$d' = \left(\frac{m - A}{i}\right)$

$m \rightarrow$ mid point

$i \rightarrow$ Class interval

Example 6

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Students	15	20	25	24	12	31	71	52

Steps: 1) Find the mid-points of each class.
2) Take assumed mean from the mid-points.
3) Find d' and fd'.
4) Find $\sum fd'$ and apply the formula.

Marks (x)	Students (f)	Mid-points (m)	$d' = \left(\frac{m - 35}{10}\right)$	fd'
0-10	15	5	-3	-45
10-20	20	15	-2	-40
20-30	25	25	-1	-25
30-40	24	35	0	0
40-50	12	45	1	12
50-60	31	55	2	62
60-70	71	65	3	213
70-80	52	75	4	208
	$\sum f = 250$			$\sum fd = 385$

Let $A = 35$ $i = 10$

$$\bar{x} = A + \frac{\sum fd'}{\sum f} \times i$$

$$= 35 + \frac{385}{250} \times 10$$

$$= 35 + 1.54 \times 10$$

$$= 35 + 15.4$$

$$= 50.4$$

Mean $= 50.4$ marks

Check your progress

1) Find arithmetic mean by both direct and step-deviation methods.

Variable	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	10	20	35	40	25	25	15

(Ans: 36.88)

2)

Weights in grams	110-120	120-130	130-140	140-150	150-160	160-170
No. of Apples	5	8	12	18	22	9

Weights in grams	170-180	180-190
No. of Apples	7	4

8.2.1.4 Merits and demerits of Arithmetic Mean

Merits:

1) It is the easiest measure of central tendency both for understanding and calculation. Hence, it is most widely used.

2) It is based on each and every observation in the data.

3) It is defined in such a way that everyone gets the same result for a single problem.

4) Further algebraic treatment is possible in case of arithmetic mean.

Demerits:

1) Since the value of arithmetic mean is based on each and every observation in the data, it is affected by the extreme values.

2) In a data-set with open-end classes, the value of mean can not be calculated without making assumptions.

8.2.2 Median

Median is the middle value in a data. It divides the data in such a way that half of the observations in the data have value less than the median and the rest half have value more than the median. Median is called a positional average.

For example, if the ages of 5 children in a class are given below:

Child	Ramesh	Suresh	Bhavesh	Nimesh	Girish
Age in yrs.	05	07	10	11	12

then, median age is Bhavesh's age i.e. 10 years. Which has a middle position in the data. Two children are younger and two children are older than Bhavesh. Thus, median is a positional average.

8.2.2.1 Calculation of median for individual observations

To calculate median for ungrouped data, following steps should be followed.

1. Arrange the data in ascending (smallest to largest) or descending (largest to smallest) order.

2. If the number of observations is odd, median is the size of $\frac{N+1}{2}$th observation in the data (where N- number of observations).

3. If the number of observations is even, the value of median is estimated by finding the mean of two middle observations.

Example 7

Calculate median weight for the following:

Weights in (kg)	12	17	30	23	15	18	10

Solution:

Arranged data	10	12	15	17	18	23	30

N = 7 (odd value)

Median is the size of $\frac{N+1}{2}$th observation $= \frac{7+1}{2} = 4^{th}$ observation.

The size of 4^{th} observation is 17 kg.

$$\boxed{\text{Median weight} = 17\,\text{kg}}$$

Example 8

Calculate median for following data of marks of 10 students in a class.

Marks	15, 25, 30, 32 ,10, 12, 22, 38, 41, 31

Solution:

Arranged data	10 12 15 22 25 30 31 32 38 41

N = 10 (Even value)

Median is the size of $\frac{N+1}{2}$th observation $= \frac{10+1}{2} = 5.5^{th}$ observation.

$$\text{Size of } 5.5^{th} = \frac{5^{th}\text{observation} + 6^{th}\text{observation}}{2}$$

$$= \frac{25+30}{2} = 27.5$$

$$\boxed{\text{Median} = 27.5 \text{ marks}}$$

Check your progress

Calculate median for the following

i) Height (inches) 13 17 25 38 8 11 24

ii) Profits (%) 18 25 27 5 10 2 12 16

iii) Wages (Rs) 101 150 85 95 160 205 148

8.2.2.2 Calculation of median for the discrete data

To calculate median for the discrete data, following steps should be followed:

1. Arrange the data in ascending or descending order (if not arranged).
2. Find "less than" cumulative frequencies.
3. Apply the formula "Median is $\frac{N+1}{2}^{th}$ observation in the data".
4. From the column of cumulative frequencies, find the value equal to or next higher than the value obtained by using $\frac{N+1}{2}$ formula.
5. The value of variable corresponding to this obtained value is median.

Example 9

Given the wages (Rs.) for 70 workers in a factory, find median wages.

Wages (Rs) (x)	No. of workers (f)	Cumulative frequency (c.f.)
100	7	7
150	12	19
200	15	34
250	18	52
300	10	62
350	8	70
	$\sum f = 70$	

Solution:

Median is the size of $\frac{N+1}{2}^{th}$ observation.

$= \frac{70+1}{2}^{th}$ observation = 35.5^{th} observation

Size of 35.5^{th} observation is 250 [first look for value 35.5 in the c.f. column. If that value is not there, then look for next higher value i.e. 52. Median is value corresponding to 52.]

Wages corresponding to 35.5 th observation are Rs. 250

Median wage = Rs.250

Check your progress

1. Find median for the following data.

i)

Weights	5	7	9	11	13	15
Students	10	18	20	22	10	5

ii)

Income (Rs.)	80	100	150	180	200	250
No. of persons	16	24	26	30	20	6

8.2.2.3 Calculation of median for continuous data

While calculating median for a continuous data, following steps should be followed.

1. Determine the median class using the formula: median class is the one where $\frac{N}{2}$ th observation lies.
2. Then apply the another formula given below to calculate actual median.

$$\text{Median} = L_1 + \left(\frac{N/2 - c.f}{f} \right) \times i$$

where L_1 = Lower class limit of median class

c.f. = cumulative frequencies of the class before median class.

f = frequency of the median class.

N = number of observations.

i = class interval.

Example 10

Calculate median for the following data:

Income	Frequency	Less than cumulative frequency
80-100	5	5
100-120	12	17
120-140	30	47
140-160	20	67
160-180	16	83
180-200	10	93
200-220	7	100
	N = 100	

Solution:

Median is the size of $\frac{N}{2}^{th}$ item $= \frac{100}{2} = 50^{th}$ item.

Median Class is 140 – 160

$$\text{Median} = L_1 + \left(\frac{N/2 - c.f}{f} \right) \times i$$

$$= 140 + \frac{100/2 - 47}{20} \times 20$$

$$= 140 + \frac{50 - 47}{20} \times 20$$

$$= 140 + \frac{3}{20} \times 20 \quad = 140 + 3$$

$$= 143$$

$$\boxed{\text{Median} = \text{Rs.}143}$$

Check your progress

Find median for the following

i)

x	0-10	10-20	20-30	30-40	40-50	50-60
f	7	12	15	20	16	10

ii)

x	5-15	15-25	25-35	35-45	45-55	55-65

f	15	20	30	18	12	5

8.2.3. Quartiles

As seen in the earlier section, median divides the data into two equal parts. When the data is divided into four equal parts, there will be three points of division. These points are called quartiles. Q_1 is the first quartile, which divides the data in such a way that 25% observations in the data have lower value and 75% observations have higher value than Q_1. Q_2 or median divides the data is such a way that 50% observations have lower and 50% have higher value than Q_2. Finally Q_3 divides the data in such a way that 75% observations have lower and 25% observations have higher value than the third quartile.

Quartiles are calculated in the same way as the median. Following example will illustrate the same.

Example 10

For the following data, compute first and the third quartile.

15 18 25 20 22 12 13

Solution

Arranged data: 12 13 15 18 20 22 25

Q_1 is the size of $\frac{N+1}{4}^{th}$ observation

$= \frac{7+1}{4} = \frac{8}{4} = 2^{nd}$ observation

$\boxed{Q_1 \text{ First Quartile } = 13}$

Q_3 is the size of $\frac{3\ N+1}{4}^{th}$ observation

$= \frac{3\ 7+1}{4}^{th}$ observation $= \frac{24}{4} = 6^{th}$ observation

$\boxed{Q_3 \text{ Third Quartile } = 22}$

Example 11

Calculate Q_1 and Q_3 for the following:

x	f	Cumulative frequencies
5	9	9
7	15	24
9	18	42
11	20	62
13	17	79
15	5	84
	N = 84	

Q_1 is the size of $\frac{N+1}{4}$ th item in the data

$$=\frac{84+1}{4}=\frac{85}{4}=21.25$$

Size of 21.25 th item is 7

$$\boxed{Q_1=7}$$

Q_3 is the size of $\frac{3\ N+1}{4}$ th item in the data

$$=\frac{3\ \ 84+1}{4}=\frac{255}{4}=\ 63.75$$

Size of 63.75th item is 13

$$\boxed{Q_3=13}$$

Example 12

Calculate first and the third quartiles for the following data.

Income (000)	8-10	10-12	12-14	14-16	16-18	18-20	20-22
Families	5	12	30	20	16	10	7

Solution:

Income (,000)	Families	c.f.
8-10	5	5
10-12	12	17
12-14	30	47
14-16	20	67
16-18	16	83
18-20	10	93
20-22	7	100
	100	

Q1 is the size of $\frac{N}{4}^{th}$ item $= \frac{100}{4} = 25^{th}$ item.

$\therefore$ First Quartile class is 12-14.

$$Q_1 = L_1 + \left(\frac{N/4 - c.f}{f} \right) \times i$$

$$= 12 + \left(\frac{100/4 - 17}{30} \right) \times 2$$

$$= 12 + \left(\frac{25 - 17}{30} \right) \times 2$$

$$= 12 + \frac{8}{30} \times 2$$

$$= 12 + 0.27 \times 2$$

$$= 12 + 0.54 \quad = 12.54$$

$$\boxed{Q_1 = 12.54}$$

Q3 is the size of $\frac{3N}{4}^{th}$ item $= \frac{300}{4} = 75^{th}$ item

$\therefore$ Third quartile class is 16-18

$$Q_3 = L_1 + \left(\frac{3N/4 - c.f}{f} \right) \times i$$

$$Q_3 = 16 + \left(\frac{75-67}{16}\right) \times 2$$

$$= 16 + \frac{8}{16} \times 2$$

$$= 16 + 0.5 \times 2$$

$$= 16 + 1 = 17 \qquad \boxed{Q_3 = 17}$$

Check your progress

Calculate quartiles for the following data:

1. 25 27 20 27 21 28 19
2. 85 87 89 92 75 65 77 80 81 72 70
3.

x	0-5	5-10	10-15	15-20	20-25	25-30
f	10	20	40	65	50	15

8.2.4 Merits and demerits of median

Merits

1. Median is a useful measure of central tendency particularly for the open ended classes. This is because, it is the positional average.

2. It is not influenced by the extreme items in the data i.e. very small or very large values in the data.

3. It is the most suitable measure for the qualitative type of data where ranks are given.

4. In the skewed distribution, it is regarded as more representative types of average.

Demerits

1. Arrangement of the data is required for calculating median.

2. Median can not be determined for combined groups. So it is not capable of further statistical treatment.

8.2.5 Mode

Mode is that value in the data which occurs with highest frequency. Sometimes mean and median fail to represent certain characteristics of data. For example, most common wages, most common height, etc. In such situations, mode is the most suitable measures. Calculation of mode, for different types of data is shown below.

Example 13

What is the modal wage for the following?
150 150 170 180 190 150 170 170 150.
Since 150 repeats 4 times, it is the mode. (Mode is the value with highest frequency). $\therefore$ $\boxed{\text{Mode} = \text{Rs.}150}$

Example 14

What are the modal marks?

Marks	15	16	17	18	19	20	21	22
No. of Students	8	15	25	30	32	28	25	10

Modal marks are 19, since the frequency related to number 19 is the highest (30 times).

Example 15

Calculate mode for the following data.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Students	4	6	20	32	18	12	8

To calculate mode for the continuous data, following steps are necessary.

1) Find the modal class i.e. a class with high frequency (30-40 in this example)

2) Apply the formula

$$\text{Mode} = L + \frac{\Delta 1}{\Delta 1 + \Delta 2} \times i$$

Where L is the lower limit of the modal class

$\Delta 1$ difference between the frequency of the modal class (f1) and the pre-modal class (f0)

$\Delta 2$ is a difference between the frequency of the modal class (f1) and the post modal class (f2)

i is class interval.

Solution:

In the above example modal class is 30-40
L is 30
i is 10
$\Delta 1$ is 12 (f1-f0 = 32-20 = 12)

$\Delta 2$ is 14 (f1-f2 = 32-18)

By substituting the values

$$\text{Mode} = 30 + \frac{12}{12+14} \times 10$$

$$= 30 + 0.46 \times 10$$
$$= 30 + 4.6$$
$$= 34.6$$

Check your progress

Calculate mode for the following data

1. 12 12 23 24 23 23 34 12 24 25 26 24 25 12 11
2. 56 57 58 54 56 57 53 54 57 56 54 57 54 57

Calculate mode for the following sets of data.

1

X	12	13	14	15	16
F	15	20	22	24	20

2

X	45	46	47	48	49
F	25	28	32	36	25

3

X	5-10	10-15	15-20	20-25	25-30	30-35
F	8	12	20	15	10	5

4

X	20-30	30-40	40-50	50-60	60-70	70-80	80-90
F	12	15	22	28	24	23	18

5

X	50-100	100-150	150-200	200-250	250-300	300-350

F	3	7	10	15	12	8

6

X	2-4	4-6	6-8	8-10	10-12	12-14
F	12	15	18	20	12	8

7

X	45	46	47	48	49
F	8	10	12	15	5

8.3 GRAPHICAL DETERMINATION OF MEDIAN, QUARTILES AND MODE

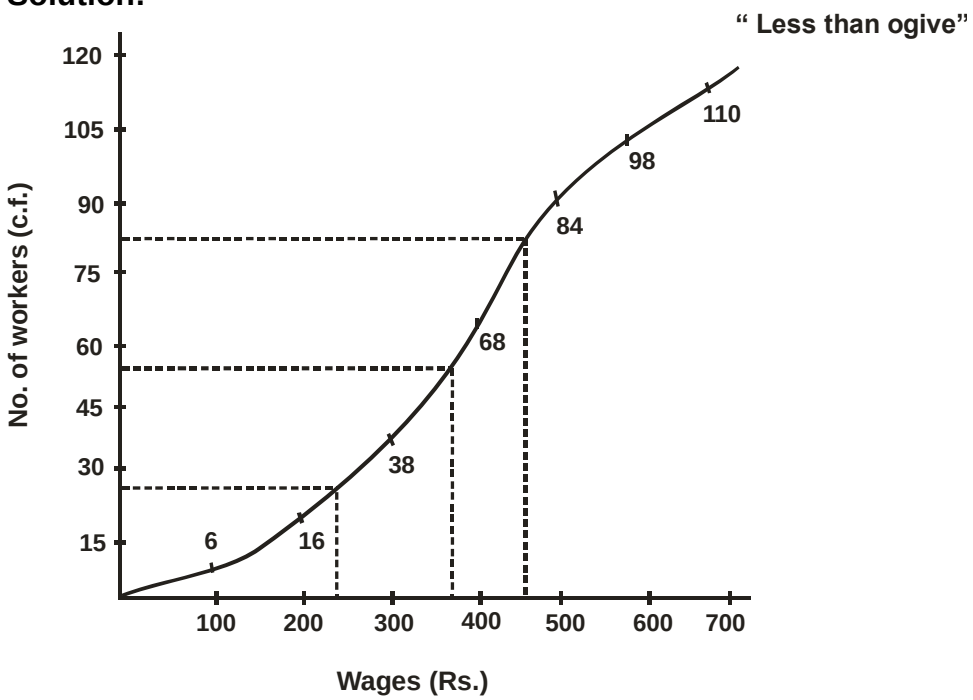
Median and quartiles can be graphically determined by drawing ogive by “less than” method. It is shown below.

Example 16

For the following data, determine median, Q_1 and Q_3 graphically.

Wages (Rs)	Workers	Less than c.f.
0-100	6	6
100-200	10	16
200-300	22	38
300-400	30	68
400-500	16	84
500-600	14	98
600-700	12	110
	110	

Solution:



Median is the size of $\frac{N}{2}^{th}$ item $= \frac{110}{2} = 55^{th}$ item. Locate this value on Y – axis (55) and draw a perpendicular on less than ogive. From that point, draw a perpendicular on x-axis. The point where it meets x-axis is median (Approximately 356).

Q_1 is the size of $\frac{N}{4}^{th}$ item $= \frac{110}{4} = 27.5^{th}$ item. Locate the point (27.5) on Y-axis, draw a perpendicular on less than ogive. From the point of where it meets the ogive, draw perpendicular to X-axis. The point where it meets X-axis is Q_1 (approximately 252).

Q_3 is the size of $\frac{3N}{4}^{th}$ item $= 82.5^{th}$ item

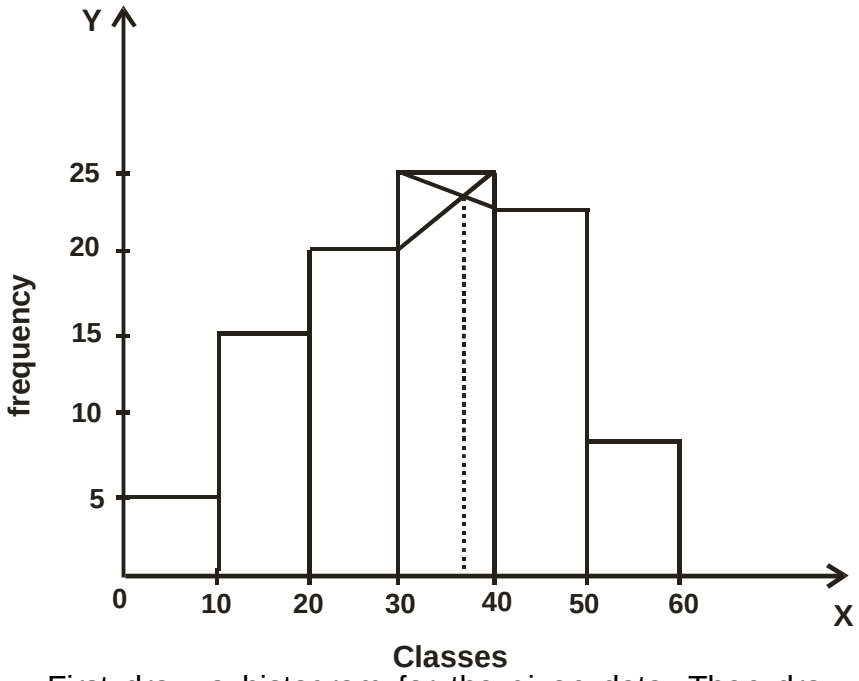
Follow the steps mentioned earlier to get the value of Q_3 (approximately 490).

Mode can be graphically determined by using histograms. This is done as follows.

Example 17

For the following data, determine mode graphically.

X	0-10	10-20	20-30	30-40	40-50	50-60
Y	5	15	20	25	18	7



First draw a histogram for the given data. Then draw two lines diagonally inside the modal class bar (the bar representing highest frequency). Then draw a perpendicular from the point of intersection on the X-axis. Value of mode is approximately 38 in this example.

8.4 SUMMARY

In this unit, we learned about important measures of the central tendency. These are widely used in statistical analysis of data in all branches of knowledge.

8.5 QUESTIONS

1) Explain the various methods of calculating Arithmetic Mean.

2) Explain briefly Arithmetic Mean. Explain its merits and demerits.

3) Discuss various methods of calculation of Median.

9

MEASURES OF DISPERSION

Unit Structure:

9.0 OBJECTIVES

- To learn about different measures of dispersion and their relative importance.
- To learn the meaning of skewness.
- To understand simple measures of skewness.

9.1 INTRODUCTION

We have learnt the important central tendency measures in the last unit. As we already know, these measures give us a single value that represents the entire data set. Sometimes, however, we need some more information about the data. Measures of dispersion are an additional step forward in understanding another important feature of the statistical data. They aim at measuring the degree of variability or scatteredness of various items in the data from the central value. Dispersion supplements the measures of central tendency by revealing some more features of given data.

The degree to which numerical data tends to spread about an average value is called the variation or dispersion of data Spiegel. In other words, dispersion measures the extent to which the given values vary from their central value.

An importance of the measures of dispersion may be well understood from the following example.

	Series A	Series B	Series C
	50	50	1
	50	47	245
	50	53	2
	50	55	1
	50	45	1
Total	250	250	250
Arithmetic Mean	50	50	50

All the three data sets given above have same arithmetic mean. But a closer look at each set will reveal that each of those groups are different from each other to a great extent. First group is homogenous, second group is also more or less homogenous but the third group is quite heterogeneous. But each set has the same central value. Thus, to know more about the given data, central tendency measure are not sufficient. We need to further understand the degree of dispersion related to that data.

Following are important measures of dispersion:

1) Range
2) Quartile Deviation
3) Mean Deviation
4) Standard Deviation

9.2 RANGE

Range is the simplest measures of dispersion. Range is defined as the difference between the highest observation (H) and the lowest observation (L) of the variable given in the data.

Range = H – L

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

9.2.1 Range for raw data / individual observation:

Ex. 1
From the following data, find range and its coefficient.
25, 27, 17, 38, 19, 10, 30

H = 38 L = 10
Range = H – L
= 38 – 10 = 28

Co-efficient of Range = $\frac{H-L}{H+L}=\frac{38-10}{38+10}=\frac{28}{48}$

= 0.58

9.2.2 Range for discrete data

Ex. 2
From the following data of weights of students, calculate range and its co-efficient.

Weights (in kg)	No. of students
25	50
27	75
29	102
31	87
33	71
35	60

H = 35 L = 25

Range = H – L
= 35 -25
= 10

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

$=\frac{35-25}{35+25}=\frac{10}{60}$

= 0.17

9.2.3 Range for continuous data

Ex. 3
From the following data set, calculate range and its coefficient.

Class	Frequency
0 – 10	7
10 – 20	12
20 – 30	18
30 – 40	25
40 – 50	38
50 – 60	23
60 – 70	15
70 – 80	8
80 – 90	3

H = 90 L = 0

Range = H – L
= 90 – 0 = 90

Co-efficient of Range $=\frac{H-L}{H+L}=\frac{90-0}{90+0}$

$=\frac{90}{90}=1$

Check your progress

What is the range and its co-efficient for the following data?

1) 7, 15, 8, 10, 6, 17, 3

2) 256, 537, 238, 407, 305, 487, 500

3)

x	15	20	25	30	35	40
f	9	19	29	18	8	5

4)

x	2.5	5.5	8.5	11.5	14.5	17.5
f	10	15	25	21	12	8

5)

Marks	15-25	25-35	35-45	45-55	55-65
Students	3	5	8	12	10

6)

Class	250-500	500-750	750-1000	1000-1250
Frequency	25	38	23	10

9.3 QUARTILE DEVIATION

This measure of dispersion is based on the quartiles. As we have studied in the earlier unit, there are three quartiles. Q_1 is called lower quartile, Q_3 is called upper quartile and Q_2 is more popularly known as median. Following three measures are computed using first and the third quartile.

Inter-quartile Range - $Q_3 - Q_1$

Quartile Deviation - $\dfrac{Q_3 - Q_1}{2}$

Coefficient of quartile deviation - $\dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

We have already learnt the computation of Q_1 and Q_3 for different types of data individual or raw data, discrete data and continuous data.

9.3.1 Quartile deviation for individual observation

Ex. 4
Find quartile deviation (Q.D.) coefficient of quartile deviation for the following data.

30	40	53	55	60	63	65

Solution:

Number of observations N = 7

Q_1 is the size of $\frac{N+1}{4}$th observation in the data

$\frac{7+1}{4}$th observation or 2nd observation in the data.

$\therefore Q_1 = 40$

Q_3 is the size of $\frac{3\ N+1}{4}$th observation in the data, which is $\frac{3\ 7+1}{4}$th or 6th observation in the data.

$\therefore\ Q_3 = 63$

Inter quartile Range = $Q_3 - Q_1$

$= 63 - 40 = 23$

Quartile Deviation $= \frac{Q_3 - Q_1}{2} = \frac{63-40}{2} = \frac{23}{2}$

$= 11.5$

Co-efficient of Quartile deviation = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{63-40}{63+40} = \frac{23}{103} = 0.22$

9.3.2 Quartile deviation for discrete data:

Ex. 5

x	f	Cumulative frequency
80	12	12
85	18	30
90	26	56
95	15	71
100	5	76
105	4	80
N =	80	

Q_1 is the size of $3\left(\frac{N}{4}\right)$ observation

$= \frac{80}{4} = 20\text{th}$ observation $\quad \therefore Q_1 = 85$

Q_3 is the size of $3\left(\frac{N}{4}\right)$th observation

$= \frac{240}{4} = 60^{th}$ observation

$\therefore Q_3 = 95$

Interquartile Range $= Q_3 - Q_1$
$= 95 - 85$
$= 10$

Quartile Deviation $= \frac{Q_3 - Q_1}{2} = \frac{10}{2} = 5$

Co-efficient of quartile deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{95 - 85}{95 + 85} = \frac{10}{180} = 0.06$

9.3.3 Quartile Deviation for continuous data:

Ex. 6
Compute Quartile Deviation for the following data

x	f	Cumulative frequency
0 – 10	12	12
10 – 20	18	30
20 – 30	26	56
30 – 40	15	71
40 – 50	9	80
	80	

Q_1 is the size of $\frac{N}{4}$th item $= \frac{80}{4} = 20^{th}$ item.

$\therefore$ First quartile class is 10 – 20

$Q_1 = L_1 + \left(\frac{N/4 - c.f}{f}\right) \times i$

$$= 10 + \left(\frac{80/4 - 12}{18} \right) \times 10$$

$$= 10 + \left(\frac{20 - 12}{18} \right) \times 10 = \boxed{Q_1 = 14.44}$$

Q_3 is the size of $\frac{3N}{4}$ th item $= \frac{240}{4} = 60^{th}$ item.

$\therefore$ Third Quartile class is 30 – 40

$$Q_3 = L_1 + \left(\frac{3N/4 - c.f}{f} \right) \times i$$

$$= 30 + \left(\frac{240/4 - 56}{15} \right) \times 10$$

$$= 30 + \left(\frac{60 - 56}{15} \right) \times 10 = 32.67$$

$$\boxed{Q_3 = 32.67}$$

Inter quartile Range = $Q_3 - Q_1$

$= 32.67 - 14.44$

$= 18.23$

Quartile Deviation $= \frac{Q_3 - Q_1}{2}$

$$= \frac{32.67 - 14.44}{2}$$

= 9.11

Coefficient of Quartile Deviation $= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{32.67 - 14.44}{32.67 + 14.44}$

= 0.39

Check your progress

Compute inter-quartile range, quartile deviation and its coefficient for the following data.

1) 56, 58, 60, 62, 65, 68, 72

2) 5, 7, 10, 12, 8, 6, 9, 15, 18, 12, 11

3)

x	7	14	21	28	35	42	49
f	5	12	30	20	16	10	7

4)

x	5	10	12	15	16	18	20
f	7	10	14	13	4	6	6

5)

Class	25-30	30-35	35-40	40-45	45-50	50-55	55-60
Frequency	6	7	10	20	12	10	5

6)

Class	0-20	20-40	40-60	60-80	80-100
Frequency	24	34	38	30	24

9.4 STANDARD DEVIATION (S.D.) & COEFFICIENT OF VARIATION (C.V)

It was introduced by Karl Pearson in 1823. It is the most widely used measure of dispersion. It is denoted by a Greek letter σ (sigma). Standard deviation is an absolute measure of dispersion. Corresponding relative measure is known as co-efficient of variation. It is used to compare the dispersion or variability of two or more groups. A group or a data set for which the value of coefficient of variation (C.V.) is high is said to be having more variability. A group which has lower value of coefficient of variation is said to be more consistent.

9.4.1 Standard Deviation for Individual Observations:

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$$

where

$\sigma =$ Standard deviation

d = Deviations from assumed mean (X-A)

N = Number of observations

Co-efficient of variation

$$C.V. = \frac{\sigma}{\bar{x}} \times 100$$

Ex. 7

Compute standard deviation and coefficient of variation for the following data:

x	d = x - A	d^2
10	10-20 = -10	100
15	15-20 = -5	25
18	18-20 = -2	4
21	21-20 = 1	1
25	25-20 = 5	25
20	20-20 = 0	0
12	12-20 = -8	64
N = 7	$\sum d = -3$	$\sum d^2 = 219$
$\sum x = 121$		

Let A = 20

$$\sigma = \sqrt{\frac{\sum d^2}{N} - \left(\frac{\sum d}{N}\right)^2}$$

$$= \sqrt{\frac{219}{7} - \left(\frac{-3}{7}\right)^2}$$

$$= \sqrt{31.286 - -0.429^{\ 2}}$$

$$= \sqrt{31.286 - 0.184}$$

$$= \sqrt{31.102} = 5.58$$

$$\boxed{\sigma = 5.58}$$

Co-efficient of variation $= \frac{\sigma}{\bar{x}} \times 100$

Let us first find arithmetic mean $\bar{x}$ for the given data.

$$\bar{x} = \frac{\sum x}{N} = \frac{121}{7} = 17.29$$

$$\boxed{\begin{array}{l}\bar{x} = 17.29 \\ \sigma = 5.58\end{array}}$$

By substituting the values

$$\text{c.v.} = \frac{5.58}{17.29} \times 100 = 32.27\%$$

9.4.2 Standard deviation and co-efficient of variation for discrete data:

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

Where
N – frequency
D – deviations from assumed mean

Formula for co-efficient of variation is the same as seen in the earlier section. i.e.

$$\text{c.v.} = \frac{\sigma}{\bar{x}} \times 100$$

Ex. 8
For the following data, compute standard deviation and co-efficient of variation.

Solution:

x	f	d = x - A	d^2	fd	fd^2	fx
5	14	-3	9	-42	126	70
6	40	-2	4	-80	160	240
7	54	-1	1	-54	54	378
8	46	0	0	0	0	368
9	26	1	1	26	26	234
10	12	2	4	24	48	120
11	6	3	9	18	54	66
12	2	4	16	8	32	24
	$\sum f = 200$			$\sum fd = -100$	$\sum fd^2 = 500$	$\sum fx = 1500$

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

$$= \sqrt{\frac{500}{200} - \left(\frac{-100}{200}\right)^2}$$

$$= \sqrt{2.5 - \ -0.05\ ^2}$$

$$= \sqrt{2.5 - 0.25} \ = \sqrt{2.25} \ = 1.5$$

$$\boxed{\sigma = 1.5}$$

For calculating co-efficient of variation, we need to compute arithmetic mean $\bar{x}$.

$$\bar{x} = \frac{\sum fx}{N} = \frac{1500}{200} = 7.5$$

$$\boxed{\bar{x} = 7.5}$$

By substituting the value

$$c.v. = \frac{\sigma}{\bar{x}} \times 100 \quad = \frac{1.5}{7.5} \times 100 = 20\%$$

$$\boxed{c.v = 20\%}$$

9.4.3 Calculation of standard deviation for continuous data:

For continuous series, standard deviation is calculated using step-deviation method. Accordingly, the formula is

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

Where $d = \left(\frac{x - A}{i}\right)$

i = class interval

The coefficient of variation is calculated by the same formula, as given earlier.

Ex. 9
Calculate standard deviation and co-efficient of variation for the following data.

Class Interval	Frequency	Mid point	$d = \frac{x - A}{i}$	d^2	fd	fd^2
5-10	6	7.5	-4	16	-24	96
10-15	5	12.5	-3	9	-15	45
15-20	15	17.5	-2	4	-30	60
20-25	10	22.5	-1	1	-10	10
25-30	3	27.5	0	0	0	0
30-35	4	32.5	1	1	4	4
35-40	3	37.5	2	4	6	12
40-45	2	42.5	3	9	6	18
	N = 50				-63	245

Let A = 27.5 and i = 5

$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} \times i$$

$$= \sqrt{\frac{245}{50} - \left(\frac{-63}{50}\right)^2} \times 5$$

$$= \sqrt{4.9 - \ -1.26\ ^2} \times 5$$

$$= \sqrt{4.9 - 1.5876} \times 5$$

$$= \sqrt{3.3124} \times 5 \quad = 1.82 \times 5 \ = 9.1$$

$$\boxed{\sigma = 9.1}$$

For co-efficient of variation, we need to find the value of arithmetic mean $\bar{x}$

$$\bar{x} = A + \frac{\sum fd}{N} \times c$$

$$= 27.5 + \frac{-63}{50} \times 5$$

$$= 27.5 + \ -1.26\ \times 5$$

$$= 27.5 + \ -6.3$$

$$= 27.5 - 6.3 \quad = 21.2$$

$$\boxed{\bar{x} = 21.2}$$

By substituting the values

$$c.v = \frac{\sigma}{\bar{x}} \times 100 = \frac{9.1}{21.2} \times 100 \ = 42.92\%$$

Check your progress

1. Calculate coefficient of variation for the following data.

Age	10-20	20-30	30-40	40-50	50-60	60-70
No. of Persons	10	14	12	9	8	7

2. Calculate standard deviation and coefficient of variation for the following

Life in Years	0-2	2-4	4-6	6-8	8-10	10-12
No. of Refrigerators	5	16	13	7	5	4

3. What is standard deviation and c.v. for the following data.

Height	150	151	152	154	156	158	160
Students	5	10	15	65	40	25	23

4. Calculate standard deviation and c.v. for the following data.

i) 40 42 45 47 50 51 54 55 57
ii) 82 56 75 70 52 80 68

5) Calculate standard deviation of salaries of the employees.

Salaries (in ,000)	45	50	55	60	65	70	75	80
Persons	3	5	8	7	9	7	4	7

9.5 SKEWNESS

Measures of central tendency and dispersion do not reveal all characteristics of the given data. They do no tell us whether distribution is symmetrical or not. Symmetrical distribution is that which, if plotted on a graph, gives a normal or ideal curve. Tow distributions may have same arithmetic mean or standard deviation. But still they may differ in their graphical representation as follows:

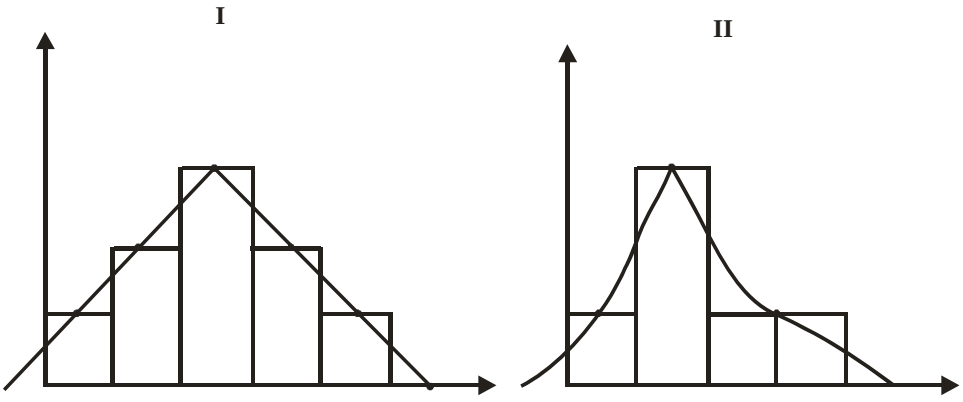


Figure 9.1

In both these data sets, mean and standard deviation are the same. But they are not alike in nature. The left hand side distribution is symmetrical and the right hand side distribution is asymmetrical in nature. Skewness is a measure of lack of symmetry. Measures of central tendency tell us about the central value of distribution, measures of dispersion tell us about the concentration of the value around the central value and measures of Skewness tell us whether the dispersal of the items from average is symmetrical or asymmetrical.

Following figures will explain the concept of Skewness more neatly. There are three major kinds of distribution.

1) Symmetrical
2) Positively Skewed
3) Negatively Skewed

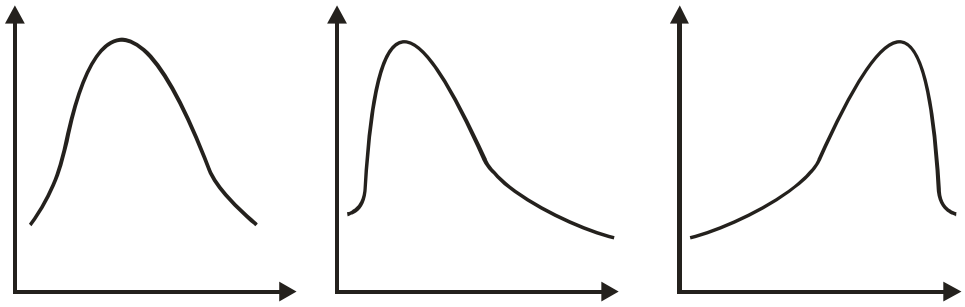


Fig. a Fig. b Fig. c

Figure 9.2

In figure (a), values of mean, median and mode will be the same. The spread of frequencies is exactly same on both the sides. Such a distribution is called symmetrical distribution.

In figure (b), value of mode is the least and that of the mean is highest. Median lies in between the two. Such a distribution is called positively Skewed distribution.

In figure (c), value of mode is highest, that of mean is the lowest and median lies in the middle of the two. Such a distribution is called negatively Skewed.

9.5.1 Measures of Skewness

The measures of Skewness tell us the extent and direction of asymmetry in a series of data.

1) Karl Pearson's measure of Skewness

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

Sk_p – Pearson's measure of Skewness

The value of Skewness by this method lies between +1 and -1. If the distribution is perfectly symmetric, mean and mode will be equal and so the numerator will be zero. Hence the value of Sk_p will also be zero indicating the absence of Skewness Higher the value of Sk_p, larger is the Skewness of distribution.

2) Bowley's Coefficient of Skewness Sk_B

$$Sk_B = \frac{Q_3 + Q_1 - 2 \text{ Median}}{Q_3 - Q_1}$$

Where

Q_1 - first quartile

Q_3 - third quartile

This measure is called quartile measure of Skewness and its value always lies between +1 and -1.

Note: Since we have already learnt the calculation of mean, median, quartiles and standard deviation, the students can compute measures of Skewness by using the formulae given above.

Check your progress

1) Calculate Pearson and Bowley's Co-efficient of Skewness.

x	12	17	22	27	32	37	42	47
f	28	42	54	108	129	61	45	33

2) Calculate Pearson and Bowley's co-efficient of Skewness.

x	0-10	10-20	20-30	30-40	40-50	50-60
f	5	8	12	18	7	5

9.6 SUMMARY

In this unit, we learnt two new steps in the analysis of data. How to measure the scatteredness of the data from its central value and how to measure Skewness of the data in comparison with the symmetrical distribution. All these measures are considered to be important in revealing the nature of data.

9.7 QUESTIONS

1) What is dispersion? Explain briefly the important measures of dispersion.
2) Write notes on the following:
   a. Range
   b. Quartile deviation
   c. Standard deviation and coefficient of variation
3) Discuss in detail the concept of Skewness.

10

Module 5
ADVANCED ANALYSIS OF DATA

Unit Structure:

10.0 OBJECTIVES

- To learn and understand the relationship between two variables.
- To explore the relationship between the variables through scatter diagram.
- To mathematically derive a coefficient of correlation by Karl Pearson.
- To understand and compute rank correlation by Spearman.
- To learn technique to estimate the value of one variable given the value of another through regression analysis.

10.1 INTRODUCTION

Most of the variables under study are correlated to each other. For example price and demand, income and consumption, interest rate and investment, use of fertilizers and productivity of agriculture, father's height and sons height, etc. The nature and degree of correlation may be different. That means, some variables may be very closely related, some may not be related so closely. The correlation analysis helps in measuring the kind of such relationships. Once a nature of relationship between two variables is measured, it is possible to estimate the value of one variable, given the value of the other. For example, once we know that a price and the supply of commodity are correlated, we can estimate the expected supply for a given price. Regression analysis helps in estimating the unknown value of one variable from the known value

of other variable. In this unit, we will learn the technique of correlation and regression.

10.2 CORRELATION ANALYSIS

Correlation is a statistical method which helps in analyzing the relationship between two or more variables. The study of correlation is useful due to following reasons:

1) Since most of the variables have some kind of relationship, quantification of it is necessary to learn more about them.
2) Correlation is a first step towards estimation or prediction of unknown values of the variables.
3) An understanding of the degree and nature of correlation between two or more variables helps in reducing uncertainties about the economic behaviour of important variables like price level and money supply, interest rate and investment, taxation and willingness to work, etc.

Correlation is classified into three ways:

1) Positive and Negative correlation (Depends upon the direction of change) : When both the variables change in the same direction, (i.e. they increase or decrease together) it is positive correlation. For example when price rises, supply also increases, when income falls, consumption also declines. When increase in one variable is accompanied by a fall in other, it is negative correlation. For example, increase in price leads to fall in demand, increase in interest rate is accompanied by a fall in investment.

2) Simple and Multiple correlation (Depends upon number of variables under study) : Simple correlation is the relationship between two variables like height and weight of a person, or wage rate and employment in the economy. Multiple correlation, on the other hand, examines relationship between three or more variables. For example a relationship between production of rice per acre, rainfall and use of fertilizers is multiple in nature.

3) Linear and non-linear (Depends on the ratio of change between two variables) : When a change in one variable is in constant ratio with a change in other, it is linear relationship. For example doubling the amount of fertilizers used exactly doubles the yield per acre, it is linear relationship. Non-linear relationship exists when a change in one variable is not in constant ratio with a change in other. In this case doubling the amount of fertilizers may not exactly double the output per acre.

10.3 METHODS OF STUDYING CORRELATION

Following important method of studying correlation between two variable will be discussed in this unit.

- Scatter diagram method.
- Karl Pearson's Coefficient of Correlation.
- Rank Correlation Coefficient.

10.3.1 Scatter diagram

It is the simplest method of studying correlation, by using graphical method. Under this method, a given data about two variables is plotted in terms of dots. By looking at the spread or scatter of these dots, a quick idea about the degree and nature of correlation between the two variables can be obtained. Greater the spread of the plotted points, lesser is an association between two variables. That is, if the two variable are closely related, the scatter of the points representing them will be less and vice versa. Following are different scatter diagrams explaining the correlation of different degrees and directions.

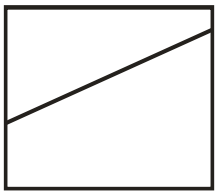


Fig. 1

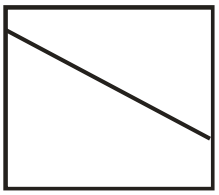


Fig. 2

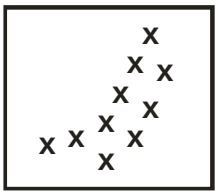


Fig. 3

Fig. 4

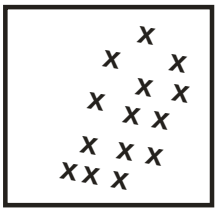


Fig. 5

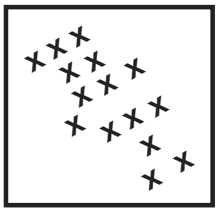


Fig. 6

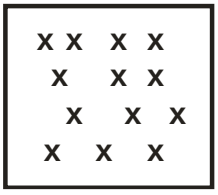


Fig. 7

Fig 10.1

1) Figure 1 represents positive perfect correlation where coefficient of correlation (r) = 1.
2) Figure 2 represents perfect negative correlation where coefficient of correlation (r) = -1
3) Figure 3 indicates high degree positive correlation where r = + 0.5 or more.
4) Figure 4 indicates high degree negative correlation where r = - 0.5 or more.
5) Figure 5 represents low degree positive correlation where the scatter of the points is more.
6) Figure 6 represents low degree negative correlation where the scatter for the points is more in negative direction.
7) Figure 7 indicates that there is no correlation between two variables. Here r = 0.

Thus, the closeness and direction of points representing the values of two variables determine the correlation between the same.

Advantages and Limitations of this method.

- It is a simple method giving very quick idea about the nature of correlation.
- It does not involve any mathematical calculations.
- It is not influenced by the extreme values of variables.
- This method, however, does not give exact value of coefficient of correlation and hence is less useful for further statistical treatment.

10.3.2 Karl Pearson's Coefficient of Correlation (r) :

This is the most widely used method of studying a bi-variate correlation. Under this method, value of r can be obtained by using any of the following three ways.

I) Direct Method of finding correlation coefficient

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

Where N = No. of observations

II) Taking deviations from actual mean

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Where $x = X - \overline{X}$, $y = Y - \overline{Y}$

III) Taking deviations from assumed mean

$$r = \frac{N\sum dx\,dy - \sum d_x \times \sum d_y}{\sqrt{N\sum d_x^2 - \left(\sum d_x\right)^2}\sqrt{N\sum d_y^2 - \left(\sum d_y\right)^2}}$$

Ex.1 Calculate Karl Pearson's coefficient of correlation using direct method.

X	Y	X^2	Y^2	XY
9	15	81	225	135
8	16	64	256	128
7	14	49	196	98
6	13	36	169	78
5	11	25	121	55
4	12	16	144	48
3	10	9	100	30
2	8	4	64	16
1	9	1	81	9
$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - \left(\sum X\right)^2}\sqrt{N\sum Y^2 - \left(\sum Y\right)^2}}$$ N – No. of observations

As per the table,

$N = 9, \sum X = 45, \sum Y = 108, \sum XY = 597, \sum X^2 = 285$ and $\sum Y^2 = 1356$

By substituting the values in the formula,

$$r = \frac{9 \times 597 - 45 \quad 108}{\sqrt{9 \times 285 - 45^2}\sqrt{9 \times 1356 - 108^2}}$$

$$= \frac{5373 - 4860}{\sqrt{2565 - 2025}\sqrt{12204 - 11664}}$$

$$= \frac{513}{\sqrt{540}\sqrt{540}} = \frac{513}{540} = +0.95$$

Since r = 0.95, there is high degree positive correlation between x and y.

Ex. 2 Calculate Karl Pearson's coefficient of correlation by taking deviations from actual mean.

x	y	$x = x - \bar{x}$	$y = y - \bar{y}$	xy	x^2	y^2
6	9	6-6 = 0	9-8 = 1	0	0	1
2	11	2-6 = -4	11-8 = 3	-12	16	9
10	5	10-6 = 4	5-8 = -3	-12	16	9
4	8	4-6 = -2	8-8 = 0	0	4	0
8	7	8-6 = 2	7-8 = -1	-2	4	1
$\sum x = 30$	$\sum y = 40$			$\sum xy = -26$	$\sum x^2 = 40$	$\sum y^2 = 20$

Formula for this method:

$$r = \frac{\sum xy}{\sqrt{\sum x^2 \times \sum y^2}}$$

Since $x = x - \bar{x}$ and $y = y - \bar{y}$, it is necessary to find arithmetic mean $\bar{x}$ and $\bar{y}$.

$$\bar{x} = \frac{\sum x}{N} \qquad \bar{y} = \frac{\sum y}{N}$$

$$= \frac{30}{5} \qquad = \frac{40}{5}$$

$$= 6 \qquad = 8$$

Where N – number of observations

From the table, it is clear that

$\sum xy = -26,\ \sum x^2 = 40,\ \sum y^2 = 20$

By substituting the values in the formula

$$r = \frac{-26}{\sqrt{40 \times 20}} = \frac{-26}{\sqrt{800}} = \frac{-26}{28 \cdot 28}$$

$$= -0.92$$

Since r = - 0.92, the correlation between x and y is high degree negative.

Ex.3 Compute Karl Pearson's coefficient of correlation by taking deviations from assumed mean.

(This method is used when the actual means are in fractions)

X	Y	$dx = X - A$	$dy = Y - A$	dx^2	dy^2	dxdy
2	25	2-9 = -7	25-29 = -4	49	16	28
5	27	5-9 = -4	27-29 = -2	16	4	8
7	26	7-9 = -2	26-29 = -3	4	9	6
9	29	9-9 = 0	29-29 = 0	0	0	0
19	34	19-9 = 10	34-29 = 5	100	25	50
16	39	16-9 = 7	39-29 = 10	49	100	70
		$\sum dx = 4$	$\sum dy = 6$	$\sum dx^2 = 218$	$\sum dy^2 = 154$	$\sum dxdy = 162$

For the above data, actual means $\bar{X}$ and $\bar{Y}$ will be in fraction. So we can take assumed means for both the variables and then find the deviations dx and dy .

Let assumed means for X = 9
Let assumed mean for Y = 29

$$r = \frac{N\sum dx\,dy - \sum dx \times \sum dy}{\sqrt{N\sum dx^2 - (\sum dx)^2}\sqrt{N\sum dy^2 - (\sum dy)^2}}$$

From the table,

$N = 6, \sum dx\,dy = 162, \ \sum dx = 4, \ \sum dy = 6, \sum dx^2 = 218, \sum dy^2 = 154$

By substituting these values in the formula,

$$r = \frac{6 \times 162 - 4 \times 6}{\sqrt{6 \times 218 - 4^2}\sqrt{6 \times 154 - 6^2}}$$

$$= \frac{972 - 36}{\sqrt{1308 - 16}\sqrt{924 - 36}}$$

$$= \frac{948}{\sqrt{1292}\sqrt{888}} = \frac{948}{\sqrt{1147296}} = \frac{948}{1071 \cdot 12}$$

$$= 0.89$$

Since r = 0.89, there is high degree positive correlation between X and Y.

Check your progress

1) Find correlation coefficient for the following data.

X	10	6	9	10	12	13	11	9
Y	9	4	6	9	11	13	8	4

Ans : $\boxed{r = 0.896}$

2)

X	45	70	65	30	90	40	50	75	85	60
Y	35	90	70	40	95	40	60	80	80	50

Ans: $\boxed{r = 0.903}$

3)

X	15	12	16	15	17	14	18
Y	17	14	20	25	20	24	22

Ans: $\boxed{r = 0.214}$

10.3.3 Rank Correlation:

For certain categories like beauty, honestry, etc quantitative measurement is not possible. Also sometimes the population under study may not be normally distributed. In such cases, instead of Karl Pearson's co-efficient of correlation, Spearman's Rank correlation coefficient is calculated. This method is used to the determine the level of agreement or disagreement between two judges. The calculations involved in this method are much simpler than the earlier method. Rank correlation is calculated using the following formula.

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

When D- difference between rank 1 and rank 2. N- No. of observations

Rank correlation is computed in following two ways:
1) When ranks are given.
2) When ranks are not given.

10.3.3.1 Rank correlation when ranks are given:

Ex.4 Following are the ranks given by two judges in a beauty contest. Find rank correlation coefficient.

Ranks by judge 1 R_1	Ranks by judge 2 R_2	$D = R_1 - R_2$	D^2
1	4	-3	9
2	5	-3	9
3	6	-3	9
4	7	-3	9
5	8	-3	9
6	2	4	16
7	3	4	16
8	1	7	49
$N = 8$			$\sum D^2 = 126$

$$R = 1 - \frac{6 \times \sum D^2}{N\ N^2 - 1}$$

By substituting the values from the table

$$= 1 - \frac{6 \times 126}{8\ 8^2 - 1}$$

$$= 1 - \frac{6 \times 126}{8 \times 63} = 1 - \frac{756}{504}$$

$$= 1 - 1.5 = -0.5$$

Since rank correlation co-efficient is -0.5, there is a moderate negative correlation between the ranking by two judges.

10.3.3.2 Calculation of rank correlation co-efficient, when the ranks are not given:

Ex.4 Calculate rank correlation for the following data.

X	Y	R_1 Rank to X	R_2 Rank to Y	$D = R_1 - R_2$	D^2
67	78	2	2	0	0
42	80	8	1	7	49
53	77	7	3	4	16
66	73	3	6	-3	9
62	75	4	4	0	0
60	68	5	7	-2	4
54	63	6	8	-2	4
68	74	1	5	-4	16
					$\sum D^2 = 98$

When the ranks are not given, we have to assign ranks to the given data. The ranks can be assigned in ascending (Rank 1 to the lowest value) or descending (Rank 1 to the highest value) order.

In this example, ranks are given in descending order.
The highest value gets rank 1 and so one.

$$R = 1 - \frac{6\sum D^2}{N\ N^2 - 1} = 1 - \frac{6 \times 98}{8 \times\ 8^2 - 1}$$

$$= 1 - \frac{6 \times 98}{8 \times 63} = 1 - \frac{588}{504} = 1 - 1$$

$$R = -0.167$$

Since rank correlation coefficient is -0.167, the relationship between X and Y is low degree negative.

Check your progress
Find rank correlation coefficient for the following data

1)

Rank by Judge A	7	6	5	8	3	1	2	4
Rank by Judge B	6	8	3	7	1	2	4	5

Ans: 0.76

2)

X	75	88	95	70	60	80	81	50
Y	120	134	150	115	110	140	142	100

Ans: 0.929

10.4 REGRESSION

10.4.1 Introduction:

Establishing relationship between two variables may be a first step towards statistical analysis. Once such a relationship is established, one may be interested in predicting the value of one variable given the value of other. Economists, businessmen policy makers or researchers are always interested in estimating profits, sales, demand, income, population, etc. Regression analysis helps in estimating or predicting the value of unknown variable with the help of known variable. The variable which is unknown or whose value is to be predicted is called dependent variable and the variable whose value is known is called independent variable.

10.4.2 Usefulness of Regression Analysis

1) With the help of regression, we can estimate the value of dependent variables from the values of independent variables. For example, if we know the price of commodity, we can estimate its demand.

2) A standard error can be calculated from the regression analysis which measures the scatter of observations around the regression lines.

3) Regression coefficients may be used for calculating co-efficient of correlation and coefficient of determination.

10.4.3 Regression lines and Regression Equations:

There are two regression lines with two variables X and Y.

i) Regression line of Y and X which is used to estimate the value of Y for any given value of X.

ii) Regression line of X on Y which is used to estimate the value of X for any given value of Y.

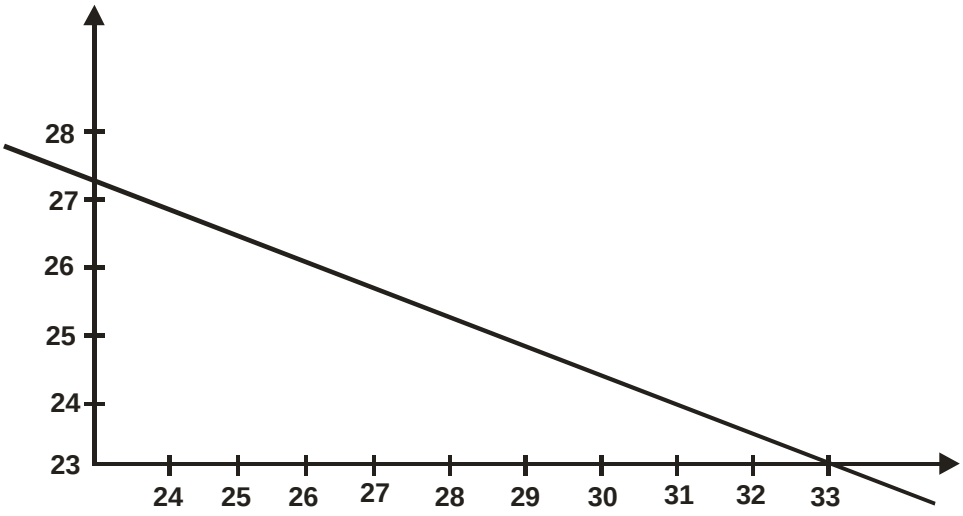


Fig 10.2

Regression equations are algebraic expressions of the regression lines. There are two regression equations.

1) Regression equation of Y on X – It is expressed as
   $Y_c = a + bX$ where Y – dependent variable
   X – independent variable
   a – y intercept, a point when regression line crosses y-axis
   b – is the slope of line.
   In the graph, a and b are shown which also are called constants.

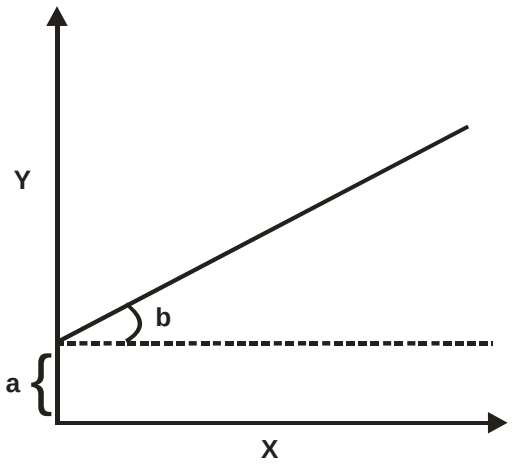


Fig 10.3

Regression analysis helps in determining the values of 'a' and 'b'. By solving following two equations simultaneously, one can determine the values of a and b.

For regression equation Y on X i.e. Y = a + bX.

$\sum Y = Na + b\sum X$ ---------- (1)

$\sum XY = a\sum X + b\sum X^2$ ---------- (2)

2) Regression equation X on Y. It is expressed as

$X_c = a' + b'Y$

To determine values of a' and b', following two equations should be solved simultaneously.

$\sum X = Na + b\sum Y$ ---------- (3)

$\sum XY = a\sum Y + b\sum Y^2$ ---------- (4)

This method is known as the least squares.

Ex.5 From the following data, obtain two regression equations, using the method of the least squares. Also estimate value of Y when X = 12.

X	Y	X^2	Y^2	XY
1	1	1	1	1
3	2	9	4	6
4	4	16	16	16
6	4	36	16	24
8	5	64	25	40
9	7	81	49	63
11	8	121	64	88
14	9	196	81	126
$\sum X = 56$	$\sum Y = 40$	$\sum X^2 = 524$	$\sum Y^2 = 256$	$\sum XY = 364$

Solution:

I) Regression equation of Y on X.

$Y_c = a + bX$

$\sum Y = Na + b\sum X$ ---------- (1)

$\sum XY = a\sum X + b\sum X^2$ ---------- (2)

From the table, N = 8, $\sum X = 56, \sum Y = 40, \sum XY = 364, \sum X^2 = 524$ and $\sum Y^2 = 256$

By substituting the values in (1) and (2)

$40 = 8a + 56b$ ----------(3)

$364 = 56a + 524b$ ---------- (4)

By solving (3) and (4) simultaneously (for this by multiplying equation (3) by 7 frist and then by subtracting (6) from (5)

$280 = 56a + 392b$ ----------(5)

$364 = 56a + 524b$ ----------(6)

$-84 = -132b$

$\therefore b = \frac{-84}{-132} = 0.636$

$\boxed{b = 0.636}$

By substituting the value of b in equation (3) we can calculate the value of a.

$40 = 8a + 56b$

$40 = 8a + 56\ (0.636)$

$40 = 8a + 35.616$

$40 - 35.616 = 8a$

$4.384 = 8a$

$\therefore a = \frac{4.384}{8} = 0.548$

$\boxed{a = 0.548}$

Regression equation of Y on X

$Y_c = a + bX$

$\boxed{Y_c = 0.548 + 0.636\,X}$

II) Regression equation of X on Y

$X_c = a' + b'Y$

$\sum X = Na' + b'\sum Y$ ---------- (1)

$\sum XY = a'\sum Y + b'\sum Y^2$ ----------(2)

By substituting the values from the table,

$56 = 8a' + 40b'$ ---------- (3)

$364 = 40a' + 256b'$ ----------(4)

By solving equations (3) and (4) simultaneously. [for this equation (3) is multiplied by 5 and then equation 6 is subtracted from equation (5)]

$$280 = 40a' + 200b' \quad \text{----------(5)}$$
$$364 = 40a' + 256b' \quad \text{----------(6)}$$
$$-84 = -56b'$$

$$\therefore\ b' = \frac{84}{56} = 1.5$$

$$\boxed{b' = 1.5}$$

By substituting the value of b' in equation (3)

$$56 = 8a' + 40\ 1.5$$
$$56 = 8a' + 60$$
$$56 - 60 = 8a'$$
$$-4 = 8a'$$
$$\therefore a = \frac{-4}{8} = -0.5$$
$$\boxed{a = -0.5}$$

So regression equation of X on Y

$$X_c = a' + b'X$$
$$\boxed{X_c = 0.5 + 1.5Y}$$

II) In order to estimate the value of variable Y when X = 12, we will have to use equation of Y on X.

$Y_c = 0.548 + 0.636X$

$Y_c = 0.548 + 0.636(12) \rightarrow$ by substituting X = 12

= 0.548 + 7.632

= 8.18 $\qquad \boxed{\therefore \text{ when } X = 12\ \ Y = 8.18}$

10.4.4 Calculation of Regression equations by taking deviations from arithmetic means of X and Y.

The method of finding regression equations as per the method discussed above is more complicated. By another method, the regression equations are written as follows:

I) Regression equation of X on Y.

$$X - \bar{X} = r\frac{\sigma x}{\sigma y}\ Y - \bar{Y}$$

Where $\bar{X}$ is mean of X-series.
$\bar{Y}$ is the mean of Y-series.

and

II) Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma y}{\sigma x}\ X - \bar{X}$$

σx - standard deviation of X
σy - standard deviation of Y

$r\frac{\sigma x}{\sigma y}$ is known as regression coefficient of X on Y.

$r\frac{\sigma y}{\sigma x}$ is known as regression coefficient of Y on X.

The regression coefficient of X on Y (bxy) can also be found as follows:

$$bxy = r\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y^2}$$

The regression coefficient of Y on X (byx) can also found as follow:

$$byx = r\frac{\sigma y}{\sigma x} = \frac{\sum xy}{\sum x^2}$$

It is should also be noted that the value of Pearson's correlation coefficient can be found using the following formula:

$$r = \sqrt{bxy \times byx}$$

Ex.6 From the following data, obtain
i) Regression equation of X on Y.
ii) Regression equation of Y on X.
iii) Correlation coefficient.
iv) Estimation of value of X when Y = 20.

X	Y	$x = X - \bar{X}$ (5)	$y = Y - \bar{Y}$ (12)	x^2	y^2	xy
1	9	-4	-3	16	9	12
2	8	-3	-4	9	16	12
3	10	-2	-2	4	4	4
4	12	-1	0	1	0	0
5	11	0	-1	0	1	0
6	13	1	+1	1	1	1
7	14	2	2	4	4	4
8	16	3	4	9	16	12
9	15	4	3	16	9	12
$\sum X = 45$	$\sum Y = 108$	$\sum x = 0$	$\sum Y = 0$	$\sum x^2 = 60$	$\sum y^2 = 60$	$\sum XY = 57$

I) Regression equation of X on Y

$$X - \bar{X} = r\frac{\sigma x}{\sigma y}\ Y - \bar{Y}$$

$$\bar{X} = \frac{\sum X}{N} = \frac{45}{9} = 5 \qquad \bar{Y} = \frac{\sum Y}{N} = \frac{108}{9} = 12$$

$$r\frac{\sigma x}{\sigma y} = \frac{\sum xy}{\sum y^2} = \frac{57}{60} = 0.95$$

By substituting the values in equation.

X - 5 = 0.95 (Y-12)

X - 5 = 0.95Y – 11.4

X = 0.95Y – 11.4 + 5

X = 0.95Y – 16.4

Regression equation of X on Y $\rightarrow$ $\boxed{X = -16.4 + 0.95Y}$

II) Regression equation of Y on X.

$$Y - \bar{Y} = r\frac{\sigma y}{\sigma x}\ X - \bar{X}$$

$$r\frac{\sigma y}{\sigma x}=\frac{\sum xy}{\sum x^2}=\frac{57}{60}=0.95$$

By substituting the values in equation

$$Y-12=0.95\ \ X-5$$

$Y - 12 = 0.95\ X - 4.75$
$Y = 0.95X - 4.75 + 12$
$Y = 0.95X + 7.25$

Regression equation of Y on X $\rightarrow$ $\boxed{Y=7.25+0.95X}$

III) Correlation coefficient

$r=\sqrt{bxy\times byx}$ $\quad bxy=0.95$

$r=\sqrt{0.95\times 0.95}$ $\quad byx=0.95$

$=0.95$ $\quad \boxed{r=0.95}$

IV) Value of X when Y = 20

X = -16.4 + 0.95 (20) $\rightarrow$ by substituting value of Y.
X = - 16.4 + 19
X = 2.6

Ex. 7 If you are given the following information, find
i) Two regression equations.
ii) Estimate value of X when Y = 75

	X	Y
Arithmetic Mean	36	85
Standard Deviation	11	8
Coefficient of Correlation Between X and Y	0.66	

i) Regression equation of X on Y.

$$X-\bar{X}=r\frac{\sigma x}{\sigma y}\ \ Y-\bar{Y}$$

By substituting the given values

$$X-36=0.66\left(\frac{11}{8}\right)\ \ Y-85$$

$$X-36=\ 0.66\times 1.375\quad Y-85$$

$$X-36=0.908\ \ Y-85$$

$X - 36 = 0.908Y - 77.18$
$X = 0.908Y - 77.18 + 36$
$X = 0.908Y - 41.18$

Regression equation of X on Y $\rightarrow$ $\boxed{X=-41.18+0.908Y}$

ii) Regression equation of Y on X

$$Y - \bar{Y} = r\frac{\sigma y}{\sigma x}\ X - \bar{X}$$

By substitution of the given values

$$Y - 85 = 0.66\frac{8}{11}\ X - 36$$

$$Y - 85 = 0.66 \times 0.73\ X - 36$$

$$Y - 85 = 0.48\ X - 36$$

$$Y - 85 = 0.48X - 17.28$$

$$Y = 0.48X - 17.28 + 85$$

$$Y = 0.48X + 67.72$$

Regression equation of Y on X $\rightarrow$ $\boxed{Y = 67.72 + 0.48X}$

Iii When Y = 75, value of X may be estimated as follows:

$$X = -41.18 + 0.908Y$$

$$= -41.18 + 0.908\ 75$$

$$= -41.18 + 68.1$$

$$= 26.92$$

Check your progress

1) By using the method of least squares, find out
   i) Regression equation of X on Y.
   ii) Regression equation of Y on X.
   iii) Estimate value of Y when X =10

X	1	2	3	4	5
Y	2	5	3	8	7

2) You are given the data of purchases and sales. Obtain two regression equations and estimate the likely sales when purchases equal to 100.

Purchases (X)	62	72	78	76	81	56	76	92	88	49
Sales (Y)	112	124	131	117	132	96	120	136	97	85

3) Given the following data, calculate the expected value of Y when X = 12

	X	Y
Average (Mean)	7.6	14.8
Standard Deviation	3.6	2.5
r	0.99	

4) Given the following data, estimate the marks in Maths obtained by a student who has scored 60 marks in English.

Mean marks in Maths	80
Mean Marks in English	50

S.D. of marks in Maths	15
S.D. of marks in English	10
Coefficient of correlation	0.4

5) Calculate regression equations from the following data

X	2	4	6	8	10	12	14
Y	5	2	5	10	4	11	12

10.5 SUMMARY

In this unit we have learnt correlation and regression analysis, which have a very important place in the statistical analysis of the data. Some important technique we learnt are:

a) Karl Pearson's correlation of coefficient.
b) Spearman's Rank correlation coefficient.
c) Regression equations by the method of least squares.
d) Regression equations by the method of deviations from arithmetic mean.
e) Regression coefficients.

All these concepts and calculations should be carefully studied. The students are required to refer to the books on statistics for variety of exercise.

10.6 QUESTIONS

1. Explain correlation and its classification.
2. Discuss the important methods of studying correlation between two variables.
3. Explain the meaning and usefulness of Regression analysis.

11

TIME SERIES ANALYSIS

Unit Structure:

11.0 OBJECTIVES

- To understand the meaning and importance of time series data.
- To estimate the trends in the data using three, four and five yearly moving averages.
- To learn how linear trend can be found out using regression analysis for time series data.

11.1 INTRODUCTION

Time series data is the data collected on the basis of some element of time. Wheat production for last 20 years, movement of share price of a company over last 15 days. Weekly sales of a manufacturing firm are some of the examples of time series data. In other words, a time series consists of statistical data which are collected, recorded and observed over successive increments of time. Economists and business men are generally interested in finding out future trends about say demand, for a commodity, incomes of consumer or likely demand for a particular service. Such an estimate is necessary for long-term planning.

11.1.1 Significance of time series analysis:

1) Analysis of time series helps in understanding the past trends in the behaviour of a particular variable. For example, trends in the production of wheat over last 20 years will be extremely helpful in predicting about future production of wheat.

2) Time series analysis helps in planning for the future programmes. The statistical techniques provided under time

series analysis help in predicting future values of a variable under consideration, on the basis of past data. For example the demand for residential needs in 2020 in Mumbai can be predicted from the past trends about the demand for housing in Mumbai.

3) Different time series can be compared and important observations can be made from such comparisons.

11.1.2 Components of Time Series:

There may be fluctuations in the time series data due to many reasons. These are called the components of time series. There are four components.

a) Secular Trend b) Seasonal Variations
c) Cyclical Variation d) Irregular Movements

a) Secular Trend: It explains the long term changes in the time series data. Generally populations, sales, production, income, consumption, etc have a rising trend whereas, the use of bullock carts, death rates and birth rates have declining trends. It implies that all the time series data have a general tendency of either a rise or fall over period of time. This tendency is known as a secular trend.

b) Seasonal Variations: These are the movements in time series data which occur regularly every year, due to seasonal variations or changes in climatic or weather conditions. For example the demand for umbrellas is very high during rainy season and low during the other seasons. Such variation in the demand for umbrellas is repeated every year. Similarly, demand for certain goods like candles, sweets, greeting cards, etc. also shows the seasonal variation.

c) Cyclical Variations: Generally, a business activity shows periodic variations. It means, according to the phase of business cycles, the performance of business activity varies. For example, during boom, there is a rapid business activity & during depression, the business activity suffers. All economic variables like sales, production, income, consumption, etc also vary in accordance with the phases of business cycle. Thus, cyclical variations in time series data are the long term (more than one year) movements in the data presenting periodic rise and decline in the variables under consideration.

d) Irregular Variations : The changes in the time series data, which can not be explained by any of the above factors, are called as irregular variations. These may occur due to natural calamities like flood, earthquakes, etc or unforeseen events like war, etc.

11.2 ESTIMATION OF TREND

As seen earlier, an important function of time series analysis is understand the trend and behaviour of the data so that one can plan in advance. For example, if we analyse the trends in the demand for residential areas over last 20 years, we may be able to plan in advance for the future to avoid mismatch between demand and supply for the same. There are many methods through which the trends can be estimated. We will learn two of these methods.

i) Method of Moving averages.
ii) Method of least squares.

11.2.1 Method of Moving Averages:

In order to estimate trend in a particular variable, it is necessary to smoothen out the variations in the time series data. It means, there may be fall or rise in the time-series data due to cyclical, seasonal or irregular variations. The method of moving averages is used to minimize the impact of these variations. There may be 3 yearly, 4 yearly or 5 yearly moving averages.

11.2.1.1 Three yearly moving average:

Example 1:

From the following time-series data related to the production of wheat calculate 3 yearly moving average.

Year	Wheat Prod. (in tones)	3 yearly totals	3 yearly averages
1991	(1) 18		
1992	(2) 21	→ 64 (1+2+3)	63/3 = 21.00
1993	(3) 25	→ 70 (2+3+4)	70/3 = 23.33
1994	(4) 24	→ 78 (3+4+5)	78/3 = 26.00
1995	(5) 29	→ 88 (4+5+6)	88/3 = 29.33
1996	(6) 35	→ 94 (5+6+7)	94/3 = 31.33
1997	(7) 30	→ 113 (6+7+8)	113/3 = 37.67
1998	(8) 48	→ 110 (7+8+9)	110/3 = 36.67
1999	(9) 32	→ 117 (8+9+10)	117/3 = 39.00
2000	(10) 37	→ 112 (9+10+11)	112/3 = 37.33
2001	(11) 43	→ 130 (10+11+12)	130/3 = 43.33
2002	(12) 50	→ 141 (11+12+13)	141/3 = 47.33
2003	(13) 48	→ 150 (12+13+14)	150/3 = 50.00
2004	(14) 52	→ 155 (13+14+15)	155/3 = 56.67
2005	(15) 55		

11.2.1.2 Five yearly moving average:

Example 2:

Compute five yearly moving average for the following data series.

Year	Investment ` crores	5 yearly moving totals	5 yearly moving averages
1985	(1) 38		
1986	(2) 49		
1987	(3) 56	→ 241 (1+2+3+4+5)	241/5 = 48.20
1988	(4) 40	→ 268 (2+3+4+5+6)	268/5 = 53.60
1989	(5) 58	→ 297 (3+4+5+6+7)	297/5 = 59.40
1990	(6) 65	→ 314 (4+5+6+7+8)	314/5 = 62.80
1991	(7) 78	→ 360 (5+6+7+8+9)	360/5 = 72.00
1992	(8) 73	→ 394 (6+7+8+9+10)	394/5 = 78.80
1993	(9) 86	→ 423 (7+8+9+10+11)	423/5 = 84.60
1994	(10) 92	→ 450 (8+9+10+11+12)	450/5 = 90.00
1995	(11) 94	→ 487 (9+10+11+12+13)	487/5 = 97.40
1996	(12) 105		
1997	(13) 110		

Example 3:

Compute 3 yearly and 5 yearly moving average for the following time series.

Year	1971	1972	1973	1974	1975	1976
Production	47	49	55	53	59	62
Year	1977	1978	1979	1980	1981	1982
Production	68	65	73	79	81	87

Computation of 3 yearly moving average

Year	Production	3 yearly moving totals	3 yearly moving averages
1971	47		
1972	49	→ 151	151/3 = 50.33
1973	55	→ 157	157/3 = 52.33
1974	53	→ 167	167/3 = 55.67
1975	59	→ 174	174/3 = 58.00
1976	62	→ 189	189/3 = 63.00
1977	68	→ 195	195/3 = 65.00
1978	65	→ 206	206/3 = 68.67
1979	73	→ 217	217/3 = 72.33
1980	79	→ 233	233/3 = 77.67
1981	81	→ 247	247/3 = 82.33
1982	87		

Computation of 5 yearly moving average

Year	Production	5 yearly moving totals	5 yearly moving averages
1971	47		
1972	49		
1973	55	→ 263	263/5 = 52.60
1974	53	→ 278	278/5 = 55.60
1975	59	→ 297	297/5 = 59.40
1976	62	→ 307	307/5 = 61.40
1977	68	→ 327	327/5 = 65.40
1978	65	→ 347	347/5 = 69.40
1979	73	→ 366	366/5 = 73.20
1980	79	→ 385	385/5 = 77.00
1981	81		
1982	87		

Check your progress

1) Calculate three yearly moving average.

Year	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
No. of students	33	31	35	39	40	41	42	40	38	39

2) Calculate five yearly moving average.

Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005

Exports	46	50	56	63	70	74	82	90	95	102

11.1.2.3 Four yearly moving average:

If moving average is to be calculated for the even number of years, a different procedure needs to be followed. 4 yearly moving totals will fall between the two time periods. As a result, the original time periods may not be maintained. To overcome this problem, a process of 'centering' needs to be followed. Following example explains the process of finding four yearly moving average.

Example 4:

Calculate 4 yearly moving averages for the following data set.

Steps to solve:

1) Get moving totals for 4 years in the rolling manner. (Eg. 1+2+3+4, 2+3+4+5, 3+4+5+6 etc)
2) Write the total between 2nd and 3rd number.
3) Calculate four yearly moving averages.
4) Take 4 yearly centred averages, by taking the average of two periods in the rolling manner.

Year (1)	Exports in ` Cr. (2)	4 Yearly moving totals (3)	5 Yearly moving averages (4) = (3)/4	4 yearly moving averages centred (5)
1991	17			
1992	22			
	→	94	94/4 = 23.50	
1993	25			25.38 $\left[\frac{23.50+27.25}{2}\right]$
	→	109	109/4 = 27.25	
1994	30			27.00 $\left[\frac{27.25+30.75}{2}\right]$
	→	123	30.75	
1995	32			32.38
	→	136	34.00	
1996	36			35.63
	→	149	37.25	
1997	38			39.75
	→	169	42.25	
1998	43			46.25
	→	201	50.25	
1999	52			
2000	68			

Check your progress

1) Calculate 4 yearly moving average for the following time series.

Year	1	2	3	4	5	6	7	8	9	10
Value	12	18	16	25	30	32	36	41	48	53

2) Compute 3 yearly and 4 yearly moving averages.

Year	1	2	3	4	5	6	7	8	9	10
Value	25	30	33	38	29	40	42	45	50	51

11.2.2 Method of least squares :

This is the most widely used method of finding out the trend values. Under this method. A straight line is fitted to the given data and with the help of the straight line, future trend can be predicted. This method is called as the method of least squares because a trend line fitted through this method satisfies two conditions.

1) $\sum Y - Y_c = 0$

Where Y – are actual values of variable

Y_c = are computed values of variable.

2) $\sum Y - Y_c^{\,2}$ is the least

Which means sum of the square of deviations of computed values from the actual value is the least.

A straight line trend is represented by the equation

$Y_c = a + bX$ where Y_c are the computed/trend values of variable under consideration. a & b are constants.

In order to obtain values of a and b, two equations should be solved.

$\sum Y = Na + b\sum X$ ----------(1)

$\sum XY = a\sum X + b\sum X^2$ ----------(2)

Where N is number of years in the time series.

Example 5:

Below are given the number of units produced in lakhs in a factory during 2000-2006.

i) Fit a straight line trend.
ii) Plot these figures on a graph and also show trend line.
iii) Estimate the number of units produced for the year 2008.

Year (1)	No. of units (Y) (2)	X (3)	XY (4)	X^2 (5)	Trend values Y_c (6)
2000	80	-3	-240	9	84
2001	90	-2	-180	4	86
2002	92	-1	-92	1	88
2003	83	0	0	0	90
2004	94	1	94	1	92
2005	99	2	198	4	94
2006	92	3	276	9	95
	$\sum Y = 630$	$\sum X = 0$	$\sum xy = 56$	$\sum X^2 = 28$	

Steps:

1) Middle year should be taken as zero or the year of origin. Accordingly, the years before the origin will have negative values & years after that will have positive value. This will make $\sum X = 0$ (This procedure is useful when there are odd number of years in given data)

2) Fit the equation of straight line $Y_c = a + bX$.

i) $Y_c = a + bX$

$\sum Y = Na + b\sum X$ ----------(1)

$\sum XY = a\sum X + b\sum X^2$ ----------(2)

Since $\sum X = 0$, equation (1) will be

$\sum Y = Na$

$\therefore a = \frac{\sum Y}{N}$

and equation (2) will be

$\sum XY = b\sum X^2$

$\therefore\ b = \frac{\sum XY}{\sum X^2}$

By substituting the values

$a = \frac{630}{7} = 90$ $\boxed{a = 90}$

$b = \frac{56}{28} = 2$ $\boxed{b = 2}$

By substituting the values of a and b in equation $Y_c = a + bX$, we get a straight lines as follows.

$Y_c = 90 + 2X$

with the help of this equation, we can get trend values for all years.

For 2000 - Y_c = 90 + 2 (-3) = 84
For 2001 - Y_c = 90 + 2 (-2) = 86
For 2002 - Y_c = 90 + 2 (-1) = 88
For 2003 - Y_c = 90 + 2 (0) = 90
For 2004 - Y_c = 90 + 2 (1) = 92
For 2005 - Y_c = 90 + 2 (2) = 94
For 2006 - Y_c = 90 + 2 (3) = 96

ii) The actual and trend values given in column (2) and (6) of the table can be represented graphically as follows.

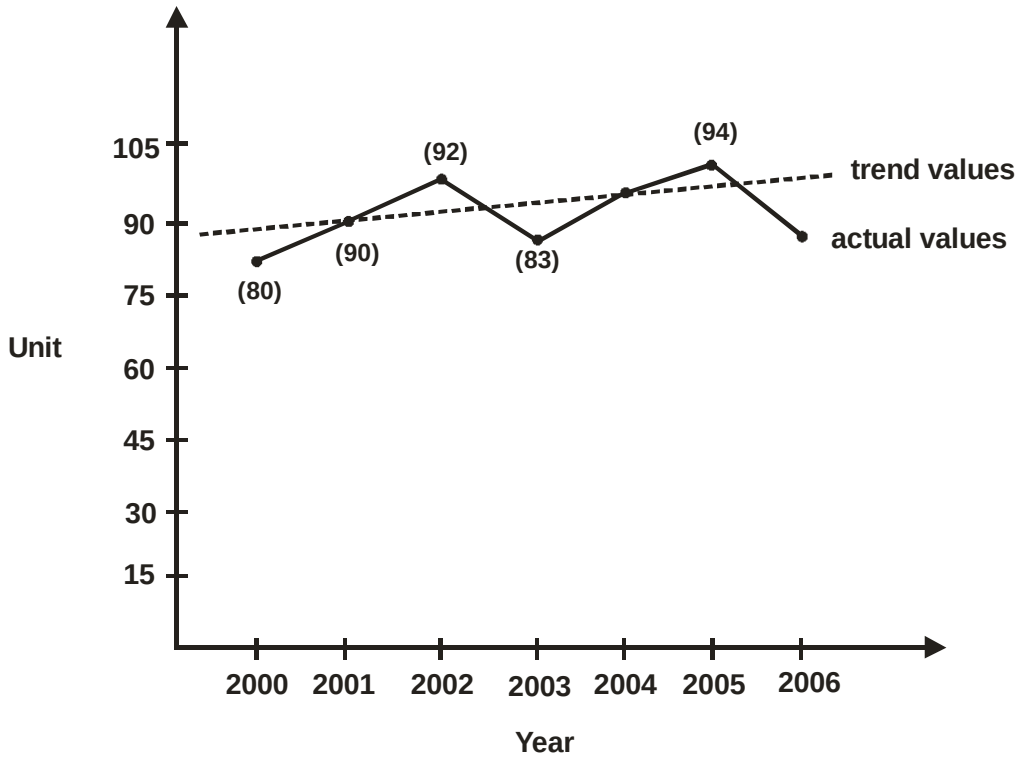


Fig 11.1

iii) For year 2008, the number of units produced will be estimated as follows

For 2008 $Y_c = 90 + 2\,(5) = 100$

Example 6:

Fit a straight line trend by the method of least square to the following data. What would be the predicted values for 2005?

Year	Earnings (Y)	X	XY	X^2	Trend values
1996	38	-7	-266	49	40.06
1997	40	-5	-200	25	47.40
1998	65	-3	-195	9	54.74
1999	72	-1	-72	1	62.08
2000	69	1	68	1	69.42
2001	60	3	180	9	76.76
2002	87	5	435	25	84.1
2003	95	7	665	49	91.44
	$\sum Y = 526$	$\sum X = 0$	$\sum xy = 616$	$\sum X^2 = 168$	

Important Note: When the number of years are even (n = even number), to convert years into X values, take the difference of two for every year. (For example, years 1999 and 2000, which are the mid-points, should be denoted number -1 and 1 respectively).

Solution:

$Y_c = a + bX$ ----------(1)

To find the value of a & b, two equations should be solved.

$\sum Y = Na + b\sum X$ ----------(2)

$\sum Y = Na$ Since $\sum X = 0$

$$\therefore a = \frac{\sum Y}{N}$$

$$\therefore a = \frac{526}{8} = 65.75$$

$$\boxed{\begin{array}{l} a = 65.75 \\ b = 3.67 \end{array}}$$

$\sum XY = a\sum X + b\sum X^2$

$\sum XY = b\sum X^2$ Since $\sum X = 0$

$$\therefore b = \frac{\sum XY}{\sum X^2} = \frac{616}{168} = 3.67$$

By substituting the values of a and b in equation (1)

$Y_c = 65.75 + 3.67\ X$

Using this equation, we can find trend values for all years.

For 1996 - Y_c = 65.75 + 3.67 (-7) = 40.06
For 1997 - Y_c = 65.75 + 3.67 (-5) = 47.40
For 1998 - Y_c = 65.75 + 3.67 (-3) = 54.74
For 1999 - Y_c = 65.75 + 3.67 (-1) = 62.08
For 2000 - Y_c = 65.75 + 3.67 (1) = 69.42
For 2001 - Y_c = 65.75 + 3.67 (3) = 76.76
For 2002 - Y_c = 65.75 + 3.67 (5) = 84.10
For 2003 - Y_c = 65.75 + 3.67 (7) = 91.44

For year 2005, the predicted value of earnings will be-

$Y_{2005} = 65.75 + 3.67\ (11) = 106.12$

(Please note, since the value of X for the year 2003 is 7, for 2004, it will be 9 and 2005 it will be 11)

Check your progress

1) Determine the equation of straight line with best fits the following data. Compute the trend values for all years. Also estimate exports for 2002.

Year	1994	1995	1996	1997	1998	1999	2000
Exports	80	90	92	83	94	99	102

2) Using the method of least squares, find the straight line trend for the following data. Also estimate values for 2005.

Year	1998	1999	2000	2001	2002	2003	2004
Production	52	53	42	60	65	67	69

3) Find trend line for the following data. Estimate the yield for 2008.

Year	2000	2001	2002	2003	2004	2005
Yield	73	72	78	86	90	93

11.3 SUMMARY

In this unit, we learnt the meaning and significance of time series data. Out of many ways to compute trends, we learnt two important techniques of estimation.

a) Method of moving averages

b) Method of least square

Time series data are very useful for economic analysis of the variables like income, exports, production, etc. Method of least squares is the most widely used method to compute linear trends and estimate the future trends in the variable under consideration.

11.4 QUESTIONS

1. Discuss the significance and components of Time Series Analysis.

2. Explain in detail the methods of Estimation of trend.

12

Module 6
INDEX NUMBERS

Unit Structure:

12.0 OBJECTIVES

- To understand meaning, nature and steps involved in calculating index numbers.
- Compute major types of weighted and un-weighted index numbers.
- Learn how the bases are shifted while constructing index numbers.
- Understand the concept of cost of living index numbers.
- Learn the problems involved in construction of index numbers.

12.1 INTRODUCTION

Index numbers are called as the barometers of economy as these are useful in understanding what is happening in the economy. For example, different indices help us understand whether industrial production has gone up or gone down in comparison with some earlier period. Whether economy is experiencing inflation or deflation, whether the cost of living has gone up and by what magnitude in comparison with the earlier period. Index number is a specialized average that provides a measurement of relative changes in price or quantity from time to time or place to place.

12.1.1 Uses of Index numbers:

Index numbers are useful for the economists, business men, policy-makers, researchers and also common people, in many ways.

1) The reveal trends and tendencies in the prices of group of commodities or production of industrial goods over time. So general business conditions can be very well studied through the technique of index numbers.

2) Many economic and business related policies are formulated on the basis of index numbers.
   For example, the dearness allowance (DA) is directly related to the cost of living index number. That means, the percentage of DA is determined on the basis of rise and fall in the index number of consumer goods. Wages and salaries are adjusted according to index numbers.

3) In order to understand the trend in time series data, it is necessary to adjust the data for price changes. This process is called deflating the data. For this, index numbers are used.

4) Index numbers are also useful in forecasting future economic activity.

12.1.2 Types of Index numbers:

There are three types of index numbers:

1) Price Index Numbers: These are useful in comparing the prices of a basket of goods and service in the current year to the prices of the same basket in the reference / base year.

2) Quantity Index Numbers: They measure the quantity produced, sold or consumed in a particular year to that in a base year.

3) Value Index Numbers: These compare total value of a group of commodities to that of the base year.

12.1.3 Problems in the construction of Index Numbers:

Construction of appropriate index numbers is not an easy task. Many problems are involved while preparing index numbers. These are as follows:

1) Purpose of the index: There are many ways to construct index. At the outset, it is important to define purpose of index number or the reason for index number. This will help in selecting the

commodities in a proper way. For example, if the cost of living index number is to be constructed, then it is necessary that the commodities of mass consumption should be included in the calculations otherwise, no fruitful outcomes would be possible.

2) Selection of base year: While constructing index number, a comparison is made with the base year. This year should be selected with a lot of care. Index number for the base year is always 100. A care should be taken that the base year is a normal year and is not a year with natural calamities or any other abnormal situations, the base year should not be too distinct in the past. Comparison should be with more recent period.

3) Selection of commodities: Appropriate commodities should be included in index numbers, depending upon the purpose of index number. For example, in case of construction of index numbers for the workers. Commodities like cars, air conditioners should not be included. In case of general price index, all the commodities of mass consumption should be considered.

4) Prices: Another very serious problem that has to be tackled before constructing index numbers, is to take appropriate prices for the commodities. The prices of commodities vary from place to place and from shop to shop. So it is necessary to obtain some representative price for the purpose of index numbers.

5) Selection of formula: There are different methods of constructing index numbers. Each of these methods, has merits and demerits. It is important to do a proper selection of formula for preparing index numbers.

12.2 METHODS OF CONSTRUCTING INDEX NUMBERS

Index numbers can be:

1) Unweighted
2) Weighted

The unweighted index numbers we are going to learn in this unit are:

a) Method of Simple Aggregative
b) Method of Price Relative

The weighted index numbers we are going to learn are:

a) Laspeyre's Method
b) Paasche's Method

c) Fisher's Method

12.2.1 Unweighted Index Numbers:

In this kind of index numbers, no weights are assigned to the commodities. That means, each commodity is considered to be of equal importance for the consumer.

1. Simple Aggregative Method:

This is the simplest method of constructing index number and it uses following formula:

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

Where 1 – current year
0 – base year
P – price

Ex.1

Commodity	Price in 2008 (P_0)	Price in 2010 (P_1)
A	10	15
B	25	31
C	18	20
D	7	9
E	12	14
	$\sum P_0 = 72$	$\sum P_1 = 89$

$$P_{01} = \frac{\sum P_1}{\sum P_0} \times 100$$

$$= \frac{89}{72} \times 100$$

= 123.61

The base year index number is always 100. Since the current year index is 143.55, there has been 23.61% increase in the price level.

2. Method of Price Relative:

In this method, the price-relatives are obtained for each commodity and then following formula is applied.

$$P_{01} = \frac{\sum\left(\frac{P_1}{P_0} \times 100\right)}{N}$$ Where N - number of commodities

Ex.2

Construct index number using the method of price relative.

Commodity	Price in 2000 P_0	Price in 2005 P_1	Price relative $\frac{P_1}{P_0} \times 100$
A	10	15	$\frac{15}{10} \times 100 = 150.00$
B	25	31	$\frac{31}{25} \times 100 = 124.00$
C	18	20	$\frac{20}{18} \times 100 = 111.11$
D	7	9	$\frac{9}{7} \times 100 = 128.57$
E	12	14	$\frac{14}{12} \times 100 = 116.67$
			$\sum \frac{P_1}{P_0} = 630.35$

$$P_{01} = \frac{\sum\left(\frac{P_1}{P_0} \times 100\right)}{N} = \frac{630.35}{5} = 126.07$$

There is 26.07% increase in the price level.

12.2.2 Weighted Index Numbers:

As seen earlier, the unweighted index numbers, assign equal importance to all the items included in the index. But it is more realistic to assign weights to the commodities on the basis of their importance in the consumption basket of a consumer. There are three such methods that assign weight to the commodities.

These are:

1) Laspeyre's Index Number:

Where P – price
q – quantity
0 – base year
1 – current year

$$P_{01} = \frac{\sum P_1 q_0}{\sum P_0 q_0} \times 100$$

2) Paasche's Method

$$P_{01} = \frac{\sum p_1 q_1}{p_0 q_1} \times 100$$

3) Fisher's Method

$$p_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

Ex. 3

Construct index number with the help of following information, by using –

a) Laspeyre's Method

b) Paasche's Method

c) Fisher's Method

Commodity	2001		2008		p_1q_0	p_0q_0	p_1q_1	p_0q_1
	p_0	q_0	p_1	q_1				
A	7	12	9	15	108	84	135	105
B	5	10	3	15	30	50	45	75
C	12	5	15	4	75	60	60	48
D	10	6	12	7	72	60	84	70
					$\sum p_1q_0$	$\sum p_0q_0$	$\sum p_1q_1$	$\sum p_0q_1$
					= 285	= 254	= 324	= 298

From the given data, we have to first get the values of p_1q_0 i.e. price in the current year multiplied by quantity in the base year for each commodity, p_0q_0 price & quantity in base year for each commodity and so on. All these values are required for using different formulae.

a) Laspeyre's Index Number:

$$p_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{285}{254} \times 100 = 1.1220 \times 100 = 112.20$$

b) Paasche's Index Number:

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{324}{298} \times 100 = 1.0872 \times 100 = 108.72$$

c) Fisher's Index Number:

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{285}{254} \times \frac{324}{298}} \times 100$$

$$= \sqrt{1.1222 \times 1.0872 \times 100}$$

$$= \sqrt{1.2201} \times 100$$

$$= 1.1046 \times 100$$

= 110.46

Ex. 4

Construct the index number by

i) Laspeyre's Method

ii) Paasche's Index Number

iii) Fisher's Index Number

Commodity	Base Year		Current Year		p_1q_0	p_0q_0	p_1q_1	p_0q_1
	Price p_0	Quantity q_0	Price p_1	Quantity q_1				
M	6	50	10	56	500	300	560	336
N	2	100	2	120	200	200	240	240
Q	4	60	6	60	360	240	360	240
R	10	30	12	24	360	300	288	240
S	8	40	12	36	480	320	432	288
					1900	1360	1880	1344

a) Laspeyres's Index Number:

$$P_{01} = \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$$

$$= \frac{1900}{1360} \times 100 = 1.3971 \times 100 = 139.71$$

b) Paasche's Index Number

$$P_{01} = \frac{\sum p_1 q_1}{\sum p_0 q_1} \times 100$$

$$= \frac{1880}{1344} \times 100 = 1.3988 \times 100 = 139.88$$

c) Fisher's Index Number

$$P_{01} = \sqrt{\frac{\sum p_1 q_0}{\sum p_0 q_0} \times \frac{\sum p_1 q_1}{\sum p_0 q_1}} \times 100$$

$$= \sqrt{\frac{1900}{1360} \times \frac{1880}{1344}} \times 100$$

$$= \sqrt{1.3971 \times 1.3988} \times 100$$

$$= \sqrt{1.9543} \times 100$$

$$= 1.3980 \times 100 = 139.80$$

12.3 DEFLATOR

Deflating means making allowances or adjustments for the effects of changes in the price-level. An increase in price level reduces the value or purchasing power of money. This means that to maintain the same standard of living or to purchase same amount of goods, more money is required than before. So with rising prices, there is a difference between money wages and real wages. The salary earners or workers are more interested in knowing what their incomes will buy than what income they are earning. The purchasing capacity of wages is real wages. And for calculating real wages, deflators are used.

$$\text{Real wages} = \frac{\text{Money wage}}{\text{Price index}} \times 100$$

Ex. 5

Following table gives annual income of a person and the price index for 7 years. Prepare the index number to show the changes in the real income of the person.

Year	Income	Price Index Number	Real Income	Real Index Number
2001	15,000	100	$\frac{15,000}{100} \times 100 = 15000$	100.00
2002	18,000	104	$\frac{18,000}{104} \times 100 = 17307.70$	115.38
2003	24,000	115	$\frac{24,000}{115} \times = 100 = 20869.57$	139.13
2004	25,000	160	$\frac{25,000}{160} \times 100 = 15625$	104.17
2005	28,000	280	$\frac{28,000}{280} \times 100 = 10000$	66.67

We can obtain the last column of Real Index number in the following way:

2001 $\frac{15000}{15000} \times 100 = 100$

2002 $\frac{17307.70}{15000} \times 100 = 115.38$

2003 $\frac{20869.57}{15000} \times 100 = 139.13$

12.4 BASE SHIFTING

Shifting the base means changing the reference year or the base year to some other year. Base shifting needs to be done when the previous base has become too old and is not much useful for making comparisons for the current year. In case of base shifting, all index numbers for the previous period (which were based on the old base year) should be divided by the index numbers corresponding to the new base period.

Ex. 7

Following are the index numbers of prices with 2002 as a base. Shift the base from 2002 to 2009 and rewrite new index numbers.

For shifting the base, formula is

$$\text{Index No. of 2002} = \frac{\text{Index No. of 2002}}{\text{Index No. of 2009}} \times 100$$

$$\text{Index No. of 2003} = \frac{\text{Index No. of 2003}}{\text{Index No. of 2009}} \times 100$$

$$\text{Index No. of 2004} = \frac{\text{Index No. of 2004}}{\text{Index No. of 2009}} \times 100$$

and so on.

Year	Index No. 2002 =100	Index No. with 2009 = 100 2009 as base
2002	100	$\frac{100}{380} \times 100 = 26.32$
2003	110	$\frac{110}{380} \times 100 = 28.95$
2004	120	$\frac{120}{380} \times 100 = 31.58$
2005	200	$\frac{200}{380} \times 100 = 52.63$
2006	400	$\frac{400}{380} \times 100 = 105.63$
2007	410	$\frac{410}{380} \times 100 = 107.89$
2008	400	$\frac{400}{380} \times 100 = 105.26$
2009	380	$\frac{380}{380} \times 100 = 100.00$
2010	370	$\frac{370}{380} \times 100 = 97.37$
2011	340	$\frac{340}{380} \times 100 = 89.47$

12.5 COST OF LIVING INDEX NUMBER

These index numbers are also known as consumer price index numbers. These include the commodities and the change in their prices, which are consumed by the masses. The consumption pattern of different income groups is different. The consumption habits of the poor, middle class and the rich people may not be the same. So the cost of living index numbers are prepared to study the effects of change in prices on the cost of living of a particular class.

Certain precautions should be taken while constructing these index numbers.

1) The class of people for whom the index is being prepared should be properly defined.

2) Geographical area of study should be decided.

3) Commodities selected for index should be grouped into food, clothing, fuel, house rent and others.

There are two methods of constructing cost of living index numbers.

1) Family Budget Method:

Consumer Price Index $= \frac{\sum pv}{\sum v}$

Where $p = \frac{p_1}{p_0} \times 100$ for each commodity

V = Value of weights or $p_0 q_0$

2) Aggregate expenditure method:

Consumer Price Index $= \frac{\sum p_1 q_0}{\sum p_0 q_0} \times 100$

This method is most popular and it is the laspeyres's method discussed earlier.

Ex. 8

Construct the cost of living index Numbers by using
1) Family Budget Method
2) Aggregate expenditure Method

Commodity	Quality in 2003	Price in 2003	Price in 2007
A	6	5	6
B	7	8	9
C	2	5	6
D	3	6	7
E	5	2	3

Solution:

Family Budget Method

Commodity	q_0	p_0	p_1	p $\left(\frac{p_1}{p_0}\times 100\right)$	v $p_0 q_0$	pv
A	6	5	6	$\frac{6}{5}\times 100 = 120.0$	30	3600.00
B	7	8	9	$\frac{9}{8}\times 100 = 112.5$	56	6300.00
C	2	5	6	$\frac{6}{5}\times 100 = 120.0$	10	1200.00
D	3	6	7	$\frac{7}{6}\times 100 = 116.67$	18	2100.06
E	5	2	3	$\frac{3}{2}\times 100 = 750.0$	10	1500.000
					124	14700.06

Cost of Living = $\frac{\sum pv}{\sum v}$

Index Number

$$= \frac{14700.06}{124}$$

= 118.55

So there has been 18.55% rise in the cost of living over the reference period.

Solution:

By aggregate expenditure method, we have already learnt the Laspeyres method of constructing index number. The students are to apply the same method here, to construct cost of living index number by aggregate expenditure method.

Check your progress

1) Calculate index numbers by the simple aggregative method and the method of price relatives (using arithmetic mean), from the following:

Commodity	Base price	Current price
Rice	35	42
Wheat	30	35
Pulse	40	38
Fish	107	120

2) Calculate consumer price index numbers from the following data, using the family budget formula:

Price (Rs.) per unit

Commodity	Base period	Current period	Weight
M	80	110	14
N	10	15	20
Q	40	56	35
R	50	95	15
S	12	18	16

3) Given below are the data on prices of some consumer goods and the weights attached to the various items compute price index number for the year using family budget method.

Price (Rs.)

Item	1984	1985	Weight
Wheat	0.50	0.75	2

Milk	0.60	0.75	5
Egg	2.00	2.40	4
Sugar	1.80	2.10	8
Shoes	8.00	10.00	1

4) Calculate the Fisher's index number using the following data.

Commodity	P_0	Q_0	P_1	Q_1
X	10	40	15	60
Y	15	80	20	100
Z	20	20	25	40

5) Calculate the Fisher's Index number for the following:

Commodity	P_0	Q_0	P_1	Q_1
A	10	5	15	5
B	5	10	5	12
C	8	4	10	5
D	12	5	15	5
E	6	15	12	10

6) For the following data, calculate the price index number by the method of simple average of price relatives.

Commodity	Price in year 1	Price in year 2
Bread	10	14
Milk	15	20
Eggs	10	16

7) For the data, calculate the Laspeyre's, Paasche's and Fisher's index number.

Commodity	Price in 1979	Quantity in 1979	Price in 1980	Quantity in 1980

X	10	40	15	60
Y	15	80	20	100
Z	20	20	25	40

12.6 SUMMARY

Technique of construction of index numbers is very useful in statistical analysis. It is used in many types of calculations such as national income accounts, calculations of dearness allowances for the workers and salaried people, industrial production prices, etc. We have learnt various formulae related to index numbers in this unit.

12.7 QUESTIONS

1. What is Index Number? Discuss the Uses and problems in the construction of Index Number.
2. Explain the method of Unweighted Index Numbers.
3. Explain the different methods of Weighted Index Numbers.
4. What is Deflator?
5. Explain Cost of living Index Number.

13

Module 7
HYPOTHESIS: NATURE AND ROLE IN RESEARCH

Unit Structure:

13.0 OBJECTIVES

- To know meaning, role and types of hypothesis.
- To acquaint with the sources of hypothesis.
- To understand characteristics of a good hypothesis.
- To know basic concepts in hypothesis testing.

13.1 MEANING OF HYPOTHESIS

Hypothesis is usually considered as the principal instrument in research. Once the research problem is decided, researcher proceeds to formulate tentative solutions or answers to the

problem. These proposed solutions or explanations constitute the hypothesis, i.e. hypotheses are tentative propositions relating to investigative questions. They are formulated so that a researcher can test them on the basis of the facts already known to him or which he collects in the course of his investigation. Hypothesis provides him a direction for investigation of research problem. Hypotheses aim at answering research questions and guide the researchers to see and select the relevant fact.

1. Definitions:

(1) Lundberg: Hypothesis as a tentative generalization, the validity of which remains to be tested.

(2) Goode and Hatt: Hypothesis as a proposition which can be put to a test to determine its validity.

(3) Webster's New International Dictionary of English Language: The word hypothesis is a proposition, condition or principal, which is assumed, perhaps without belief, in order to draw out its logical consequences and by this method to test its accord with facts which are known or may be determined.

(4) Cohen and Nagel: Formulation of prepositions are the hypotheses.

(5) Coffey: A hypothesis is an attempt at explanation; a provisional supposition made in order to explain scientifically some facts or phenomenon.

13.2 ROLE OF HYPOTHESIS

In social science research, hypothesis serves several important functions such as:

1. A hypothesis guides the direction of study or investigation. It states what we are looking for.
2. Its purpose is to include in the investigation all available and pertinent data either to prove or disprove the hypothesis.
3. Research becomes unfocussed or random without a hypothesis and useless data may be collected in the hope that important data is not omitted.
4. Thus, a hypothesis specifies the sources of data, which shall be studied and in what context they shall be studied.
5. It also determines the data needs and prevents a blind search.
6. A hypothesis can suggest the type of research which is likely to be appropriate to study a given problem.
7. It determines the most appropriate technique of data analysis.

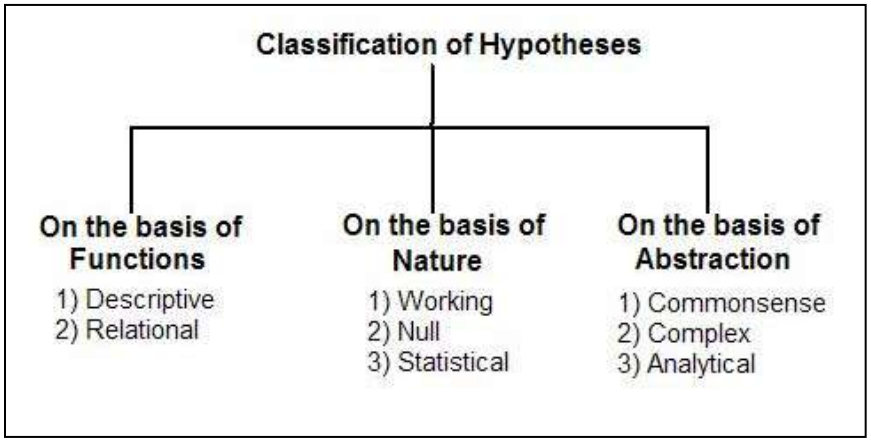
8. A hypothesis can contribute to the development of theory by testing various hypotheses relating to a stated theory. It is also likely, in some cases that a hypothesis helps in constructing a theory.

Check your progress:

1) What do you mean by hypothesis?
2) How does hypothesis help researchers?
3) Prepare a note on role of hypothesis.

13.3 TYPES OF HYPOTHESIS

Researchers in social sciences have to work with many kinds of hypotheses. Hence, they can be classified in several ways:



13.3.1 On the basis of their functions:

1. **Descriptive hypotheses:** These are propositions which describe the characteristics of a variable such as income, expenditure, production, hours of work, etc.

2. **Relational hypothesis:** These are propositions which describe the relationship between two variables. This type of hypotheses state that something is greater or less than

something. For example, increase in income tends to increase in expenditure; families with small income spend large proportion of their income on necessities, etc.

13.3.2 On the basis of their nature:

3. W**orking hypothesis:** Hypotheses are formulated while planning the study of a problem. They may not be specific in the initial stages. In such cases, they are called working hypotheses. Such hypotheses are subject to change or modification in the course of investigation.

4. **Null hypotheses:** These are hypothetical statements and they deny what is stated in working hypotheses. They do not ever expect to exist in reality. For example, statement like 'education does not increase the earning capacity of an individual' is a null hypothesis.

5. **Statistical hypotheses:** These are statements about a statistical population. The statements are derived from a sample drawn from a given population. Statistical hypotheses are quantitative in nature as they are numerically measurable. These hypotheses can be hypotheses of difference or association i.e. we can formulate them as null hypotheses or casual hypotheses.

13.3.3 On the basis of their level of abstraction: (Goode and Hatt):

6. C**ommonsense hypotheses:** At the lowest level we have simple description which gives rise to commonsense hypotheses. They state existence of certain empirical uniformities and hence expect verification of commonsense propositions.

7. **Complex hypotheses:** At a relatively higher level of abstraction, we have logical derivations which give rise to complex hypotheses. These aim at testing the existence of logically derived relationships between empirical uniformities. They are purposeful distortions of empirical reality. Hence, they are also called 'Ideal types'. Such hypotheses try to create tools and problems for further research in complex areas of investigations.

8. **Analytical hypotheses:** The category of hypothesis at the highest level of abstraction is concerned with the relation of analytic variables. Hence, they are called analytical hypotheses. These are statements about how changes in one property will affect another property. For example, statements about relation between level of education and

migration, level of income and social mobility are some such abstractions.

Check your progress:

1. How do you classify the types of hypothesis?
2. Prepare a chart showing classification of hypotheses.

_____

_____

_____

_____

_____

13.4 SOURCES OF HYPOTHESIS

Hypotheses may be developed from a variety of sources. Some of them are as follows:

9. **Observation:** Hypotheses can be derived from observation. Relation between production, cost and output of goods or relationship between price variation and demand are hypothesized from observation.

10. **Culture:** A very important and major source of hypotheses is the culture in which a researcher has grown. Hypotheses regarding relationship between caste and family size, income level and education level depend on the socio-economic background.

11. **Analogies:** They are often a source of meaningful hypotheses. For example, the hypotheses that similar human types or activities may be found occupying the same territory has come from plant ecology. Analogy is very suggestive. But one has to be careful in adopting models from other disciplines. Economic theory has adopted a few models from physics also.

12. **Theory:** Theory is an extremely fertile seed bed of hypotheses. A theory represents what is a known and logical deduction from the theory lead to new hypotheses, which must be true if the theory is true. For example, various hypotheses are derived from the theory based on profit maximization as the aim of a private enterprise. New

hypotheses may be derived from the established theory by method of logical induction or logical deduction.

13. **Findings of other studies:** Hypotheses may also be developed from the findings of other studies. This can happen when a study is repeated under different circumstances or different type of population. The findings of an exploratory study may be formulated as hypotheses for other structured studies which aim at testing a hypothesis for other structured studies which aim at testing a hypothesis. For example, the concept of trickle down effect of economic growth, later on becomes a testable hypothesis.

14. **Level of knowledge:** An important source of hypotheses is the state of knowledge of any particular science. Hypotheses can be deduced from existing formal theories. If the hypotheses are rejected, the theory can be modified. If formal theories do not exist, hypotheses are generated from formal conceptual framework. This leads to the growth of theory.

15. **Continuity of research:** Continuous research in a field is itself an important source of hypotheses. The rejection of hypotheses leads to the formulation of new ones. These new hypotheses explain the relationship between variables in the subsequent studies on the same subject.

In short, an ideal source of fruitful and relevant hypotheses is a fusion of two elements (i) past experience and (ii) imagination in the disciplined mind of the scientist.

Check Your Progress:

1. List down the sources of hypotheses.

_____

_____

_____

_____

_____

_____

13.5 CHARACTERISTICS OF A GOOD HYPOTHESIS

An acceptable or a useable good hypothesis should fulfill following certain conditions:

16. **Testability:** A hypothesis should be empirically testable and should not be just a moral judgement i.e. concepts included in the hypothesis must have empirical correspondence. In other words, it should be possible to collect empirical evidence to test the hypothesis.

17. **Conceptual clarity:** A hypothesis must be conceptually clear. The concepts used in the hypothesis should be clearly defined, not only formally but also operationally. For example, in a hypothesis, if concept of unemployment or educated unemployment is used, the terms must be clearly defined.

18. **Specificity:** A hypothesis should be specific and must explain the expected relation between variables. Broad, generalized statement does not form a specific hypothesis. Specificity ensures that the research is practicable and significant. It also helps in increasing the validity of results as the predictions are specific.

19. **Observable:** Hypothesis must be formulated in such a manner that deductions can be made from it and consequently a decision can be reached as to whether it does or does not explain the facts.

20. **Consistency:** Hypothesis should be logically consistent i.e. two are more propositions logically derived from the same theory should not be contradictory.

21. **Objectivity:** Hypothesis should be free from researcher's own value judgements.

22. **Simplicity:** Hypothesis should be simple and should involve fewer assumptions. However, simplicity demands insight and it does not mean that it is obvious.

23. **Theoretical relevance:** A hypotheses should be related to a body of theory. Thus, it can help to quality, support or refute theory. It can then become a new leap into new areas of knowledge. Theory not only formulates what we know, but also tell us what we want to know. If hypothesis is based on theory, it would have power of prediction.

24. **Availability of techniques:** Hypotheses should be related to available techniques; otherwise it will not be researchable.

The research cannot formulate useable questions, if the does not know about the available techniques or the required techniques are not available at all.

Check your progress:

1. List down the points showing characteristics of a good hypothesis?
2. What are characteristics of a good hypothesis?

_____

_____

_____

_____

13.6 BASIC CONCEPTS IN HYPOTHESIS TESTING

Basic concepts in the context of testing of hypotheses need to be explained. Those are:

13.6.1 Null and Alternative hypotheses:

In the context of statistical analysis, we often talk about null hypothesis and alternative hypothesis. If we are to compare method ***A*** with method ***B*** about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the *null hypothesis*. As against this, we may think that the method ***A*** is superior or the method ***B*** is inferior, we are then stating what is termed as *alternative hypothesis*. The null hypothesis is generally symbolized as H_0 and the alternative hypothesis as H_a. Suppose we want to test the hypothesis that the population mean (**μ**) is equal to the hypothesized mean (**μ**H_0) = 100. Then we would say that the null hypothesis is that the population mean is equal to the hypothesized mean 100 and symbolically we can express as:

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

If our sample results do not support this null hypothesis; we should conclude that something else is true. What we conclude rejecting the null hypothesis is known as alternative hypothesis. In other words, the set of alternatives to the null hypothesis is referred to as the alternative hypothesis. If we accept H_0, then we are rejecting H_a and if we reject H_0, then we are accepting H_a. For $H_0 : \mu = \mu H_0 = 100$**,** we may consider three possible alternative hypotheses as follows:

If a hypothesis is of the type $\mu = \mu H_0$, then we call such a hypothesis as simple (for specific) hypothesis but if it is of the type $\mu \neq \mu H_0$ or $\mu > \mu H_0$ or $\mu < \mu H_0$ then we call it a composite (or nonspecific) hypothesis.

Alternative hypothesis	**To be read as follows**
$Ha : \mu \neq \mu H_0 \neq 100$	The alternative hypothesis is that the population mean is equal to 100 i.e., it may be more or less than 100.
$H_a : \mu > \mu H_0$	The alternative hypothesis is that the population mean is greater than 100.
$H_a : \mu < \mu H_0$	The alternative hypothesis is that the population mean is less than 100.

The null hypothesis and the alternative hypothesis are chose before the sample is drawn (the researcher must avoid the error of deriving hypotheses from the data that he collects and then testing the hypotheses from the same data.) In the choice of null hypothesis, the following considerations are usually kept in view:

1) Alternative hypothesis is usually the one which one wishes to prove and the null hypothesis is the one which one wishes to disprove. Thus, a null hypothesis represents the hypothesis we are trying to reject and alternative hypothesis represents all other possibilities.

2) If the rejection of a certain hypothesis when it is actually true involves great risk, it is taken as null hypothesis because then the probability of rejecting it when it is true is a (the level of significance) which is chosen very small.

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

   Generally, in hypothesis testing we proceed on the basis of null hypothesis, keeping the alternative hypothesis in view. Why so? The answer is that on the assumption that null hypothesis is true, one can assign the probabilities to different possible sample results, but this cannot be done if we proceed with the alternative hypothesis. Hence, the use of null hypothesis (at times also known as statistical hypothesis) is quite frequent.

13.6.2 Parameter and Statistic:

The main objective of sampling is to draw inference about the characteristics of the population on the basis of a study made on the units of a sample. The statistical measures calculated from

the numerical data obtained from **population units** are known as ***Parameters.*** Thus, a parameter may be defined as a characteristic of a population based on all the units of the population. While the statistical measures calculated from the numerical data obtained from **sample units** are known as ***Statistics.*** Thus a statistic may be defined as a statistical measure of sample observation and as such it is a function of sample observations. If the sample observations are denoted by $x_1, x_2, x_3, \ldots\ldots, x_n$. Then, a statistic T may be expressed as $\mathbf{T = \mathit{f}\ (x_1, x_2, x_3, \ldots\ldots, x_n)}$.

Measure	Mean	Variance	Proportion	Unit
Parameter	μ	σ^2	P	Population
Statistics	$\overline{X}$	SD^2	P	Sample

13.6.3 Type I and Type II errors:

In the context of testing of hypothesis, there are basically two types of errors we can make. We may reject H_0 when H_0 is true and we may accept H_0 when in fact H_0 is not true. The former is known as Type I error and the latter as Type II error. In other words, Type I error means rejection of hypothesis which should have been accepted and Type II error means accepting the hypothesis which should have been rejected. Type I error is denoted by α (alpha) known as α error, also called the level of significance of test; and Type II error is denoted by β (beta) known as β error. In a tabular form the said two errors can be presented as follows:

	Decision	
	Accept H_0	**Reject H_0**
H_0 (True)	Correct decision	Type I error (α error)
H_0 (False)	Type II error (β error)	Correct decision

The probability of Type I error is usually determined in advance and is understood as the level of significance of testing the hypothesis. If type I error is fixed at 5 per cent, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true. We can control Type I error just by fixing it at a lower level. For instance, if we fix it at 1 per cent, we will say that the maximum probability of committing Type I error would only be 0.01.

But with a fixed sample size, n, when we try to reduce Type I error, the probability of committing Type II error increases. Both types of errors cannot be reduced simultaneously. There is a trade-off between these two types of errors which means that the probability of making one type of error can only be reduced if we

are willing to increase the probability of making the other type of error. To deal with this trade-off in business situations, decision-makers decide the appropriate level of Type I error by examining the costs or penalties attached to both types of errors. If Type I error involves the time and trouble of reworking a batch of chemicals that should have been accepted, whereas Type II error means taking a chance that an entire group of users of this chemical compound will be poisoned, then in such a situation one should prefer a Type I error to a Type II error. As a result one must set very high level for Type I error in one's testing technique of a given hypothesis. Hence, in the testing of hypothesis, one must make all possible effort to strike an adequate balance between Type I and Type II errors.

13.6.4 The level of significance:

It is a very important concept in the context of hypothesis testing. We reject a null hypothesis on the basis of the results obtained from the sample. When is such a rejection justifiable? Obviously, when it is not a chance outcome. Statisticians generally consider that an event is improbable, only if it is among the extreme 5 per cent or 1 per cent of the possible outcomes. To illustrate, supposing we are studying the problem of non attendance in lecture among college students. Then, the entire number of college students is our population and the number is very large. The study is conducted by selecting a sample from this population and it gives some result (outcome). Now, it is possible to draw a large number of different samples of a given size from this population and each sample will give some result called statistic. These statistics have a probability distribution if the sampling is based on probability. The distribution of statistic is called a 'sampling distribution'. This distribution is normal, if the population is normal and sample size is large i.e. greater than 30. When we reject a null hypothesis at say 5 per cent level, it implies that only 5 per cent of sample values are extreme or highly improbable and our results are probable to the extent of 95 per cent (i.e. 1 – .05 = 0.95).

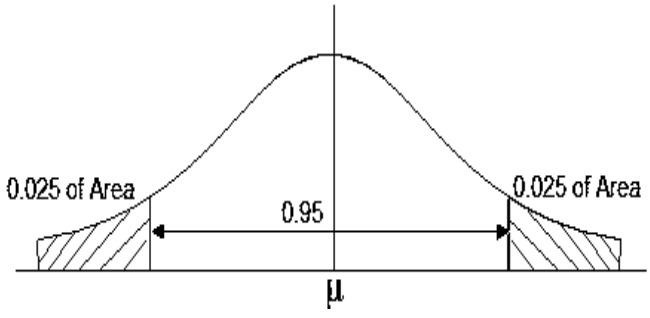


Fig 13.1

For example, above *Figure* shows a normal probability curve. The total area under this curve is one. The shaded areas at both extremes show the improbable outcomes. This area together is 0.05 or 5 per cent. It is called the region of rejection. The other area is the acceptance region. The percentage that divides the entire area into region of rejection and region of acceptance is called the **level of significance**. The acceptance region, which is 0.95 or 95 per cent of the total area, is called the **level of confidence.** These are probability levels. The level indicates the confidence with which the null hypothesis is rejected. It is common to use 1 per cent or 5 per cent levels of significance. Thus, the decision rule is specified in terms of a specific level of significance. If the sample result falls within the specified region of rejection, the null hypothesis is rejected at that level of significance. It implies that there is only a specified chance or probability (say, 1 per cent or 5 per cent) that we are rejecting H_0, even when it is true. i.e. a researcher is taking the risk of rejecting a true hypothesis with a probability 0.05 or 0.01 only. The level of significance is usually determined in advance of testing the hypothesis.

13.6.5 Critical region:

As shown in the above figure, the shaded areas at both extremes called the *Critical Region*, because this is the region of rejection of the null hypothesis H_0, according to the testing procedure specified.

Check your progress:

1. Which basic concepts regarding hypothesis testing have you studied?
2. Define:
   i. Null Hypothesis
   ii. Alternative Hypothesis

3. What do you mean by parameter and statistic?
4. What are the Type I and Type II errors?
5. What are level of significance and level of confidence?
6. What is Critical Region?

_____

_____

_____

_____

_____

_____

13.7 SUMMARY

25. Hypotheses are tentative propositions relating to investigative questions. They are formulated so that a researcher can test them on the basis of the facts already known to him or which he collects in the course of his investigation.

26. Hypotheses can be classified on the basis of their functions, on the basis of their nature, on the basis of their level of abstraction.

27. An ideal source of fruitful and relevant hypotheses is a fusion of two elements i) past experience and ii) imagination in the disciplined mind of the scientist.

28. An acceptable or a useable good hypotheses should fulfill certain conditions like testability, conceptual clarity, specificity, consistency, objectivity, simplicity etc.

29. When we compare method A with method B about its superiority and if we proceed on the assumption that both methods are equally good, then this assumption is termed as the null hypothesis. As against this, we may think that the method A is superior or the method B is inferior, we are then stating what is termed as alternative hypothesis. The null hypothesis is generally symbolized as H_0 and the alternative hypothesis as H_a.

30. A parameter may be defined as a characteristic of a population based on all the units of the population. A statistic may be defined as a statistical measure of sample observation and as such it is a function of sample observations.

31. Type I error means rejection of hypothesis which should have been accepted and Type II error means accepting this hypothesis which should have been rejected. Type I error is denoted by α known as α (alpha) error, also called the level of significance of test and Type II error is denoted by β (beta) known as β error.

32. The percentage that divides the entire area into region of rejection and region of acceptance is called the level of significance. The acceptance region, which is 0.95 or 95 per cent of the total area, is called the level of confidence. These are probability levels. The level indicates the confidence with which the null hypothesis is rejected. It is common to use 1 per cent or 5 per cent levels of significance.

33. Critical region is the region of rejection of the null hypothesis H0, according to the testing procedure specified.

13.8 QUESTIONS

1. What is hypothesis? Explain various types of hypothesis.
2. Discuss the role of hypothesis in the field of economic research.
3. What are the characteristics of a good hypothesis?
4. Write notes on:
    a) Role of hypothesis.
    b) Sources of hypothesis.
    c) Types of hypothesis.
    d) Criteria of a good hypothesis.
    e) Type I and Type II errors.
    f) Level of significance.
    g) Null and Alternative hypotheses.

14

Module 8
REPORT WRITING

Unit Structure:

14.0 OBJECTIVES

- To know the meaning and significance of report writing.
- To familiar the students about the various types of research report.
- To know the general structure of a research report.

14.1 MEANING AND SIGNIFICANCE OF A RESEARCH REPORT

1. Meaning:

Writing the research report is the final and very important step in the process of research work. The research report is a means for communicating one's research experiences to others. Of course, it requires different type of skills. *Research report is a narrative but authoritative document on the outcome of a research work.* It presents highly specific information for a clearly targeted audience. A well written research report is a means of presenting the studied problem, the methods of data collection and analysis, findings, conclusions and recommendations in an organized manner.

It is a means of judging the quality of research. Also, it is a means for evaluating the researcher's ability and competence to do

research. It provides factual base for formulating policies and strategies relating to subject matter studied. It provides systematic knowledge on problems and issues analyzed.

2. Significance of Report Writing:

Research report is considered a major component of the research study for the research work remains incomplete till the report has been written. As a matter of fact even the most brilliant hypothesis, highly well designed and conducted research study and the most striking generalizations and findings are of little value; if they are not effectively communicated to others. The purpose of research is not well served unless the findings are made known to others. All this explains the significance of writing research report. There are people who do not consider writing of report as an integral part of the research process. But the general opinion is in favour of treating the presentation of research results or the writing of report is the last step in a research study and requires a set of skill somewhat different from those called for in respect of the earlier stages of research. This task should be accomplished by the researcher with most care; he may seek the assistance and guidance of experts for the purpose.

Check your progress:

1. What do you mean by research report?
2. What is the significance of writing research report?

_____

_____

_____

_____

_____

14.2 TYPES OF RESEARCH REPORT

Research reports vary greatly in length and type. In each individual case, both the length and the form are largely dictated by the problems at hand. Let us see few details about some main types of research report:

1) Technical Report: In the technical report the main emphasis is on (i) the methodology used, (ii) assumptions made in the course of the study, (iii) the detailed presentation of the findings

including their limitations and supporting data. A general outline of a technical report is as under:

1. Summary of results
2. Nature of the study
3. Methodology used
4. Data
5. Analysis of data and presentation of findings
6. Conclusions
7. Bibliography & webliography
8. Technical appendices
9. Index Summary of results
10. Nature of the study
11. Methodology used
12. Data
13. Analysis of data and presentation of findings
14. Conclusions
15. Bibliography & webliography
16. Technical appendices
17. Index

However, it is to be remembered that even in a technical report, simple presentation and ready availability of the findings remain an important consideration and as such the liberal use of charts and diagrams is considered desirable.

2) **Popular Report:** This type of report is generally designed for non-technical users like executives, administrators, etc. It gives emphasis on simplicity and attractiveness. The emphasis is also given on practical aspects and policy implications.

   The format of this report is different from that of a technical report. There can be a liberal use of margins and blank spaces. The style may be more journalistic, but precise. While writing it, possibly it is made easy to rapid reading and quick comprehension.

3) **Interim Report:** An interim report is published when there is long time lag between data collection and the presentation of the results in the case of a sponsored project. In such a case, the study may lose its significance and usefulness. This report is short and may contain either the first results of the analysis or the final outcome of the analysis of some aspects which are completely analyzed. The interim report contains a narration of what is completed so far and its results are given. It presents a summary of the findings of that part of analysis, which has been completed.

4) **Summary Report:** A summary report is generally prepared for the use of general public. When the findings of a study are of general interest, a summary report is desirable. It is written in non-technical and simple language. It also contains large number of charts and pictures. It contains a brief reference to the objective of the study, its major findings and their implications. It is a short report which can be published in a newspaper.

5) **Research Abstract:** It is a short summary of the technical report. It contains a brief presentation of the statement of the problem, objectives of study, methods and techniques used and an over-view of the report. A brief summary of the results of the study can also be added. This abstract is primarily meant for the convenience of examiner, who can decide whether the study belongs to his area of interest. Results of a research can also be published as articles in research journals. A professional journal may have its own special format for reporting research.

   Thus, research results can be reported in a number of ways. In academic fields, the usual practice is to write the technical report and then prepare several research papers. In

practical field and problems having policy implications, it is more common to write a popular report. Researches conducted on behalf of Government or private or public organizations are usually presented in the form of technical report.

Check your progress:

1. What are the main types of research report?
2. Give an outline of a technical research report.

_____

_____

_____

_____

_____

14.3 FORMAT OR STRUCTURE OF A RESEARCH REPORT

A report has a number of clearly defined sections in certain order. The order of the headings and sub-headings may vary according to nature and type of research. Yet, it is possible to suggest a general sequence of contents in general format of a research report as following:

I. Introductory Items: (*Preliminary Pages*)

1) Title page.
2) Researcher's declaration.
3) The certificate of the research guide or supervisor.
4) Preface / Acknowledgement.
5) Contents.
6) List of tables.
7) List of graphs and charts.
8) Abstract or Synopsis.

II. Body of the report: (*Main Text*)

1) Introduction. (This may include the following items).

i. Theoretical background of the topic.
ii. Statement of the Problem.
iii. Review of Literature.
iv. The Scope of the study.
v. Objectives of the study.
vi. Hypothesis.
vii. Definition of concepts used.
viii. Model or Chapter Scheme.

2) The Design.

i. Methodology, including overall type and methods used for data collection.
ii. Sources of data.
iii. Sampling plan.
iv. Instruments of data collection.
v. Field work.
vi. Data processing and analysis (plans).
vii. An overview of findings.
viii. Limitations of the study.

3) Results

4) Summary, Conclusions and Recommendations.

III. Concluding Items: (*End Matter*)

1) Bibliography and Webliography.

2) Appendix.

i. Copies of data collection instruments (like interview schedules, questionnaire).
ii. Technical details on sampling plan.
iii. Complex tables, Primary tables.
iv. Supporting documents.
v. Statistical computations.
vi. Glossary of new terms used in report.

Check your progress:

1) Which are the main parts of the research report?
2) Prepare general structure for a research report.

_____

_____

_____

_____

_____

14.4 STEPS IN PLANNING REPORT WRITING

It is necessary to write research report with careful pre-planning. This planning consists of following steps:

1. Effectiveness of communication: A research report is a means of communication and it is necessary to first consider the basic questions which determine the effectiveness of communication. The considerations of effective communication are basically linked with the target audience for whom the report is written and who writes this report i.e. the agency or individual conducting the research. The manner in which the research findings are expressed i.e. style of writing, is also equally important.

2. The identification of target audience: The form and type of reporting and other aspects depend upon the type of reader or the user of the report. The identification of the target audience depends on who is the researcher and what is his intention. The target audience can be academic community, the sponsor of the researcher or the general public. The communication characteristics, i.e. the level of knowledge, the type of language that is understood and appreciated, the expectation form the report are not identical for different groups of audiences.

3. Logical analysis of the subject matter: The subject matter can be developed logically or chronologically. This is because logical analysis implies development of the subject from simple matter to the complex. It is also based on logical connections or associations between different factors. Therefore, planning for logical presentation is important.

4. Preparation of the final outline: Outline is a framework on which the long written report is constructed. It is an aid to decide the logical arrangement of the material to be included in the report

and the relative importance of various points. Outline is drawn after preparation of the format of the report. It gives cohesiveness and direction to report writing. The outline can be according to topic or sentence. In the topic outline, the topic headings and the sub-topic headings are noted and the points to be discussed under each head are noted in short forms or with key words. In case of sentence outline, it gives more details about the points to be included in the report.

5. Preparation of the rough draft and final draft: The rough draft follows the outline and the research should write down the broad findings and generalizations. The rough draft can also include various suggestions which help in improving the final writing. A rough draft is essential to avoid mistakes or omission in the final draft. It is possible to polish the language of the rough draft in the final draft. Final draft is written after a careful scrutiny of the rough draft.

6. Preparation of Bibliography and Webliography: Bibliography is a list of books which provide references to the work undertaken and webliography is a list of website addresses where the researcher visited for references and consultancy. Both are appended to research report in a systematic manner. The bibliography should be arranged alphabetically and may be divided into three parts. Fist part may consist of books, second part may contain magazines, periodical and newspaper articles and the third part may contain web-addresses. The entries in the bibliography should be according to a certain order like name of the author, title of the book in *italics*, place, publisher and date of publication, edition, page number if required, etc. For example, see the list of suggested reading given below this.

Check your progress:

1. What is importance of planning for report writing?
2. Which steps are necessary in planning for report writing?
3. What is outline?
4. Differentiate between draft and final research report.

_____

_____

_____

_____

_____

14.5 SUMMARY

1. Writing the research report is the final and very important step in the process of research work. Research report is a narrative but authoritative document on the outcome of a research work.

2. In the technical report the main emphasis is on the methodology used, assumptions made in the course of the study and the detailed presentation of the findings including their limitations and supporting data.

3. In Popular report there can be liberal use of margins and blank spaces. The style may be more journalistic, but precise. The emphasis is given on practical aspects and policy implications.

4. An Interim report is published when there is long time lag between data collection and the presentation of the results in the case of a sponsored projects. It presents a summary of the findings of that part of analysis, which has been completed.

5. Summary report is written in non-technical and simple language. It also contains large number of charts and pictures. It contains a brief reference to the objective of the study, its major findings and their implications.

6. Research abstract is a short summary of the technical report. It contains a brief presentation of the statement of the problem, objectives of study, methods and techniques used and an over-view of the report.

7. General format of a research report includes the following:

   a. Introductory items

   b. Body of the report : Introduction, The design, Results, Summary, Conclusions and Recommendations

   c. Concluding items : Bibliography and Webliography, Appendix

8. Steps in planning Report writing consists of different steps.

14.6 QUESTIONS

1. Explain the main types of research report.
2. Explain the meaning of research report and explain its structure.
3. What are the steps involved in planning for writing research report?

15

ORGANIZATION AND STYLE OF RESEARCH REPORT

Unit Structure:

15.0 OBJECTIVES

- To know the principles of writing research report.
- To understand the organization structure and style of report writing.
- To understand that which precautions are to be taken while writing research report.

15.1 PRINCIPLES OF WRITING THE RESEARCH REPORT

Considering the previously discussed general format of the research report, there are certain principles of standard practices which should be observed in writing a research report. Those principles comprise organization, style of research report and essentially some precaution in writing a research report.

1. Organization of Report:

In the organization of research report, following points are essentially be considered:

a. Size and physical design: Accordingly prescribed size of the paper and the given general instructions are to be maintained throughout the writing of report. Ofcourse, writing should be in double-space and on one side of the page.

b. Procedure: Various steps in writing the report which are explained before in previous unit should be strictly followed.

c. Layout: Keeping in view the objective and nature of the problem, the layout of the report should be suitable according to the type of research report.

d. Treatment of quotations: Quotations should be placed in quotation marks and double-spaced, forming an immediate part of the text. But if a quotation is of a considerable length then it should be single-spaced and indented at least half an inch to the right of the normal text margin.

e. Footnotes: Regarding footnotes one should keep in view the following things:

i) It should provide proper cross references, data sources.

ii) It should be written at the bottom of the page and separate from the main text.

iii) It should be numbered consecutively beginning with 1 in each chapter separately and such number should be typed a little above the line at its end.

iv) It is always be typed in single space and make separate form one another by double space.

f. Documentation Style: Regarding documentation, the first footnote reference to any given work should be complete with all its essential facts about the edition used.

g. Punctuation and Abbreviations: The punctuation marks should be proper for meaningful reading and the abbreviations used should be most familiar with all.

h. Use of Statistics, Charts and Graphs: For the more clarification and simplification, use of statistics in research report is of great important. One may well remember that a good picture is often worth more than a thousand words. Statistics are usually presented in the form of tables, charts, bars and line-graphs and pictograms. Such presentation should be self explanatory and complete in itself. It should be suitable and appropriate looking to the problem at hand. Finally, statistical presentation should be neat and attractive.

i. The final draft: Revising and rewriting the rough draft of the report should be done with great care before writing the final draft. For the purpose, the researcher should put to himself questions like: Are the sentences written in the report clear? Are they grammatically correct? Do they say what is meant? Do the various points incorporated in the report fit together

logically? And finally, having at least one colleague read the report just before the final revision is extremely helpful.

j. Bibliography and Webliography: These should be prepared and appended to the research report as discussed earlier.

k. Preparation of the index: At the end of the report, giving a neatly prepared index plays a role of a good guide to the readers. Index may be prepared both as subject index and as author index. The former gives the names of which they have appeared or discussed in the report, whereas the latter gives the similar information regarding the names of authors. The index should always be arranged alphabetically.

2. Style of Report:

There are certain basic aspects relating to the style of a research report. These aspects are based on the principles of writing research report. The style of the research report is closely related with arrangement of the content, grammatical aspect, quotations, index, calculations, use of graphical presentation, arrangement of bibliography, etc. The style of the research report can be maintained by keeping the following precautions in mind while writing the research repot.

3. Precautions are to be taken while writing Report:

There are certain principles of standard practice which should be observed and take precautions while writing a research report. Those are as under:

a) Organization of Report: The layout should be well throughout and must be appropriate with the objective of the research problem. Each chapter may be divided into two or more sections with appropriate headings and in each section; margin headings and paragraph headings may be used. Physical presentation should also be neat. A page should not be filled from top to bottom. It should have enough margins on all sides. The report should be long enough to cover the subject, but short enough to maintain interest.

b) Style: A research report should not be dull. It should be so written as to sustain a reader's interest. A report requires a style different from other academic writings like essays, stories, etc. It is a formal presentation of an objective, unbiased investigation. Therefore, it should be written in formal, Standard English without making it uninteresting.

A research report does not require elegant words. It just needs a plain discussion which is accurate, concise and also readable

and coherent. The writer should keep in mind the reading ability of the target audience and avoid unclear writing. To make the report clear and readable:

i) Avoid jargon and pompous style.

ii) Avoid offensive words.

iii) Do not use adjectives that exaggerate the facts e.g. awful, gigantic, immeasurable, etc.

iv) Avoid tautology or repetition like return back, the true facts and so on.

v) Avoid unnecessary words or descriptions. For example, one can say 'many' instead of 'large number'.

vi) Try not to use stereotyped phrases, like 'as a rule', 'it goes without saying', 'it seems', 'last but not least', 'by and large', etc.

vii) Use of **slang** also does not make good reading in a report.

In short, the report should convey the meaning in as simple words as possible.

c) Content: The report must present the logical analysis of the subject matter. Report must contain necessary charts, graphs and statistical tables in addition to the important summary tables.

d) Grammar: Presentation in a report should be free from spelling mistakes and grammatical errors. The important rules of grammar relate to: Spelling of words, punctuations, capitalization and other standard rules, etc.

e) Quotations: Footnotes, documentation, abbreviations are used strictly according to the convention or rules of incorporating them. Every quotation used should be acknowledged with a footnote. Do not use abbreviations in the text of the report. However, abbreviations are desirable in footnotes, tables and appendices.

f) Index: It is also an essential part of a report and it should be properly prepared.

g) Calculations: Calculated confidence limits for statistical results must be mentioned. The various difficulties experienced in the conduct of study may also be mentioned.

h) Bibliography and Webliography: Bibliography should be arranged according to the rules. Intellectual honesty demands that all source material should be acknowledged by the

researcher. Also, consulted web addresses are to be noted properly for the interest of the readers.

i) **Other essential Considerations:** Appendices should be enlisted in respect of all the technical data on the report. Index is also an essential part of a report and it should be properly prepared.

15.2 SUMMARY

1. There are certain principles of standard practices which should be observed in writing a research report.
2. In the Organization of research report some points are essentially be considered such as size and physical design, procedure, layout, treatment of quotations, footnotes, documentation style, use of statistics, charts and graphs etc.
3. The style of the research report is closely related with arrangement of the content, grammatical aspect, quotations, index, calculations, use of graphical presentation, arrangement of bibliography etc.
4. Precautions are to be taken while writing report such as organization of report, style, content, grammer, quotations, index, calculations, bibliography and webliography etc.

15.3 QUESTIONS

1) Explain the principles of writing research report.
2) Write a note on:
    i. Organization of Report.
    ii. Style of Report.
    iii. Importance of Bibliography.
3) Explain the precautions to be taken in writing a research report.