# Yashwantrao Chavan Maharashtra Open University

## Research Methods and Statistical Analysis

*Index*

# Yashwantrao Chavan Maharashtra Open University

**Vice-Chancellor :** Dr. B. P. Sable  **Director :** Dr. Surya Gunjal




## Expert Writer

**Dr. S. L. Sananse**
Asso. Prof. of Statistics,
Zonal Agril. Research
Station, Karjat

**Dr. A. M. Degaonkar**
Asso. Prof. of Statistics,
College of Agriculture,
Latur

**Prof. P. R.Waghmare**
Asso. Prof. of Statistics,
Marathwada Agricultural
University, Parbhani

**Prof. S. W. Jahagirdar**
Asso. Prof. of Statistics,
College of Agriculture,
Akola

**Dr. S. M. Kareppa**
Asso. Prof. of Statisti cs,
Mahatma Phule Krishi
Vidyapeeth, Rahuri

## Editing Team

**Dr. V. D. Deshmukh**
Ex Professor & Head
Department of Statistics
MPKV, Rahuri

**Dr. Surya Gunjal**
Director,
School of Agricultural Sciences,
YCMOU, Nashik

**Dr. Somnath Suryawanshi**
Co-ordinator (PG),
School of Agricultural Sciences,
YCMOU, Nashik

**Dr. Prakash Atkare**
Reader,
School of Agricultural Sciences,
YCMOU, Nashik



**Production :** Shri. Anand Yadav, Production Officer, Print Production Centre, YCMOU, Nashik





❑ First Edition : June 2003 ❑ Publication No. : 1180

**Typeseting :** Akosis, Nashik
**Cover Design :** Nitin Mahamuni, YCMOU, Nashik

**Printer :** M/s. Vikas Electricals &Ceramics Products Pvt. Ltd., M.I.D.C. Area, Satpur, Nashik - 7

**Publisher :** Dr. N.R. Kapadnis, Registrar, Yashwantrao Chavan Maharashtra Open University, Nashik

**ISBN : 81-8055-042-7**

## FOREWORD

*Dear Students*

*The School of Agricultural Sciences of our university has always remained at the forefront in planning and implementing need based and innovative programmes in Agricultural Sciences through the distance mode. They have worked out a unique and flexible model of multiple entry and multiple exit that provide for vertical mobility for each and every aspirant in the farming community.*

*This university has received an International Award of Excellence in Distance Education from the Commonwealth of Learning, Canada, for its innovative approach in operating distance education to suit the common man and woman in every sphere of life. The university has catered to over six hundred thousand learners in Maharashtra including 40000 farmers since its establishment in 1989. The School of Agricultural Sciences has contributed substaintially in this endeavour.*

*The School of Agricultural Sciences would now like to move further ahead and cater to the educational needs of village leaders, extension specialists and policy makers in the field of Agricultural Education and Development. Taking into account the national perspective, these are the nontraditional areas that need to be addressed with the help of the powerful tools of Information and Communication Technology (ICT) in order to reach the unreached.*

*The Post-Graduate and Research programmes in Agricultural Communication, Agricultural Extension and Agricultural Development would definitely equip and arm the agricultural experts in using tools of Information and Communication Technology and promote Virtual Education in Agriculture and Rural Development. The virtual education would proceed from Class Education to Mass Education and would cover distance learners from un-reached communities including farmers and farm women.*

*I am quite confident that the learners striving to complete these innovatively developed educational programmes would soon establish themselves as pillars of the distance education movement in India.*

*I wish you all a happy and rewarding learning experience in this university.*

*Professor B. P. Sabale*
*Vice-Chancellor*
*Yashwantrao Chavan Maharashtra Open University, Nashik*

## WEL COME

*Dear Friends*

*On behalf of the School of Agricultural Sciences, I would like to welcome you to our Post Graduate and Research programmes in Agricultural Communication, Agricultural Extension and Agricultural Development. The programmes would meet the long felt need of extension workers and policy makers in the field of Agriculture and Rural Development.*

*The term distance education refers to Intentional Processes of Teaching and Learning in which physical space separates teachers and learners. Teachers and learners communicate through various media and an educational organization exists to design, facilitate and evaluate the educational process.*

*The Food and Agriculture Organization (FAO) is an international catalyst involved in promoting distance learning for diverse and globally distributed learners, organizations and communities, whose capacities and actions ensure the achievement of Food Security and Rural Development*

*The role of FAO in distance education and learning was spelt out in the Rome declaration on World Food Security, in 1996. The international community leaders had made the core commitment thus "We recognize the need to adopt policies conducive to investment in human resource development, research and infrastructure for achieving food security. Our sustainable development policies will promote full participation and empowerment of people, especially women and equitable distribution of income, access to health care and education and opportunities for youth." Consistent with their commitment, FAO has recognised the best practices in pedagogy of distance education and distance learning.*

*With this commitment from FAO, it may safely be assumed that at flexible mode will soon be capable of reaching every home and will empower each individual to achieve a better standard of living. Since you are the 'early bird' having an access to distance education, your involvement as a student in our post-graduate programme, will certainly go a long way in benefiting the distance learning system in the country.*

*I wish you all the best in your learning endeavours.*

*Professor Surya Gunjal*
*Director*
*School of Agricultural Sciences*
*Yashwantrao Chavan Maharashtra Open University, Nashik*

# Unit 1 : Research Methods : An Introduction

***Index***



## 1.1 Introduction

Research is an inseparable part of human knowledge. Life would lose its taste without research, exactly in the same manner as a food without salt. A long journey from stone-age to computer-age has been the result of research. As long as man's inquisitive pursuits do not end, the onward march of human society to more advanced achievements will not stop. The shape of things and the order of the society shall continuously change. The process shall continue forever. New innovations and inventions shall be received greater excitement and the process shall continue forever.

All research is essentially socially oriented whatever branch of knowledge it may be, research has social bearing. Research in social science is the direct outcome of man's urge to understand his society, its nature and working. Human beings do not operate under controlled conditions, on the contrary they are always under the influences of as environmental, psychological and social factors. These influences interact freely with each other and seldom operate in isolation. In a huge societal laboratory it is hard to conduct precise experiment. Societal behavior, problems and aspirations are well understood by research investigator and appropriate policies formulated accordingly, to suit the society and it's various segments in particular.

**After the study of this unit, you will be able to know and understand :**

- Different types of research.
- Scientific methods of research..
- Problems in research.

## 1.2 Contents

### 1.2.1 Meaning and Objectives of Research

Research is the activity of solving problem, which adds new knowledge, and develope theory as well as gather evidences to test generalization. The Webster's international dictionary defines research as **"a careful and critical investigation and examination in seeking facts for principles and diligent investigation in order to ascertain some things"**. The people when they talk about research are not specific and precise enough to get a complete and clear idea of what research involves. Improving upon the definition given above, it may be stated that research refers to a critical and exhaustive investigation or experimentation having it as aim of revision of accepted conclusions in the light of newly discovered facts. In modern times, research is often a corporate affair in as much as the complex technique of collecting and processing data resulting in generalizations.

Another important definition is given by the encyclopedia of social sciences. It defines research as **"the manipulation of things, concepts or symbols for the purpose of generalizing to extend, correct or verify knowledge, whether the knowledge aids in construction of theory or in the practice of an art".**

Research should never be treated as a piece of compilation work. Research is always expected to be some thing original or a piece of work that advances human knowledge. Research must be based on facts. Observable data forms a sound basis of research. The modern concept of research is, therefore, broad based and provides a meaningful investigation into any field of research. According to P. V. Young, Social research may be defined as “a scientific undertaking are means of logical and systematized techniques seeks to :

(i) discover new facts or verify and test old facts;

(ii) analyze the sequences, inter-relationships and casual explanations which were derived within an appropriate theoretical frame of references; and

(iii) develop new scientific tools; concepts and theories which would facilitate reliable and valid study of human behavior.

**Characteristics of Research**

After we have gone through the characteristics of scientific thinking, we can move to the next logical step of knowing the specific characteristics of research

1. Research gathers new knowledge or data from primary or first hand sources. It is not research when one simply restates or reorganizes what is already known or what has been written.
2. Research is expert, systematic and accurate investigation. The researcher tries to secure an expertise before undertaking any investigation. He plans the procedures carefully. The data are gathered, recorded and analyzed with as complete accuracy as possible. He uses standardized and valid data-gathering tools or instruments as he can find or devise. He also employs mechanical means to improve upon the accuracy of human observation, recording and computation of data.
3. Research is logical and objective, applying every possible test to verify the data collected and the procedures employed. The researcher eliminates personal feelings and preferences from his research activity. He resists the temptation to seek only that data which support his hypothesis. There is no attempt to persuade or to prove the emphasis on testing, rather than on proving the hypothesis. He argues within the scope and relevance of his data. He builds only those generalizations, which directly emerge out of the data. Research endeavours to organize data in quantitative terms as far as possible. It further tries to express them as numerical measures.
4. Research is patient and unhurried type of task. The researcher is willing to make

painstaking effort. He works patiently towards sound conclusions. He realizes that significant findings do not come as a result of hurried and careless procedures.

5. Research requires courage. The researcher would not afraid of unpleasant consequences of his findings. He has the guts to speak and record the truth. He is willing to follow his procedures to conclusions that may be unpopular and bring social disapproval. We can mention several historic examples of the penalties which researches suffered because their findings came in conflict with dogmas and commonly accepted beliefs.
6. Research is highly purposive. It deals with a significant problem which demands a solution. Research places emphasis upon the discovery of general principles and scientific generalizations, applied to the solution of a wide range of problems.
7. Research maintains rigorous standards. The research worker is expected to be a scholarly and imaginative person possessing highest integrity. He takes such step according to expected specifications and norms. He keeps his work scrupulously free from loopholes. Research is a job of great responsibility. Its findings carry for reaching implications.
8. Research usually involves, as a step, a hypothesis or a set of hypotheses concerning the tentative explanation of a phenomenon or the solution of problem. Research is carefully recorded and reported observation and data. Every term is carefully defined, all procedures are described in detail, all limiting factors are recognized, all references are carefully documented and all results are objectively recorded. All conclusions and generalizations are cautiously arrived at with due consideration for all of the limitations of methodology, data collected, and errors or human interpretation.

**Objectives of Research**

The purpose of research is to discover answers to questions through the application of scientific procedure. The meaning of research is to find out the truth which is hidden and which has not been discovered as yet. Though each research study has its own specific purpose, we may think of research objectives as falling into a number of following broad groupings.

1. **Development of Knowledge :** Every science tries to collect systematized knowledge about the branch or subject matter that it studies which applies the research as a method. The main object of research is to add to the knowledge. Social research is an organized and scientific effort to acquire further knowledge about the social phenomenon and social facts.
2. **Scientific Study of Social Life :** Social research is an organized and scientific effort to acquire further knowledge about the social phenomenon and social facts. The researcher makes study of the collective processes, social change, social structure, social process etc. Apart from it, it demands study of human being human behavior and collects data about various aspects of social life of man and formulates laws in this regard.
3. **Welfare of Humanity :** No researcher or scientist makes study only for sake of study. He has to direct his study to some higher aim, which invariably is " welfare of humanity". To promote the welfare of humanity by the result of investigation should be the ultimate aim of a research and research scientist.
4. **Classification of Facts :** Social research aims to clarify facts. In the given universe of discourse to find the specific determined sequences and inter-relationship of facts and their social settings.
5. **Social Control and Prediction :** Through social research one makes study of social phenomenon, events and the factors that govern and guide them. The ultimate objects of many research undertakings is to make it possible, to predict the behavior of particular type of individuals under specific conditions.

Research finds out new facts and verifies the old facts on the basis of the tests applied to old facts. Social research thus studies the social values, beliefs, traditions, events etc.

## 1.2.2 Types of Research

Research has been classified differently depending upon the approach, the purpose and the nature of a research activity. Broadly speaking research can be classified into two groups such as Fundamental Research and Applied Research and various categories.

## 1. Fundamental Research

This type of research is original or basic in character. Painstaking research worker, with qualities of honesty, integrity and his lust for the search of truth, makes persistent partient efforts to discover something new to enrich human knowledge in a fundamental fashion such research is known as fundamental research. The fundamental research :

(i) helps in developing principles,

(ii) helps to find out the central factors in practical problems,

(iii) becomes standard procedure for researcher to find out an answer to the problem.

Fundamental research can be grouped into two types.

### (a) Discovery of a New Theory

Fundamental research may be entirely new discovery, the knowledge of which has not existed so far. Such a discovery may follow from the researcher's own idea or imagination. This depends upon how genius a researcher is. The researcher is thirsty for knowledge and eventually has a ocean of knowledge in his possession from this ocean emerges a jewel that enlightens the world. Newton's and Galileo's contributions are fundamental in nature as these depended upon their own imagination. Since these fundamental contributions form the basis of different theories, it may not be wrong to call such a research as theoretical research.

### (b) Development of Existing Theory

This type of research is also held to take the shape of an improvement in the existing theory by relaxing some assumptions or by developing new theory with existing one as its basis. There also exist the possibilities of re-interpretation of the theory that has been already developed. By questioning some of the assumptions of Keynesian theory, Friedman came out with new interpretations of monetary phenomenon. Theories developed in capitalist countries have often been challenged by researchers of the socialist block and they have often either re-interpreted or developed new theories akin to other theories which were already existing.

## 2. Applied Research

This type of research is based on the application of known theories and models. The applied research is conducted to test empirical content or the basic assumptions or the validity of a theory under given conditions. Applied research has practical utility in developing countries. Instead of spending lavishly on fundamental or pure research, a less developed country may benefit by applying the existing theories and models and seek plausible explanations of socio-economic behavior. Applied research often take the form of a field investigation and aims at collecting the basic data for verifying the applicability of existing theories and models in given situation. If a field selected is not properly identified for the application, the results are bound to give wrong interpretations and a judgment.

## 3. Various fCategories of Research

The research these days are categorized into 8 groups as follows :

**(a) Ex-post Facto Research :** Ex-post facto research is systematic empirical investigation in which the scientists do not have direct control on independent variables because they are inherently not manipulable in its nature. In this kind of research, researcher's control on the behavior of independent variables is very weak and in many cases no control is possible. Social science does not afford a possibility of controlling the changes in the behavioral pattern of independent variables. These are usually affected by complex social phenomenon and one can only examine how the dependent variable is affected by such situations. In social and scientific research, there always exists a possibility of identifying dependent and independent variables either singly or in-groups. Thus, this helps in postulation of models for carrying out their analysis for researcher's utility.

This type of research has some weaknesses. These weaknesses are.

(i) The inability to control the changing pattern of independent variable.

(ii) The ex-post facto research findings carry the risk of improper interpretations. This is mainly due to the fact that control on independent variable is very weak.

(iii) The ex-post facto research may not have any particular hypothesis, as there is likelihood that such a hypothesis may predict a spurious relationship between independent and dependent variables.

However, despite of these weaknesses there is a considerable scope for conducting research of this type in social science. Since, most of the social behavioral problems cannot be subjected to experimental investigations; ex-post facto research is a good explanatory instrument of changes that take place in the dependent variables. If the set of independent variables is carefully selected and proper techniques adopted for analyses, meaningful insights can be gained into changing pattern of dependent variables.

**(b) Experimental Research :** Experimental research is considered to be the photo type of the scientific method. It involves procedure for gaining knowledge by collecting new or fresh observations under controlled conditions. Further it reveals casual relation and interactions. Hence, it calls for precision and accurate observation. In experimental research the independent variable as well as dependent variable are involved. The independent variable can be manipulated as an experimental variable in order to see the effect on dependent variable.

Laboratory experiments usually help to discover relations under pure and uncontaminated conditions. An experiment conducted in the created space under special conditions is called “Laboratory experiment”. e.g. if an experiment is made with students in the college itself would be field experiment, but created a classroom and experiment upon it, it will be ‘laboratory experiment’. The laboratory experiment : (i) Attempts to discover relations under uncontaminated conditions, (ii) Tests predictions derived from theory primarily and other researchers, secondarily and (iii) Refines theories and hypotheses.

The experimental research is most often carried out in the field of physical sciences. Social science theories and models can be verified, refined and redesigned only when field situations are taken into account.

**(c) Field Investigation Research :** A field experiment is a research study in a realistic situation in which one or more independent variables are manipulated by the experiments under controlled conditions, as the situation will permit. The most common feature of the field experiment is practical orientation.

The weaknesses of field experiment are of practical nature. The investigator himself works under several influences and may be faced with unpleasant situations. Uncontrolled environmental influences may significantly affect the independent variables. Many times a combination of dependent variable independent variables may be such that the dependent variables are not responsive to the changes occurring in independent variables. A practical difficulty in the field of this research is that of randomization of units in the given population. Unless the field of operation has been selected in scientific manner, the investigations may not proceed on the lines desired by the researcher.

A field experiment is generally credited with a few virtues which are supposed to be unique to this category of research. These virtues may be listed as :

(i) The variables in a field experiment operate more strongly than those used in laboratory experiment. This is because of the fact that field situation takes stock of realistic natural operations.

(ii) Field experiments have the advantage of investigation of more fruitfully the dynamics of inter-relationships of small groups of variables.

(iii) Field experimental studies are also ideal to testing of the theory and to the solution of the real world problems.

Field experimental studies, therefore, are important part of the applied research which play an important role in pointing out the nature and direction of the refinements required for an existing doctrine.

**(d) Survey Research :** This type of research has become vary popular these days as a scientific method for discovering relevant impact and inter relationships of social and psychological variables from given populations. In survey research studies samples are chosen from the population to discover the relative incidence and inter relations of sociological and psychological variables. Survey research as a matter of fact has developed as a research activity along with the development of sampling theory and its diverse procedures. The advantage of this type of research is that it links sample investigations with population and thereby offers an easy opportunity of studying population behavior through sample survey research assessments. Survey research is mostly devoted to the study of characteristics of the population under investigation. These days this type of research is held to be significant in making valuable contribution to social sciences research methodology. It has directly helped in the improvement of the sampling procedures and their applicability to real world solutions besides also suggesting improvements in resolving the complex situations to unambiguous experimentation with a view to obtaining research findings.

This type of research has the advantage of greater scope in the sense that a large volume of information can be collected from a very large population. Survey research is more expensive but the amount and quality of information that is collected makes such investigation very economical.

This information is also accurate, within the range of sampling errors because trained and technically knowledgeable personnel are employed for the job.

This type of research, however, suffers from some limitations, which could be listed as follows:

(i) It is charged that surveys information touches only the surface of the research field and does not make a deeper thrust into it.

(ii) Sometimes if sample information have not been collected very carefully, the magnitude of sampling error may be too large to render the sample results reasonably accurate.

(iii) Since the sample research is based on the respondents interviews, the problems of personal inhibitions, indifferences and unawareness of the nature and purpose of investigation renders survey information invalid or at least imprecise.

Despite the weaknesses mentioned above, survey research holds a greater promise for social researchers in future for exploring socio-economic information in diverse population.

**(e) The Case Study Research :** A case study is deep and intensive study of a particular social unit, confined to a very small number of cases. Thus the field of study in the case study method is limited but it aims at studying all aspects of a social unit. It also seeks to determine social process; it reveals the complexity of factors and indicate their sequences and their relationships. It is also a diagnostic study oriented towards finding out what is happening and why is it happening and what can be done about it. Case study, says Charles Colley, says that case study is "depends on our perception and gives us a clearer insight into life.".

***Sources of Data***

Case studies are not limited to any single sources of data collection. A number of sources of data collection should be employed by the researchers. Mr. Thomas made extensive use of personal documents considered them a chief instrument in reaching the actual experience and attitudes of persons and groups as well as in securing a cross-section of the entire process of their social becoming. Many others have made exhaustive case studies with preliminary interviews, conferences at specified intervals, conversations, observations, etc. Sometimes a researcher has to use different techniques and tools to collect data about the different aspects of the cases.

In the case study method the main sources of data are Personal documents and Life History

*(I) Personal Documents :* Most of the people keep personal records, documents, letters and write their autobiographies or memories. These documents play an important role in the case study as they contain description of the important events of the life of the writer as well as his relations towards them. These documents may also contain the description of even those events in which the narrator has played his part only as a witness. Thus, the personal documents are very helpful is studying the personality of the writer and his relations to different circumstances of life as the writer is an identical part of the group; they may represent not only the reaction of the person but of any typical number of the group.

*(II) Life History* : Life history is the study of various events of respondent's life together with an attempt to find their social significance. Life history data is generally written material about his life, conference at specified intervals, experimental studies, observations, post-experimental interviews, various tests on the analysis of facts so collected in order to draw vivid generalizations from them.

The case study method is very useful in helping the researcher to develop and formulate scientifically sound hypothesis for more research on broader level. As already mentioned, researcher may not start with a given hypothesis but may desirably undertake a case study for formulating such hypothesis for further research. Case study method has also an advantage in making a multi-dimensional exploration of the same unit and thus enriches the knowledge pertaining to a particular case for further use in policy formulations.

When a case study is undertaken, some of the areas of research may not have occurred to the researcher's mind and the very case study, may open out new avenues of research where fruitful investigations can be undertaken either by the same researcher or other researchers.

**(f) Evaluation Research :** This type of research is primarily directed to evaluate the performance of the developmental projects and other economical programmes that have been already implemented. This evaluation is carried out with respect to the broader changes and major objectives of a programme. Evaluation does not only aim at the physical achievements of a project but at the size and direction of variables that the project has resulted in attitudes, interests, thinking patterns, work habits, socio-cultural adaptability and psycho-dynamics of the population around which project has been launched.

The evaluation of research is of there types, mainly (a) Concurrent evaluation (b) Periodic evaluation and (c) Terminal evaluation.

(a) ***Concurrent evaluation*** is a continuous process of an inspection of the project that has been launched. The word concurrent itself points out to the fact that the evaluation goes on side by side with the implementation of a project. Such type of research not only evaluates the performance but also stimulates it and gives direction and control as and when possible.

(b) ***Periodic evaluation*** takes place at different phases or stages of performance of the project. If, for example, a project has been launched in three different phases, evaluation research is conducted at the completion of each stipulated phase of the project thus launched. It enables us to evaluate the performance of the completed phase and make adjustments in the subsequent phases after keeping in view the failures and successes of the previous phase.

(c) ***Terminal evaluation*** is the evaluation of the final phase of the project. Once the project has been completed, an over all assessment is made to see how best a project has served the objectives for which it was launched.

While carrying out the evaluation research, a researcher can significantly look into the performance of the project in relation to the objective that has been laid down and in line with the processes that has been recommended for its implementation.

**(g) Action Research :** It is the study undertaken with a goal of immediate application. In other words, it refers to study of 'on the job' problem. In this type of research, the researcher appears as a participant rather observer and therefore is actively and even emotionally involved in these results and their application. Action research is spread over different phases such as :

(i) A base line survey in order to get possible information relevant to the subject matter should be collected to get an idea of the existing situation.

(ii) Launching the action research project.

(iii) The periodical assessment of action research project.

(iv) Making changes and modifications in the working of the project.

The action research is not much different to concurrent evaluation of a project. Examples of action reports are: A teacher conducts action research to improve his own teaching. The practitioner attempts to study his problem scientifically for guiding, correcting and evaluating action.

**(h) Interdisciplinary Research :** Interdisciplinary research is also called as coordinated research. It is characterized by exchange of views and ideas from one discipline to another. Because of contribution of various disciplines, the hidden facts become uncovered quicker and discoveries are made fruitful and thereon integrated development. Due to interdisciplinary research much time, money and effort are saved for example research on agriculture. The research in agricultural sector needs a coordinated approach because this sector's progress depends upon the others, viz., banking sector, from service sector, subsidiary occupations such as diary, sheep, goat, piggery etc. Which supplements income to the agriculture. The essential characteristic features of interdisciplinary research are:

(i) It is a cooperative venture and a common object.

(ii) Different disciplines must work together.

(iii) The research is utilitarian in nature.

### 1.2.3 Research and Scientific Methods

Scientific method is collective term denoting the various processes by the aid of which the sciences are built up. In a broad sense, any method of investigation by which scientific or any other impartial systematic knowledge is acquired, is called a scientific method. Scientific method consists of systematic observation, classification and interpretation of data.

**Forms of Scientific Method**

Following six primary forms of scientific methods used for research are :

1. **Historical Method** : Historical method is a backward movement in knowledge to trace the antecedent cause or causes of a phenomenon. Darwin's theory of evolution and Karl Marx's law of Economic Determinism are examples of this method.
2. **Comparative Method :** Analysis and comparison of two contemporary phenomenon is known as comparative method e.g. studies of two or more constitutions governments or political systems. The comparison is always undertaken in order to discover new facts or relations.
3. **Structural Method :** A study of what a thing is, what its outlines and significant features is a structural study. Anatomy is a structural study.
4. **Functional Method :** This method studies the processes and their causes i.e. 'how' and 'why' of a thing. Physiology is a functional study.
5. **Deductive Method :** Deductive method is based on the principle that out of same generalizations, some conclusions are drawn. In other words these are the decisions which are derived from general to the particular. In actual practices it means that some inferences are made from some general observations. This

method is used in testing hypothesis and in finding out of the material already collected or research findings already made still hold good or not. This method thus helps in revising previous knowledge. This research method is also useful and helpful in the sense that those researches who have no time or no capacity to undertake any deep and minute research study, can also conduct research with the help of this method.

6. **Inductive Method :** This method of research is just the opposite of deductive method. In this method some facts are first collected and on the basis of those facts some conclusions in the broader and wider sense are drawn.

## 1.2.4 Research Process

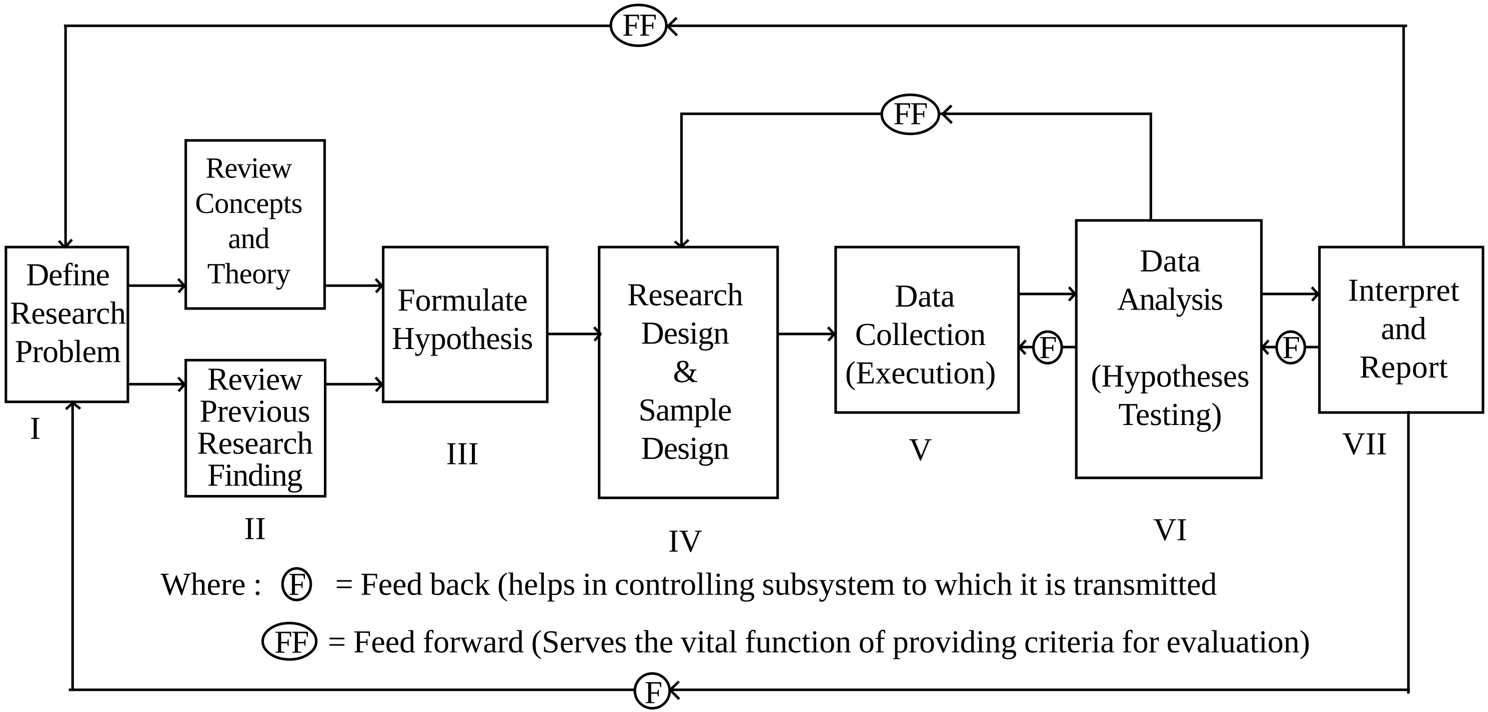
Before embarking on the details of research methodology and techniques, it seems appropriate to present a brief overview of the research process. Research process consists of series of action or steps necessary to carry out research and the desired sequencing of these steps. The chart given illustrates a research process.

The chart indicates that the research process consists of a number of closely related activities, as shown through I to VII. One should remember that the various steps involved in a research process are 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 hypothesis and research design; (4) determining sample design; (5) collecting the data; (6) execution of the project; (7) analysis of data; (8) hypothesis testing; (9) generalizations and interpretation and (10) preparation of the report or presentation of the results.

**(1) Formulating the Research Problem :** There are two types of research problems, viz., those which relate to states of nature and those which relate to relationships between variables. At the very outset the researcher must single out the problem he wants to study i.e., he must decide the general area of interest or aspect of a

**RESEACH PROCESS IN FLOWCHART**




Where : (F) = Feed back (helps in controlling subsystem to which it is transmitted

(FF) = Feed forward (Serves the vital function of providing criteria for evaluation) subject-matter that he would like to investigate into. Initially the problem may be stated in a broad general way and then the ambiguities, if any, related to the problem be resolved. Then, the feasibility of a particular solution has to be considered before a working formulation of the problem can be set up. The formulation of a general topic into a specific research problem, thus, constitutes the first step in a scientific investigation. 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.

The best way of understanding the problem is to discuss it with one's own colleagues or with those having some expertise in the matter. In an academic institution the researcher can seek the help from a research guide who is usually an experienced man and has several research problems in mind. Often, the research guide putforth the problem in general terms and it is up to the researcher to narrow it down and phrase the problem in operational terms. In private business units or in governmental organizations, the problem is usually earmarked by the administrative agencies with whom the researcher can discuss as to how the problem originally came about and what considerations are involved in its possible solutions.

The researcher must at the same time examine all available literature to get himself acquainted with the selected problem. He may review two types of literature i.e. the conceptual literature concerning the concepts and theories, and the empirical literature consisting of studies made earlier which are similar to the one proposed. The basis outcome of this review will be the knowledge as to what data and other materials are available for operational purposes which will enable the researcher to specify his own research problem in a meaningful context. After this the researcher rephrases the problem into analytical or operational terms i.e., to put the problem in as specific terms as possible.

The task of formulating a research problem is a step of greatest importance in the entire research process. The problem to be investigated must be defined unambiguously for that will help discriminating relevant data from irrelevant ones. Care must, however, be taken to verify the objectivity and validity of the back ground facts concerning the problem. Professor W.A. Neiswanger correctly states that the statement of the objective is of basic importance because it determines the data which are to be collected, the characteristics of the data which are relevant, relations which are to be explored, the choice of techniques to be used in these explorations and the form of the final report. If there are certain pertinent terms, the same should be clearly defined along with the task of formulating the problem. In fact, formulation of the problem often follows a sequential pattern where a number of formulations are set up, each formulation more specific than the preceding one, each one phrased in more analytical terms, and each one realistic in terms of the available data and resources.

2. **Extensive Literature Survey :** Once the problem is formulated, a brief summary of it should be written down. It is compulsory for a research worker writing a thesis for a M.Sc. and Ph.D. degree to write a synoptic outline of research topic and submit it to the appropriate Research Committee or the Research Board for approval. At this juncture 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 should 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 and Research Design :** After extensive literature survey, researcher should state in clear terms to 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. They also affect the manner in which tests must be conducted in the analysis of data and indirectly the quality of data which is required for the analysis. In most types of research, the development of working hypothesis plays an important role. Hypothesis should be very specific

and limited to the piece of research in hand because it has to be tested. The role of the hypothesis is to guide the researcher by delimiting the area of research and to keep him on the right track. It sharpens his thinking and focuses attention on the more important facets of the problem. It also indicates the type of data required and the type of methods of data analysis to be used.

How does one go about developing working hypotheses? The answer is by using the following approach :

(a) Discussions with colleagues and experts about the problem, its origin and the objectives in seeking a solution;

(b) Examination of data and records, if available, concerning the problem for possible trends, peculiarities and other clues;

(c) Review of similar studies in the area of the studies on similar problems; and

(d) Exploratory personal investigation which involves original fields interviews on a limited scale with interested parties and individuals with a view to secure greater insight into the practical aspects of the problem.

Thus, working hypotheses arise as a result of a prior thinking about the subject, examination of the available data and material including related studies and the counsel of experts and interested parties. Working hypotheses are more useful when are stated in precise and clearly defined terms. It may as well be remembered that occasionally we may encounter a problem where we do not need working hypotheses, specially in the case of exploratory or formalities researches which do not aim at testing the hypothesis. But as a general rule, specification of working hypotheses is another basic step in most of the research problems.

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 how all these can be achieved depends mainly on the research purpose. Research purposes may be grouped into four categories, viz., **(i) Exploration, (ii) Description, (iii) Diagnosis,** and **(iv) Experimentation.** A flexible research design which provides opportunity for considering many different aspects of a problem is considered appropriate if the purpose of the research study is that of exploration. But when the purpose happens to be an experimentation suitable design will be one that minimizes bias and maximizes the reliability of the data collected and analyzed.

There are several research designs, such as, experimental and non-experimental hypothesis testing. Experimental designs can be either informal designs (such as before-and-after without control, after-only with control, before-and-after with control) or formal designs (such as completely randomized design, randomized block design, Latin square design, simple and complex factorial designs), out of which the researcher must select one for his own project.

The preparation of the research design, appropriate for a particular research problem, involves usually the consideration of the following :

(i) the means of obtaining the information ;
(ii) the availability and skills of the researcher and his staff (if any);
(iii) explanation of the way in which selected means of obtaining information will be organized and the reasoning leading to the selection;
(iv) the time available for research; and
(v) the cost factor relating to research, i.e., the finance available for the purpose.

4. **Determining Sample Design :** All the items under consideration in any field of investigation constitute a 'Universe' or 'Population'. A complete enumeration of all the items in the 'population' is known as a census investigation. 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. Even the slightest element of bias in such an investigation will get larger and larger as the number of observations increases. Moreover, there is no way of checking the element of bias or its extent except through a re-survey or use of sample checks. Besides, this type of investigation involves a great deal of time, money and energy. Not only this, census investigation 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.

The researcher must decide the way of selecting a sample or what is popularly known as the sample design. In other words, a sample design is a definite plan determined before any data are actually collected for obtaining a sample from a given population. Thus, the plan to select 12 of a city's 200 drugstores in a certain way constitutes a sample design. 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 probability of selection of units in the sample such as deliberate sampling, simple random sampling, systematic sampling, stratified sampling, cluster and area sampling whereas non-probability samples are those based on convenience or certain purpose such as judgment sampling, quota sampling multistage sampling and sequential sampling techniques. A brief mention of the important sample designs is as follows:

(i) *Deliberate sampling* : Deliberate sampling is also known as purposive sampling or non-probability sampling. This sampling method involves purposive or deliberate selection of particular units of the Universe for constituting a sample which represents the Universe. When population elements are selected for inclusion in the sample based on the ease of access, it can be called *convenience sampling*. If a researcher wishes to secure data from, say, gasoline buyers, he may select a fixed number of petrol stations and may conduct interviews at these stations. This would be an example of convenience sample of gasoline buyers. At times such a procedure may give very biased results particularly when the population is not homogeneous. On the other hand, in *judgment sampling* the researcher's judgment is used for selecting items which he considers as representative of the population. For example, a judgment sample of college students might be taken to secure reactions to a new method of teaching. Judgment sampling is used quite frequently in qualitative research where the desire happens to develop hypotheses rather than to generalize to larger populations.

(ii) *Simple random sampling :* This type of sampling is also known as chance sampling or a probability sampling where each and every item in the population has an equal chance of inclusion in the sample and each one of the possible samples, in case of finite universe. It has the same probability of being selected. For example, if we have to select a sample of 300 items from a universe of 15,000 items, then we can put the names or numbers of all the 15,000 items on slips of paper and conduct a lottery. Using the random number tables is another method of random sampling. To select the sample, each item is assigned a number from 1 to 15,000. Then, 300 five digit random numbers are selected from the table. To do this we select some random starting point and then a systematic pattern is used in proceeding through the table. We might start in the 4th row, second column and proceed down the column to the bottom of the table and then move to the top of the next column to the right. When a number exceeds the limit of the numbers in the frame, in our case over 15,000, it is simply passed over and the next number selected that does fall within the relevant range. Since the numbers were placed in the table in a completely random fashion, the resulting sample is random. This procedure gives each item as equal probability of being selected. In case of infinite population, the selection of each item in a random sample is controlled by the same probability and those successive selections are independent of one another.

(iii) *Systematic sampling :* In some instances the most practical way of sampling is to select every 15th name on a list, every 10th house on one side of a street and so on. Sampling of this type is known as systematic sampling. An element of randomness is usually introduced into this kind of sampling by using random numbers to pick up the unit with which to start. This procedure is useful when sampling frame is available in the form of a list. In such a design the selection process starts by picking some random point in the list and then every *n*th element is selected until the desired number is secured.

(iv) *Stratified sampling* : If the population from which a sample is to be drawn does not constitute a homogeneous group, then stratified sampling technique is applied so as to obtain a representative sample. In this technique, the population is stratified into a number of non-overlapping subpopulations or strata and sample items are selected from each stratum. If the items selected from each stratum are based on simple random sampling the entire procedure, first stratification and then simple random sampling, is known as *stratified random sampling*.

(v) *Cluster sampling and area sampling* : Cluster sampling involves grouping the population and then selecting the groups or the clusters rather than individual elements for inclusion in the sample. Suppose some departmental store wishes to sample its credit card holders. It has issued its cards to 15,000 customers. The sample size is to be kept say 450. For cluster sampling this list of 15,000 card holders could be formed into 100 clusters of 150 card holders each. Three clusters might then be selected for the sample randomly. The sample size must often be larger than the simple random sample to ensure the same level of accuracy because in cluster sampling procedural potential for order bias and other sources of error is usually accentuated. The clustering approach can, however, make the sampling procedure relatively easier and increase the efficiency of field work, specially in the case of personal interviews.

*Area sampling* is quite close to cluster sampling and is often talked about when the total geographical area of interest happens to be big one. Under area sampling we first divide the total area into a number of smaller non-overlapping areas, generally called geographical clusters, then a number of these smaller areas are randomly selected, and all units in these small areas are included in the sample. Area sampling is specially helpful where we do not have the list of the population concerned. It also makes the field interviewing more efficient since interviewer can do many interviews at each location.

(vi) *Quota sampling* : In stratified sampling the cost of taking random samples from individual strata is often so expensive that interviewers are simply given quota to be filled from different strata, the actual selection of items for sample being left to the interviewer's judgment. This is called quota sampling. The size of the quota for each stratum is generally proportionate to the size of that stratum in the population. Quota sampling is thus an important form of non-probability sampling. Quota samples generally happen to be judgment samples rather than random samples.

(vii) *Multi-stage sampling* : This is a further development of the idea of cluster sampling. This technique is meant for large scale investigation extending to a considerably large geographical area like an entire country. Under multi-stage sampling the first stage may be to select large primary sampling units such as states, then districts, then towns and finally certain families within towns. If the technique of random-sampling is applied at all stages, the sampling procedure is described as multi-stage random sampling.

(viii) *Sequential sampling* : This is somewhat a complex sample design where the ultimate size of the sample is not fixed in advance but is determined according to mathematical decisions on the basis of information yielded as survey process. This design is usually adopted under acceptance sampling plan in the context of statistical quality control.

In practice, several of the methods of sampling described above may well be used in the same study in which case it can be called mixed sampling. It may be pointed out here that normally one should resort to random sampling so that bias can be eliminated and sampling error can be estimated. But purposive sampling is considered desirable when the universe happens to be small and a known characteristic of it is to be studied intensively. Also, there are conditions under which sample designs other than random sampling may be considered better for reasons like convenience and low costs. ***The sample design to be used must be decided by the researcher taking into consideration the nature of the investigation and other related factors.***

5. **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 and sufficient. 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 the hypothesis. But in the case of a survey, data can be collected by any one or more of the following methods :

(i) *Observation method :* This method implies the collection of information by way of investigator's own observation, without interviewing the respondents. The information obtained relates to what is currently happening and is not complicated by either the past behavior or future intentions or attitudes of respondents. This method is no doubt an expensive method and the information provided by this method is also very limited. As such this method is not suitable in investigations where large samples are concerned.

(ii) *Personal interviews method* : The investigator follows a rigid procedure and seeks answers to a set of pre-conceived questions through personal interviews. This method of collecting data is usually carried out in a structured way where output depends upon the ability of the interviewer to a large extent.

(iii) *Telephone interviews method* : This method of collecting information involves contacting the respondents on telephone itself. This is not a very widely used method but it plays an important role in industrial surveys in developed regions, particularly, when the survey has to be accomplished in a very limited time.

(iv) *Mailing of questionnaires method* : The researcher and the respondents do not come in contact with each other if this method of survey is adopted. Questionnaires are mailed to the respondents with a request to return after completing the same. It is the most extensively used method is various economic and business surveys. Before applying this method, usually a pilot study for testing the questionnaire is conducted which reveals the weaknesses, if any, of the questionnaire. Questionnaire to be used must be prepared very carefully so that it may prove to be effective in collecting the relevant information.

(v) *Schedules method :* Under this method the enumerators are appointed and given training. They are provided with schedules containing relevant questions. These enumerators go to respondents with these schedules. Data are collected by filling up the schedules by enumerators on the basis of replies given by respondents. Much depends upon the capability of enumerators so far as this method is concerned. Some occasional field checks on the work of the enumerators may ensure sincere work.

The researcher should select one of these methods of collecting the data taking into consideration the nature of investigation, objective and scope of the inquiry, financial resources, available time and the desired degree of accuracy. Through he should pay attention to all these factors are much depends upon the ability and experience of the researcher. In this context Dr. A.L. Bowley remarks that in collection of statistical data, commonsense is the prerequisite and experience is the main teacher.

6. **Execution of the Project :** Execution of the project 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.

    Occasional field checks should be made to ensure that the interviewers are doing their assigned job sincerely and efficiently. A careful watch should be kept for unanticipated factors in order to keep the survey as much realistic as possible. This, in other words, means that steps should be taken to ensure that the survey is under statistical control so that the collected information is in accordance with the pre-defined standard of accuracy. If some of the respondents do not cooperate, some suitable methods would be designed to tackle this problem. One method of dealing with the non-response problem is to make a list of the non-respondents and take a small sub-sample of them, and then with the help of experts regorous effects can be made for securing response.

7. **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 unwieldy data should necessarily be condensed into 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, specially in large inquiries, is tabulated by computers. Computers not only save time but also make it possible to study large number of 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 tests 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 or 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.

8. **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. Statisticians have developed various tests, such as *Chi-Square test, T-test, F-test*, for this purpose. The hypotheses may be tested through the use of one or more of such tests, depending upon the nature and object of research investigation. 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.

9. **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 trigger off new questions, which in turn may lead to further researches.

10. **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 following :

    The layout of the report should be as follows : (i) preliminary pages; (ii) main text and (iii) end matter.

    *In its preliminary pages the* report should carry title of report followed by acknowledgements and foreword. Then there should be a table of contents followed by a list of tables and list of graphs and charts.

    *The main text of the report* should contain following parts :

    (a) *Abstract of Research findings :* After introduction there would appear a statement of findings and recommendations in non-technical language. If the findings are extensive, they should be summarized.

(b) *Introduction :* It should contain a clear statement of the objective of the research and an explanation of the methodology adopted in accomplishing the research. The scope of the study along with various limitations should as well be stated in this part.

(c) *Main report :* The main body of the report should be presented in logical sequence and broken-down into readily identifiable sections. This includes review of literature, materials and methodology and results and discussion etc.

(d) *Conclusion :* Towards the end of the main text, researcher should again put down the results and conslusion of his research clearly and precisely. In fact, it is the final summing up.

*At the end of the report,* appendices should be enlisted in respect of all technical data. Bibliography, i.e., list of books, journals, reports, etc., consulted, should also be given in the end. Index should also be given for a published research report.

(1) Report should be written in a concise and objective style in simple language avoiding vague expressions such as ‘it seems,’ ‘there may be’, and the like.

(2) Chart's graphs and illustrations in the main report should be used for numerical data.

(3) Calculated ‘confidence limits’ must be mentioned and the various constraints experienced in conducting research operations may as well be stated.

## 1.2.5 Problems Encountered

Researchers in India, particularly those engaged in empirical research, are facing several problems. Some of the important problems are as follows:

1. **The lack of a scientific training in the methodology of research :** It is a great impediment for researchers in our country. There is paucity of competent researchers. Many researchers take a leap in the dark without *knowing* research methods. Most of the work, which goes in the name of research, is not methodologically sound. Research to many researchers and even to their research guides, is mostly a scissors and pasts job without any insight shed on the collated

materials. The consequence is obvious, viz., the research results, quite often, do not reflect the reality or realities. Thus, a systematic study of research methodology is an urgent necessity. Before undertaking research projects, researchers should be well equipped with all the methodological aspects. As such, *efforts should be made to provide short-duration intensive courses for meeting the requirement of research methodology.*

2. **Insufficient interaction between the university research departments :** There is very little or no interaction exists among research departments on one side and business establishments, government departments and research institutions on the other side. A great deal of primary data of non-confidential nature remains untouched and untreated by the researchers for wants of proper contacts. *Efforts should be made to develop satisfactory liaison among all concerned for better and realistic researches.* There is need for developing some mechanisms of a university- industry interaction programme so that academics can apply the research in industrial organisations.

3. **Reluctancy in providing information and data :** 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 reluctant in supplying the needed information to researchers. The concept of secrecy seems to be sacrosanct to business organizations. Thus, *there is need for generating the confidence that the information and data obtained will not be misused.*

4. **Research studies offonly overlapping one another :** This results in duplication and fritters away resources. This problem can be solved by proper compilation and revision, at regular intervals, of a list of subjects on which and the places where the research is going on. Due attention should be given toward identification of research problems in various disciplines of applied science which are of immediate concern to the industries.

5. **Non existance of a code of conduct for researchers :** It is observed that 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. **Difficulties in timely assistance :** Many researchers in our country also face *the difficulty of adequate and timely secretarial assistance,* including computer assistance. This causes unnecessary delays in the completion of research studies. All possible efforts be made in this direction so; that efficient secretarial assistance is made available to researchers and that too well in time. University Grants Commission must play a dynamic role in solving this difficulty.
7. **Library management and functioning is not satisfactory :** At many places library facilities are scanty 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. There is also the problem that many of our libraries are not able to get copies of old and new Acts and Rules, reports and other government publications in time. This problem is felt more in libraries, which are away in places from big cities and the state capitals. Thus, efforts should be made for the regular and speedy supply of all governmental publications to reach our libraries. There is also the difficulty of timely availability of published data from various government and other agencies doing this job in our country. Researcher also faces the problem on account of the fact that the published data vary quite significantly because of differences in coverage by the concerning agencies.
8. **Problem of conceptualization :** There are also problems relating to the process of data collection and related things in the process and tenure of research.

## 1.3 Glossary

**Research** : It is the activity of solving the problem through systematic investigation.

**Pure Research** : The search of truth makes persistent and partient to discover something new to enrich human knowledge is referred as pure research.

**Applied Research** : Application of known theories and models is referred as applied research.

**Survey Research** : In survey research samples are chosen form population to discover the relative incidence.

**Coordinated Research** : It is interdisciplinary research. It is characterized by exchange of view of idea from one discipline to another.

**Experimental Research** : It involves the procedure for gaining the knowledge by collecting new or fresh observations under control conditions.

## 1.4 Summary

Research is never treated as a piece of compilation work, but it is something original or a piece of work that advances human knowledge. Research is based on facts and therefore it is scientific undertaking which by means of logical and systematized techniques to discover new facts or verify and test old facts. The important objectives of research falls in broad grouping such as Development of Knowledge, Scientific Study of Social life, Welfare of Humanity, Classification of Facts and Social Control and Prediction.

The research can be classified broadly into two categories as Fundamental, which helps in developing principles, helps into find out the central factors in practical problems and become standard procedure for research to find out an answer to the problem. The other category is Applied Research which is based on the application of known theories and models. It is conducted to test impirical content or the basic assumptions or the validity of a theory under given conditions.

The research these days can be categorised as (a) Ex-post facto research (b) Experimental research (c) Field investigation research (d) Survey research (e) Case study research (f) Evaluation research (g) Action research and (h) Interdisciplinary research.

For any research the Hypothesis is formulated and then appropriate research design is prepared. The important sample designs are Deliberate sampling, Simple - sampling, Systematic sampling, Stratified sampling, Cluster and Area sampling, Quota sampling, Multistage sampling and Sequential sampling.

In order to examine the hypothesis the measurement is required which can be possible by collecting required data by the following methods such as observation, personal and telephone interviews, mailing questionnaires and filling schedules. Then the appropriate analysis of data is required for conclusions and report writing. In empirical research, the research has to face many problems such as scientific training, insufficient interactions, want of adequate data, library facilities and published data.

## 1.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. Define research and state and explain objectives of research.
2. Define and differentiate between fundamental research and applied research.
3. State and explain eight categories of research.
4. State and explain setps in the process of research with the help and flowchart of research process.
5. State and explain problems and difficulties faced by researchers from developing countries particulary in empirical research.

# Unit 2: Basic Principles in Research

***Index***



## 2.1 Introduction

Research usually involves as a step, a hypothesis or a set of hypotheses concerning the tentative explanation of a phenomenon or the solution of the problem Research is carefully recorded and reported. Every term is carefully defined, all procedures are described in detail, all limiting factors are recognized and all references are carefully documented.. Hence, it is essential to know the basic principle in research.

**After the study of this unit, you will be able to know and understand :**

- Motivating Factors in Research
- Scope of Research
- Limitation in Scientific Research
- Types of Research for Agriculture
- Formulation of working hypothesis
- Observation
- Formulation of Concepts

## 2.2 Content

### 2.2.1 Motivating Factors in Research

What makes the people to understand research ? This is a question of fundamental research. The possible motives for doing research may be either one or more. P.V.Young has mentioned four motivating factors in research.

1. **Curiosity about Unknown :** "Curiosity is an intrinsic trait of human mind and a compelling drive in the exploration of man's surroundings." It is natural instinct in the making. A man is always curious about the unknown and mysterious objects those he notices around him and tries to understand them in his own. The same curiosity drives a scientist to explore, reveal and understand unknown factors behind the social and scientific phenomenon.
2. **Desire to Understand the Relationship of Social and Scientific Problems :** According to P.V. Young, "The research of cause and effect relationship has been more relentless than almost any other scientist effort upon which human energies has been spent." People want not only the results of the events but also want to know how they have happened. What was their cause and what was their effect?. The research is undertaken to remove doubts and uncertainties, which results from misconceptions about social and scientific phenomenon.
3. **Appearance of New and Unexpected Situations :** In a modern complex and dynamic world a man is often faced with many acute and difficult problems. It is the duty of the scientist to find out their real cause and suggest solutions to such problems.
4. **Desire to Discover New and Test Old Scientific Procedures :** It concerns with the technique or methods used in social research. Social scientists have been busy in devising and developing new methods and techniques in place of old ones for dealing with social and scientific problems. In order to make research results more definite and precise, the use of quantitative or statistical methods is increasingly used now a days. In this connection, it is said that scientists are professional trouble shooters; they must challenge old beliefs, create new ones and then turn to the challenge upon those new ones.

However, this is not an exhaustive list of factors motivating people to undertake research studies. Many more factors such as directives of Government, employment conditions, curiosity about new things, desire to understand casual relationship, desire to get a degree through research, and the like may as well motivate people to perform research operation.

## 2.2.2 Scope of Research

Research provides an analytical framework for the subject matter of investigation. It establishes the relationship between different variables, especially the relationship of the dependent variables with the independent variables. The cause effect relationship between different variables can also be identified, leading to valuables observations, generalizations and conclusions. Induction and deductions are also possible in a systematic research. Induction into the process of inferring a general law from particular instances, i.e., a generalization is arrived at on the basis of the observation or result of particular instances. Deduction, on the other hand, is a way of making a particular inference from a generalization. In the deductive method, particular conclusions are derived from generalizations, while a generalization is made from particular solutions in the inductive method. Empirical studies have a great potential, for they lead to induction and deductions.

Research involves collection of data from primary and secondary sources and expresses them in quantitative terms. It stimulates the process or understanding on the one hand and deepens the insight on the other. The research should be objective and logical, applying every possible test to verify the data collected and the procedures employed. Purpose of research is to solve the human problems. Researches, which do not serve mankind, are of no use.

When we look at research in its modern setting, two things have to be taken into account viz., : (a) it must take stock of changed economic, political, social and cultural institutions; (b) is must substitute inductive reasoning for deductive one to make research more meaningful and acceptable academic contribution. The concept of research in its modern setting is held to be consisting of the following things:

(i) Avenue of quest and investigation;

(ii) Specific and well defined domain;

(iii) A well laid hypothesis;

(iv) Capable to use observable data;

(v) Scientific techniques of analysis; and

(vi) Logical interpretation of results.

Research should never be treated as a piece of completion work. By reading a number of books and compiling their material is yet another book is no research. Research is always expected to be something original or a piece of work that advances human knowledge. This may be done in several ways such as re-interpretation of an existing theory, investigation or an unexplored area of development of a new theory, etc. The main thing, therefore, is that a researcher should select a field of operation that offers possibility of inquisition or quest. Researcher must be clear in his mind as to what he is looking for.

The domain of research problems must be well specified and accurately defined. Those research investigations whose operational domain is not specific always remain inconclusive. It is always useful to select a smaller area of investigation and study it more intensively micro-level analysis are more precisely researchable than the macro-level studies.

The information of hypothesis is equally important for meaningful research. Hypothesis is an assertion that is sought to be proved or disproved. If there is no hypothesis, what a researcher is to prove or disprove, remains a big question. A well laid hypothesis also keeps the researcher on the right track and saves his time and effort by not going astray.

Research must be based on facts. Observable data forms a sound basis of research. Inductive investigations lend greater support to research findings. For analyzing facts a scientific methodology of analysis must be developed and results interpreted logically.

The scope of research in social sciences covers the disciplines of (i) Economics, (ii) Education, (iii) Management, (iv) Political Science, (v) Psychology, (vi) Public Administration, (vii) History, (viii) Sociology, (ix) Anthropology, (x) Demography (xi) Geography, (xii) Law, (xiii) Linguistics and (xiv) Agricultural Sciences and Rural Development.

### 2.2.3 Limitations in Scientific Research

Social and scientific research deals with social and scientic phenomenon respectively which is different from physical phenomenon. Man has complete control over physical phenomenon because it can be put to laboratory tests under guided conditions. Because of this fundamental difference we have to study the difficulties and problems that may be faced in the application of the scientific methods in investigation of social sciences.

Following are the main difficulties faced by the Researcher in the application of scientific methods in social research.

1. **Complexity of Social Data :** Researches in social sciences are based on human behaviour and the behaviour of human beings, is influenced by so many factors such as physical, social temperamental, psychological, geographical, biological and socio-cultural factors. Because of these factors an observer is generally confused. In the case of economic problems too the relations involved can be extremely complex on account of number of known and unknown variables.
2. **Problems in Interpreting Relationship Between Cause and Effect :** In case of a social phenomenon the cause and effect are interdependent and one stimulates the other. It is very difficult to establish cause and effect relationship in Social Sciences and to find as to what is the cause and what is its effect ?
3. **Problems of Concepts :** Problems of complexity of data lead to a further inherent problem in research. All the sciences struggle with complexity of real life and abstracts, crucial aspects of complex processes. The errors in social sciences might be result of a faulty reasoning also.
4. **Dynamic Nature of Social Phenomenon :** Human society is constantly changing and improving itself by past knowledge. Thus not only the deductions based on past studies, but even the techniques used in past may prove useless for the present or future studies. Like other social sciences it is true for Economics. Now the question arises what is the utility of studying a thing, which is so frequently changing. Its conclusions can hardly be used for gathering any advantage in near future. But a close study of the problem indicates that although human being are becoming more and more intelligent through experience, still

their fundamental instinct remains the same although their application in certain circumstances has undergone a change.

Whether the economic system is a person, a family or an entire nation there is bound to be inter-action with its environment. Some present variables of social phenomenon depend upon past and future. For example, in Economics consumption not only depends on present income but also on past income. But the study of lagged variables and framing of dynamic models solve this problem of dynamic nature of social phenomenon to some extent. On account of this dynamic nature of social phenomenon our task of analyzing data becomes very much complicated and the inferences drawn are sometime wrong.

5. **Problem of Maintaining Objectivity :** The problem of impartiality is in part a problem of objectivity. It is axiomatic that if a preposition is to be true, it must be arrived at objectivity. Any personal bias vitiates the universality criterion of a scientific proposition. It has been argued that the social scientists are less objective oriented than natural scientists because their own interests are affected by the findings of their studies, hence leading to prejudice and bias. To be objective in the conduct of an inquiry one should not let one's belief be influenced by his personal motives customs and social situation of which he himself its a part.

If a flower is red, everybody must view it as red. The viewer must not be permitted to wear colored glasses. In social sciences, however, personal or group bias often tends to creep in? Lundberg has quoted the first reason of it in the following words: "The physical phenomenon may be known directly through senses (eye, ear, nose, tongue, etc.) whereas social phenomenon are known only symbolically through words representing such phenomenon as welfare, tradition, customs, attitude, values and the whole realm of so-called subjective world."

The second reason of personal or group bias is that the social scientists being himself a part of the Universe that he studies finds it extremely difficult to get away from bias and he is also surrounded with customs and environment because of this lack of objectivity the verification becomes rather difficult.

London School of Economics often criticizes socialism on the following basis: “Socialism is bad at the first stage because it does not encourage initiative, at the second because it deprives the economy of the mechanism of the market, finally at the third because it is a road to serfdom. They always try to discover demerits in socialism and merits of capitalistic economy and thus objectivity is bound to appear in their thinking.

No doubt, achieving an effective degree of objectivity in a social inequity is a very difficult task but if we wish to proceed scientifically, we must find the best means of avoiding these biases. However, Prof. A.K. Dasgupta is of the opinion that “Personally, I think that a little bias in our thinking is healthy so far as it is explicit and not concealed. Our science deals with men in society and as social scientists we can ill-afford to dissociate ourselves from man’s desires and aspirations. A little passion brings meaning to our economic thinking. There is however one kind of bias against which an economic scientist has constantly to be on guard. He must not be taken in by consideration of his own interest. Objectivity in this sense implies disinterestedness.”

A classical instance of this kind of objectivity in economic theorizing is provided by the controversy between Prof. Malthus and Ricardo on rent. While Malthus being professor gives argument in favor of rent but Ricardo being a landlord puts argument against rent.

**6. Unpredictability :** Predictability is one of the most important characteristics of science. In case of physical sciences a remarkable degree of predictability has been achieved. The laws of physical sciences are universal and true for all times under given circumstances. But it is not so in the case of social sciences because of complexity of social data as well as due to the fact that social behaviour is sufficiently irregular and unpredictable. But the statement is partially correct because a social scientist can locate rough estimates of the behaviour of the whole group. As our knowledge of social phenomenon increases we are able to judge the effect of various variables involved. Thus it may be concluded that there is every possibility of the prediction of social events with sufficient accuracy.

7. **Difficulty in the Verification of the Inferences :** Verification of the results obtained is possible in the case of physical sciences but in the cases of social sciences it is much more difficult. In physical sciences with the help of experiments we can easily verify our results but in social sciences it is difficult to isolate the facts. The events of social sciences are non-repetitive and the social scientists are ill-equipped with their tools to verify prediction.

8. **Difficulties in the Use of Experimental Method :** Most of the physical sciences can be subjected to laboratory test under specified conditions which enable them the analysis and deductions more accurate. Besides this they can be tested and verified at any time. In case of social sciences such a facility is completely lacking. It is not possible to put human beings to laboratory test. Even if it is done their responses would not be natural but subject to the awareness of the artificial conditions. Thus, the social scientist has to watch them in wide world. He has to wait for the circumstances to occur in the natural course. They cannot be created for his studies as in the case of physical sciences.

    To quote Lundberg, "The social scientist will probably never be able to bring into his laboratory a piece of society there to observe its behavour in a test tube under varied conditions. He will have to study in wide world." The other difficulty is that human behaviour is not governed by particular cause. It is governed by particular situation as a whole.

9. **Incapability of being Dealt Through Empirical Method :** According to Lundberg, "Exact science tends to become increasingly quantitative in its units, measures and terminology while most part of social sciences is qualitative and does not admit to quantitative statement. We can talk of urbanization, culture etc., but we cannot measure them quantitatively." Direct qualification of socio-economic variable is not possible we can have only rough estimates.

    In case of Economics indirect quantification of variables is possible through dummy variables. In case of social sciences the data obtained are based on human behavior while in physical sciences the data used are obtained from experiments. Hence data are more reliable in case of physical sciences. The data obtained in social sciences vary simultaneously and therefore, it is not possible to keep some variables constant and other vary while in case of physical sciences influence of one set of data on the other can be found out. But in case of Economics the problem has been solved to a great extent with the help of least square method or other methods.

Empirical method gives very accurate results when experiment on social phenomenon is repeatedly carried on. But in case of social sciences repeated experimentation is not possible. Empirical methods are methods of statistics therefore, all problems, limitations, distrusts in statistical methods are also problems of social research. For example, the problem of unbiased sampling, selection of data, etc.

10. **Problem of Inter-Disciplinary Research :** Social research in any field are inter-related, therefore, we cannot draw watertight compartments for each sector of social sciences and we cannot say whether it is purely political research, economic research or purely sociological research. But in case of physical sciences it is possible to a very great extent to state whether it is physical or chemical or biological problem.

    The main problem in inter-disciplinary research is that every branch of knowledge has its own line of approach and a methodology suited for the purpose when these are tried to be fitted in a single frame, distortions are bound to take place.

11. **Less Finance :** Prof. Goode and Scates has pointed out that social researcher gets less finance than a researcher in physical sciences. In his own words, "The rate of progress in social sciences is much less than the rate of progress in physical sciences only due to lack of finance. There must be a clear and strong recognition of the need and urgency of all social researches and the main steps must be taken by various leaders of our society.

    Prof. Mitchell has also pointed out that qualitative worker hardly requires any equipment beyond a few books and hardly any helper except a typist. A quantitative worker needs often a statistical laboratory, computers and sometimes a staff of field workers.

To sum up, social sciences are less precise in its findings than the natural sciences because it deals with human society whose group as well as individual behavior have always been more diverse, full of more surprises and less predicates. Like natural sciences we do not have nice neat equations that will yield answers for all situations. Nevertheless, there are objective principles of social sciences that can be applied. More important, the study of social sciences can furnish us with an objective method of analysis, so that in dealing with social problems we need not remain more slaves to hozry slogans or prejudices. But it is encouraging to note that tendency towards the use of scientific method is fast growing in social sciences.

### 2.2.4 Types of Research for Agriculture

Research being a fact-finding process, profoundly influences decisions in agricultural management. The farmer is interested in choosing that course of action, which is most effective in attaining the maximum production. Research not provides only facts and figures, but enables one to choose a measuring road to judge the effectiveness of each decision.

The research relevant to agriculture is the basic, applied, adoptive social, financial and environmental research. This includes production and processing national policies towards production, trade, distribution, export, import, monopoly impact, labor etc.

**(a)** **Financial Management Research :** Research in financial management includes operations of specialized financial institutions like Primary Agricultural Co–operative Credit Society, District Central Co-operative Bank, Apex Bank, NABARD, Crop Insurance etc. A list of broad areas in finance is given below:

1. Financial Analysis
2. Rationing of Resources
3. Management of funds i.e. short, medium and long term loans
4. Subsidies for export and import
5. Impact of crop insurance

**(b)** **Production Research :** New technologies reduce costs and improve production and its quality. The information from production system may include technology components.

1. Development of new technology for different crops.

2. Impact assessment of technology on different crops.
3. Yield gap analysis and constraint analysis.
4. Impact assessment of government schemes for agricultural development.
5. Development of High Yielding Varieties.
6. Assessment area, production and productivity of cereals, pulses, oilseeds, fruit crops, vegetables, cotton, sugarcane etc.

(c) **Research in Marketing :** Marketing research deals with problems of production and distribution, marketing institutions, marketing policies and practices.

1. Marketing policies which includes pricing, inventory, advertising, channels of distribution, import and export.
2. New product ideas for processed food, fruits and vegetables.
3. Packaging, consumer's preferences, product standardization, trade marks
4. Market information system at national and international level.

(d) **Research in Agricultural Economics and Social Sciences**

1. Economic forecasting.
2. Technological innovations and progress.
3. Labor economics.
4. Socio-economic studies and impact of new agricultural technology..
5. The relevance of co-operation
6. Socioeconomic factors affecting co-operation.
7. Evaluation of schemes of co-operation.
8. Co-operatives and weaker sections.
9. Identification of strengths and weakness of co-operation.
10. Export through co-operatives.
11. Identification and need of training of co-operatives.
12. Evaluation of states policy on co-operation.

## 2.2.5 General Phases of Scientific Procedures

The Scientific procedures can be devided into following Phases

(a) **Formulation of Working Hypothesis :** Research means systematic method of innovating principles or discovering new facts on the basis of data. Collection of data or information for the sake of collecting them will yield no fruits; to be fruitful, one should collect such data or information for or against some point of view or proposition. Such a point of view or proposition is termed as hypothesis. The task of research is to test its accord with data or information. In other words, research needs formulation of tentative solution or answer on which, with the help of data or information the researcher can proceed for or against it. Thus, the proposition form a viewpoint, guides to see and select those that are relevant to the research he proposes to solve.

Hypothesis is a proposition temporarily accepted as true in the light of what is, at the time, known about a phenomenon. It is adopted as a tool for action in the search for truth. Lundberg remarks "the gathering data with a hypothesis recognizes the limitations of our senses and attempts to reduce their fallibility by limiting our field of investigation so as to avoid greater concentration of attention on particular aspects which are insignificant for our purpose". Thus, the function of a hypothesis is to direct our search for the order among facts. Goode and Hatt defined hypothesis as "a proposition which can be put to test to determine validity". Lundberg defines hypothesis as "a tentative generalization, the validity of which remains to be tested". In its most elementary stages, the hypothesis may be any hunch, guess, imaginative idea which becomes a basis for action or investigation". Palmer Johnson has beautifully described hypothesis as "islands in the uncharted seas of thought to be used as bases for consolidation and recuperation as we advance into the unknown".

**(1) Characteristics of a Workable or Usable Hypothesis**

In view of the importance and intended purposes of the hypothesis in the phenomenon of investigation, a workable or usable hypothesis must possess or satisfy the following criteria.

***The hypothesis should be Empirically Testable :*** A hypothesis should be empirically testable. It should be framed so that it makes possible to deduce logically certain inferences, which in turn can be tested. The concept embodied in the hypothesis must have empirical correspondence. For example, "good parents be get good children" is a proposition that can qualify as a workable hypothesis.

***The hypothesis should be Conceptually Clear :*** The hypothesis must be conceptually clear. In other words, the concept be used should be clearly defined both formally and operationally. The formal concept will clarify what a particular concept stands for, while the operational concept is the indicator of the concept in the field. It is noted that the concepts embodied in the hypothesis / hypotheses be defined in a way commonly accepted and communicable. This would go a long way in bringing about a cumulative growth of scientific knowledge.

***The hypothesis should be Specific :*** The formulation of hypothesis should not be too vague or general. Often the researcher is persuaded to express his hypothesis so general or with grandiose a scope that it is not able to test. Stated otherwise, this temptation is suicidal. A hypothesis for which suitable and clear statement is to be developed. Therefore, specific formulation has so obvious merit of confidence leading research into practicable and significant achievement specific hypothesis of real use.

***The hypothesis should be Related to the Body of Theory :*** As we studied that hypothesis is formed from theory so hypothesis must be related to it. In other words, the roots of hypothesis would endow hypothesis with power of prediction, which is one of the valuable attributes of good hypothesis. If the hypothesis is related to some theory, research will help to qualify, support, correct or refute the theory. The usual procedure that different scientists followed is select small but correlated problems and their findings are consolidated and coordinated into a broader theory. Against background of this theory, a test for the hypothesis may be framed.

***The hypothesis should be Related to Available Technique :*** Hypothesis should be related to the available technique. In other words, after knowing the technique that has been available and applicable to measure the concepts and the variables to be incorporated, then only the hypothesis can be formulated. Therefore, a researcher must know clearly the workable technique before formulating a hypothesis. Judging the hypothesis to its research ability needs the sensible requirement of related and workable techniques. For example, to study the degree of vertical social mobility, it cannot do so because no definite technique has yet been evolved for measuring this kind of mobility. Thus, it is clear that for being really workable, the hypothesis must be capable of being measured and tested according to existing technique.

***The hypothesis should be Simple :*** The hypothesis should be simple and to the point. Insight governs the essentials of simplicity. Prof. P.V. Young stated that "The more insight the researcher has into the problem the simple will be his hypothesis about it". Forming hypothesis should neither more or less but is necessary to account for the phenomenon. William Occam an English Philosopher described hypothesis as 'razor'. This means the hypothesis should be as sharp as razor's blade.

**(2) Utility of Hypothesis**

Though hypothesis occupies a small space in the body of a thesis or dissertation, it is as important as the research problem for investigation. A researcher cannot move with his investigation without a hypothesis or two. If he could not formulate a hypothesis, his investigation becomes aimless and fruitless a mere collection of data which is likely to lead him nowhere. Thus, the hypothesis is the foundation of scientific research. If a clear, simple, specific and scientific hypothesis has been developed, it accounts for one-fourth of the research work. The utility of hypothesis in social science research is stated below.

(i) Hypothesis acts as a guide. In other words, it is investigator's eye – a sort of guiding light in the journey of research.

(ii) It prevents blind research. It spells out the difference between precision and haphazard, between fruitful and fruitless research.

(iii) It provides direction to research, identifying of which is relevant and which is irrelevant. It prevents irrelevant review of literature and collection of useless or excess data.

(iv) It focuses research without it is like a random and aimless path.

(v) It links up related factors and information in fully understandable.

(vi) It serves as a framework for drawing meaningful conclusions. Direct answer to the hypothesis being tested.

**(3) Sources**

It goes without saying that the formulation or verification of hypothesis is a goal of scientific equity. A hypothesis may be developed or formulated from (a) scientific theory, (b) the findings of previous studies, (c) culture, (d) analogies, (e) personal experience, (f) pilot study, (g) Hunch, and (h) creative thinking and imagination of the researcher. Some of these are discussed vividly below.

***Scientific Theory :*** A hypothesis may stem from body of theory or scientific laws which is available to the field of research for putting it to test its validity. A generalization is drawn from the theory itself. When it has been tested and found correct it becomes a part of the theory. In other words, a theory is known fact, logical deductions from which constitute the hypothesis which must be true if the theory is true. Thus, theory is a fertile seed-bed of hypotheses. Dubin says “hypothesis is the feature of the theoretical model closed to the things observable that the theory is trying to model.”

***Previous Studies :*** The findings of earlier research works lead in the formulation of proposition. In other words, the researcher, based on the earlier research works, may hypothesize that similar or dissimilar relationship between the variables in the present study. For example, C.H. Hanumantha Rao in his study entitled “Agricultural Production Functions, Cost and Returns in India” found that “per acre net income decreases with increase in size of holding”. On the basis of this finding Viswanath in his study entitled “A Comparative Study of Irrigated and Un-irrigated Crops in Azamagarh (UP)” noted that “per acre net income increases along with the increase in size of holding”. This is a common way of the researcher who design his study with a view to replicating another study conducted in a different or similar context. Thus, the findings of one study may be formulated as hypothesis for the study which aims at and puts it to test. This is because many a studies in social science are exploratory in character. Thus, the hypothesis is based on the findings of other studies and if the present study support the hypothesis, then the result will help to confirm this relation as a repeatedly recurring one or otherwise explains the caustic reasons.

***Culture :*** The general culture has a great influence upon the thinking process of human beings. In solving value-oriented problems, culture cannot lose sight of. While formulating hypothesis attention should be paid to recognize the culture of people. For example, if education habits of Indian children are taken for study, religion or moral bias, type of the family, play interest, cultural heritage of children, etc. should be studied. Another example is human happiness. Happiness has been related to income, education, occupation, social class and so on. The cultural emphasis on happiness has been productive of a very much range of hypotheses. We can see that a large number of commonsense propositions or proverbs could serve as a source hypotheses.

***Analogies :*** Analogies are often a source of hypothesis. Reasoning is a very fertile source of hypothesis but reasoning by analogy generally is considered unacceptable as a source of proof. It is the process of ramming hypothesis from the likeness and similarity. It is assumed that similarities between the situations is not an accidental but is the result of the operation of some laws common to the both. For example, if Indian poverty is similar in nature to a poverty problem studied in a China, we may frame the hypothesis in a similarly manner.

***Personal Experience :*** Experience is a good source for hypothesis. Sometimes, the facts are there but a right person sees it in right perspective and formulates a hypothesis. For example, every body had seen the falling of an apple from the tree before Issac Newton but he was unique who could get the idea of the force of gravitation.

### (4) Formulation of Hypothesis

For ideally formulating the hypothesis for research, the researcher has made a systematic effort . It is learnt that the research problem reduces to a set of alternative means with which the investigator is in a position to pose the questions for each mean as to what would constitute the evidence for a particular mean is the most efficient one among the alternatives. Being a specific condition it is accepted as the most efficient among the alternatives. Such specific conditions are called acceptance conditions, which are the hypotheses. Simply stated, the alternative hypotheses are the statements of acceptance conditions to the problem. The researcher does not know which of acceptance conditions is true, this is precisely what the research is designed to determine. Thus, the formulation of hypothesis involves the steps mentioned as follows.

(a) A measure of efficiency applicable to all the alternative courses of action is selected.

(b) A set of acceptance conditions for each alternative course of action is assigned on the basis of selected measure of efficiency.

(c) The acceptance conditions are reformulated as hypotheses which are mutually exclusive.

The approach to hypotheses testing is not the constitution of a simple hypothesis about the population parameter, but rather to set up different hypotheses. These hypotheses must be constituted so that of one hypothesis is accepted or other is rejected and vice-versa. The hypotheses are normally referred to as:

**(i) Null hypothesis**

**(ii) Alternative hypothesis**

A reasoning for possible rejection of proposition is called 'null hypothesis'. It is a very useful tool in testing the significance of difference. H. E. Garrett remarks "the null hypothesis is akin to the logical principle that a person is innocent until he is proved guilty. It constitutes a challenge and the function of a research is to give facts a chance to refute this challenge". A null hypothesis is its other form asserts that the results found to be expected on a probability basis in terms of certain theory. As against null hypothesis, the alternative hypothesis is formulated embracing whole range of values rather than single point. For example, an alternative hypothesis may be stated that:

$H_1$ = The males visit cinema often than the females.

The Null hypothesis may be stated as:

$H_0$ = The males and females do not differ in respect of the frequency of seeing cinema.

The null hypothesis is more useful than other hypothesis because it is exact. It is easier to disprove the category of the hypothesis than to prove it with certainty. According to Karl Popper, the real basis of science is the possibility of empirical disproof and no amount of positive confirmation of a hypothesis.

***Observation :*** Analysis of questionnaire responses is concerned with what people think to do, as revealed by what they put on paper. Interview reveals what people think and do by what they express in conversation with an interviewer. Observation seeks to ascertain what people think and do by watching them in action as they express themselves in various situations and activities.

Observation is recognized as the most direct means of studying people when one is interested in their overt behaviour. Observation is a more natural way of gathering data. Restrictions imposed in questionnaire / interview are missing in observation. Artificiality and formality of questionnaire / interview is replaced with reality and informality in observation. Data collected through observation may be often more real and true than data collected by any other method. Whenever direct observation is possible, this is preferable method to use.

Observation is the most refined of modern research techniques. Direct observation has recently come to be looked upon as a scientific procedure. It is undoubtedly the first procedure of science, as all scientific data must originate in some experience or perception. As a scientific tool it may range from the most casual and uncontrolled to the most scientific and precise, involving modern mechanical and electronic means. It can be made progressively more scientific to meet the needs of the particular situation and observation is a fundamental tool even at the most advanced levels of science.

**(a) Characteristics for Observation for Research**

***(i)*** ***Scientific Observation is Systematic :*** It is not haphazard or opportunistic. Again it is not a change dropping in, when one happens to be passing by. The length of the observation periods, the interval between them, and the number of observations are carefully planned. There are dependable arrangements for controlling the situation if special factors are to be studied, e.g. studying of honest behaviour, sportsman sprit, leadership qualities, etc.

***(ii)*** ***Observation is Specific :*** It is not just looking around for general impressions. It is directed at those specific aspects of the total situation which are assumed to be significant from the standpoint of the purpose of the study. The layman, for example, frequently overlooks what is crucial while he devotes his attention to what is irrelevant. There are definite things for which a scientific observer looks. These things are carefully defined so that there is practically that nothing left to the discretion or preference of the observer. All items of observation are ironed out in the preliminary trials of the of the technique. The phenomenon is properly dissected beforehand and its significant aspects are correctly identified.

***(iii)*** ***Scientific Observation is Objective :*** It is free bias as far as possible . It should generally be guided by a hypothesis. However there is the risk that prejudgment on the part of the observer may color his perceptions and blind him to certain aspects of the actual situation . The observer who is convinced that under-achievers are lazy is likely to find many confirming instances. Prejudgment is a liability, particularly in the early stages of observation, where the observer must maintain maximum flexibility and open-mindedness. Although it is possible that a hypothesis may orient the investigator in the wrong direction, he cannot deal adequately with a complex situation if he simply looks at everything on an opportunistic basis.

The observer must maintain his neutrality, not only must he consider hypotheses as something to be tested rather than proved, but he must maintain a flexible attitude so that he can deviate from his original plans when such deviation appears advisable.

***(iv)*** ***Scientific Observation is Quantitative :*** Although many important phenomenon cannot be quantified, it becomes almost imperative to device some means of quantifying observations in order to increase their precision and to facilitate their analysis. Even the quality should be converted into quantity. Qualitative data is subjective and quantitative data is objective in itself and can be further interpreted in objective manner. This is an important characteristic of science.

Usually the number of instances of a certain type of behavior is recorded. Sometimes it is the total duration of the particular conduct during the period of observation. As far as possible the information is gathered in quantities, e.g. the distance traveled by a child, frequency of a good conduct, intervals of rest during a day, etc.

***(v)*** ***The Record of Observation is made Immediately :*** The findings are not entrusted to the memory, but notes are made as promptly as possible. If we rely on memory, the factor of forgetting will enter and affect it.

*(vi)* ***Observer is Expert :*** The expert observer knows precisely what he is looking for in the total situation. He not only structures the phenomenon he is to observe, but he also plans his observation to prevent his overlooking any significant aspect. He is aware of the pitfalls to be avoided and he has the background of experience both in research and in the problem area. He is adequately trained in the use of this technique.

*(vii)* ***Observation is Verifiable :*** Its result can be checked and substantiated. Observation must comply with the usual criteria of reliability, validity and usability.

## (b) Formulation of Concepts

Concepts are basic elements of scientific method but by and large all concepts are abstractions and represent only certain aspects of reality. In the words of Pauline Young, "Each new class of data, isolated from the other classes on the other basis of definite characteristics, is given a name, a label-in short a concept. A concept is in reality a definition in shorthand of a class of group of facts."

Theory is a set of generalizations, which are deduced from one another and are intimately inter-related. Theory is 'a set of generalizations that are deductively related'. Theory consists of logically inter-related concepts, which are combined into propositions. In a broader perspective, theory is not only logically inter-related concepts but it is a conceptual scheme because the processes of conceptualization are identified with theorization. According to Robert. Merton, the very effort to analyze and clarify the meaning of a concept can be termed as 'theorization '. Therefore, the process of theory building also starts with a series of inter-related concepts in a series of propositions which are given, shape of a conceptual scheme or theory. Now , it is clear that theory is a set of empirical generalizations which are deductively connected. To put it the other way deductibility is the most important characteristic of theory. For a set of statements to be appropriately called 'theory', it must necessarily have the requirement of being arranged deductively.

According to Scott Greer, theory is "a constructed view of aspects of the world from which regularities can be deduced. It is logically consonant with, and implies, known laws. It explains them by including them in a larger regularity and by ordering them in a pattern metaphorically familiar. And equally important, it allows us to predict hither known laws."

Theorems are statements which are logically deduced from axioms, serving as the conclusion of a deducible argument. Theorems are thus empirical statements deducible and related to axioms. It can be said that there is no theory unless there is a body of theory. Theories can be evaluated in terms of their testability, scope and capacity to explain and predict.

Theories are constructed to provide deeper understanding of observed regularities or potential laws. Theory can help understanding, analysis and explanation to social phenomenon. Theory is thus, regarded as knowledge of general belief systems or as a conceptual scheme of interpretation of point of views.

## 2.3 Glossary

**Curiosity :** Curiosity is an intrinsic trait of human mind and a compelling drive in the exploration of man's surroundings.

**Primary data :** Data which is collected first time and is original is known as primary data, e.g. data on field, data of survey etc.

**Secondary data :** Data which was collected by some person for other purposes and used by other person for their research and study is called as secondary data, e.g. data published in research journals, data published by the Government, data published in thesis etc.

**Hypothesis :** Hypothesis is the feature of the theoretical model closed to the things observable that the theory is trying to model.

## 2.4 Summary

It is essential to know the basic principles in research. The research provide an analytical framework for the subject matter of investigation. In modern setting the research should change economic, political, social and cultural institutions and it must substitute inductive reasoning to make research more meaningful and acceptable.

There are number of limitations in scientific research which the research has to face such as complexity of social data, interpreting relationship between cause and effect, problems of concepts, dynamic nature of social phenomenon, problem of maintaining objectivity, upredictability, verification, incapability of empirical method, problem of inter-disciplinary approach, lack of finance, etc.

Relevance in Agricultural research is the social, financial and environmental. The broad categories are Financial Management Research, Production Research, Marketing Research and Research in Agricultural Economic and Social Sciences. The research does not know which of acceptance conditions is true and what the research is designed to determine, which requires formulation of hypothesis. The hypothesis are normally referred to Null hypothesis or Alternative hypothesis.

The observation is a more natural way of gathering data and it is the most refined of modern research techniques. The important characteristics for observation for research are systematic, specific, scientific, quantitative or qualitative and verifiable. After this, theories can be evaluated or constructed to provide deeper understanding of observed regularities or potential laws.

## 2.5 Exercises for Practice

**Answer the following questions in 200 words each**.

1. What are the motivating factors for research ? Explain them in detail.
2. State and explain the limiting factors in scientific research.
3. Explain general phases of scientific research.
4. Define and explain hypothesis concept and theory.

# Unit 3: Defining the Research Problem

***Index***



## 3.1 Introduction

For a new research worker one of the difficult tasks is to decide upon a suitable problem wherein his most valuable time, money and other energies are to invest in the process of research. It is an area in which 'vision' plays an important part.

The first step in designing social research is to perceive a problem either theoretical or applied. It is researcher's perception or recognition of a problem that motivates research. A problem exists when we do not have enough information to answer a question. This may happen :

(1) When there is a negotiable gap in the results of investigations. Those questions which have remained unanswered by earlier investigators may make us aware of the problem. Collection of data with a view to fill this gap is thus indicated.

(2) When the results of several investigations disagree with each other.

(3) When a fact exists in the form of a bit of unexplained information. For

example, when the production targets are not being met or the cost of production is going up without being sufficient explanation for such development.

(4) When there is desire for innovation. For example, a manufacturer may think of conducting research in new methods of production, packing or sale even when there is no problem existing in any of these areas. He wants to search out something new and better.

The problem for research should ordinarily be expressed in an interrogative form.

**After the study of this unit, you will be able to know and understand :**

- How to select a research problem?
- Sources of this problem
- Steps involved in preparation of research proposal

## 3.2 Content

### 3.2.1 Research Problem and its Scope

The problem of identifying the problem is one, which is least amenable to the formal methodological treatments. To overcome the problem of methodological issues, the following preliminary observations may be of some help to anybody that who wishes to get started with research.

The striking and shocking aspect of higher education is its extensive passivity or intellectual non-involvement. Our system is working in such a way that it is possible for a student to complete the post-graduate course and take a brilliant first class depending solely on half-a-dozen textbooks. It is worth to mention here that "a fresh post-graduate student is problem-blind. He has eyes but he does not see. He has ears but he does not hear." So, it is pertinent that when a student with a master's degree turns to research there is a sudden and big discontinuity in the academic process, which put him excessively in oppressed state. And some time to the extent of not being able to proceed at all. The only remedy to his problem is developing a sense of academic curiosity in collecting material and processing it; and getting conclusions putting them to tests. This sort of exercise though not full but pre-occupied can be of great help to get one's mind at tuned to research. All this exercise constitutes the essence of what is called 'research'.

The second preliminary problem is related to whether the research problem be suggested by the supervisor or be selected by the candidate himself. On this issue, there are no hard and fast rules which can be laid down in general terms except that if both the supervisor and the candidate are sustained their interest in a topic acceptable to both. Neither the supervisor alone suggests a problem to the candidate nor the candidate himself welcomes it from the supervisor. In the sense that the imposition by the supervisor and the dependency by the candidate destroy the elements of spontaneity and authenticity which must mark research work. A supervisor can at best suggest a topic for research and from which a research problem has to be chosen and it can only be the responsibility of the one whom actually does the research.

Identifying the research problem is a long drawn process for those who are new to research. After demarcating the broad area in which one would like to do research, he must at least posses two considerations, viz., (1) interest and (b) competence. To do sustained work on a problem and to put one's intellectual efforts fully into kit is a difficult task and will turn out to be a drudgery unless one has a genuine interest in the problem itself. Interest in the subject is being developed by various reasons. Example, an agricultural background student may have interest in some area related to agriculture. On the contrary, he may avoid agriculture altogether. Academic interest in a particular branch of subject may be the result of reading a good book or because of the teacher dealt with it extremely well. However, the interest of the candidate has come about and which makes sure that the area of work conforms to one's general inclinations. Interest and inclination must be matched by competence before one can decide on an area of research. Competence can certainly be cultivated and learning new techniques is part of any research programme.

The common tendency in the approach to research is to feel that is must result in some extraordinary 'findings'. A young researcher is often tempted to think of himself as a potential Prof. M.S. Swaminathan. Such professional ambitions may be good up to a point but it is worthwhile to recall that a very few of the stalwarts in the profession have achieved fame through a doctoral dissertations. Therefore, ambition is good but too much ambition is dangerous because it may lead to the rejection of manageable problem in search for the brilliant and path-breaking contribution.

Once the main area of research is decided upon, the next step is to narrow it down to a topic. In fact, the supervisor can be and should be of greatest assistance. A good grasp of the branch of the subject as a whole and a detailed acquaintance with work already done in it are essential to suggest a topic for research. The researcher cannot be a passive recipient of suggestions but to know the field himself. After the survey of literature, the area of research problem has to be narrowed down. It means select an area on which one can get the guidance of a few competent articles or books, but not of too many. In brief, identifying a research problem is an active but subjective process, and possibly the most important aspect of research.

Out of the most difficult phase of any research project is the choice of a suitable problem. The beginner is likely to take a very long time in making his choice. In the first step of any research the worker should not take a hasty decision. Every problem that comes to his mind or even that suggested by a more experienced person may not be a fit research problem. The identification of a good research problem should be considered a discovery in itself.

The selection of a topic for research is only half a step forward. This general topic does not help a researcher to see what data are relevant to his purpose, what method he should employ in securing them and how he would organize these. Before he can consider these aspects, he needs to formulate a specific problem. The problem defines the goal of the researcher in clear terms. It is obvious that without a clear-cut idea of the goal to be reached, research activities would only become a meaningless exercise. A research like any other activity is goal directed. If the goal itself is unknown or ill defined, the whole research operation will lead us nowhere. Thus, without a problem research cannot proceed because there is nothing to proceed from and proceed toward. There is nothing but wisdom a in the saying: “If you start from nowhere, you will generally reach there”.

The formulation of the topic into a research problem is a, really speaking, the first step in a scientific inquiry. A problem in simple words is some difficulty experienced by the researcher in a theoretical or practical situation. Solving this difficulty is the task of research.

Let us now appreciate what experiencing a difficulty in a theoretical situation means. Observations are not fitting the theoretical expectations, e.g. a theory may predict that particular type of societies will have a low rate of suicides but observations do not substantiate this prediction, gives rise to a problem in theoretical situation. This is just one incidence of difficulty or problem faced in theoretical situation. A difficulty in practical situation may be felt, for instance when the application of increased dose of Nitrogen gives increase in yield but observation does not substantiate this prediction.

### 3.2.2 Selection of Research Problem

The selection of research problem is very important job for a researcher. The range of potential topics for social research is as broad as the range of social behavior itself. As usually the case every problem may not be researchable and, therefore, if the researcher has not been able to select the problem properly, his efforts may not yield the desired results and in some cases it may lead to frustration and desperateness of the researcher.

The selection of research problem depends on whether research is being conducted as a requirement of a degree or it is for academic interest. If the research is only to fulfill some requirements for obtaining a degree, the problem may be more specific and limited in scope and may offer itself for completion within a specified time. If the researcher fails to select such a problem as would be completed in a reasonable amount of time, he would often run into difficulties and may go without the desired degree. On other hand, if the researcher is of academic interest, time should not be binding factor. Many a times entire lifetime of researcher goes towards exploring something new through his research efforts. He does not want any research degree but only does the research with the sole objective of enriching his personal knowledge as well as that of the general academic world.

The selection of problem as requirement for degree further depends on the level of research or learning. Many a times, research, problems are selected at Masters Degree level, M. Phil. or Ph.D. levels but it is obvious that the levels of learning at these stages are different and, therefore, demand different treatments to research problems. Usually the problem selected at initial stages such as M.Sc. degree level constitutes an exploratory learning process. At the next stage such as Ph.D. level, it may aim at application of advanced research methods and seek to make significant contribution to advance the existing knowledge.

Not only selection of problem is important but also its statement is equally important. If the problem is not carefully and scientifically stated, it may lead to confusion regarding the specific interest for which the research investigation has been undertaken.

**Sources of Research Problem**

The research problems may be selected from the following sources.

(i) Theory of one's interest;

(ii) Daily problems;

(iii) Technological changes;

(iv) Unexplored areas; and

(v) Discussions with supervisors.

A researcher may select a problem for investigation from a given theory in which he has considerable interest. In such situations the researcher must have thorough knowledge of that theory and should be sufficiently inquisitive to explore some unexplained aspects or assumptions of the theory. Research problem can also be selected on the basis of daily experience of a researcher. Every day problems constantly present something new and worthy of investigation and it depends on the sharpness of researcher's intellect to knit his daily experiences in to a research problem.

Technological changes in a fast developing society are constantly bringing forth-new problems and new opportunities for research. What is the impact of a changed technology on the existing socio-economic set up, always interests the researcher and tempts him to undertake such studies as are revealing regarding the impact of new technology on the existing system.

Research problems can be both abstract and have applied interest. These may also be selected from those areas, which have not been explored so far; such areas may be theoretical or empirical in nature. For example, an economic analysis of tribal development schemes in Madhya Pradesh is a useful piece of research. In a similar fashion many geographical regions about which no knowledge exists can easily be taken up for research purpose.

Sometimes the researcher while discussing his interests with his proposed supervisor may come across a problem that can be researched by the investigator. The problem may relate to any source as discussed above. In the same way reading assignments in textbooks, special assignments, research reports and term papers may also suggest some additional areas which needs research. Many research articles suggest problems for further investigation that may prove fruitful. Such specialized sources as the, Encyclopedia of social sciences, dissertation abstracts, research abstracts and similar other publications are rich sources for problem seekers.

### 3.2.3 Techniques Involved in Defining a Problem

Let us start with the question: What does one mean when he/she wants to define a research problem? The answer may be that one wants to state the problem along with the bounds within which it is to be studied. In other words, defining a problem involves the task of laying down boundaries within which a researcher shall study the problem with a pre-determined objective in view.

How to define a research problem is undoubtedly a Herculean task. However, it is a task that must be tackled intelligently to avoid the perplexity encountered in a research operation. The usual approach is that the researcher should himself pose a question or in case someone else wants the researcher to carry on research, the concerned individual, organization or an authority should pose the question to the researcher and set-up techniques and procedures for throwing light on the question concerned for formulating or defining the research problem. But such an approach generally does not produce definitive results because the question phrased in such a fashion is usually in broad general terms and as such may not be in a form suitable for testing.

Defining a research problem properly and clearly is a crucial part of a research study and must in no case be accomplished hurriedly. However, in practice this is frequently overlooked which causes a lot of problems later on. Hence, the research problem should be defined in a systematic manner, giving due weightage to all relating points. The technique for the purpose involves the undertaking of the following steps generally one after the other: (i) statement of the problem in a general way; (ii) understanding the nature of the problem; (iii) surveying the available literature; (iv) developing the ideas through discussions; and (v) rephrasing the research problem into a working proposition.

A brief description of all these points will be helpful.

***(i) Statement of the problem in a general way :*** First of all the problem should be stated in a broad general way, keeping in view either some practical concern or some scientific or intellectual interest. For this purpose, the researcher must immerse himself thoroughly in the subject matter concerning which he wishes to pose a problem. In case of social research, it is considered advisable to do some field observation and as such the researcher may undertake some sort of preliminary survey or what is often called *Pilot survey*. Then the researcher can himself state the problem or he can seek the guidance of the guide or the subject expert in accomplishing this task. Often, the guide puts forth the problem in general terms, and it is then up to the researcher to narrow it down and phrase the problem in operational terms. In case there is some directive from an organizational authority, the problem then can be stated accordingly. The problem stated in a broad general way may contain various ambiguities, which must be resolved by cool thinking and rethinking over the problem. At the same time the feasibility of a particular solution has to be considered and the same should be kept in view while stating the problem.

***(ii) Understanding the nature of the problem :*** The next step in defining the problem is to understand its origin and nature clearly. The best way of understanding the problem is to discuss it with those who first raised it in order to find out how the problem organically came about and with what objectives in view. If the researcher has stated the problem himself, he should consider once again all those points that induced him to make a general statement concerning the problem. For a better understanding of the nature of the problem involved, he can enter into discussion with those who have a good knowledge of the problem concerned or similar other problems. The researcher should also keep in view the environment within which the problem is to be studied and understood.

***(iii) Surveying the available literature :*** All available literature concerning the problem at hand must necessarily be surveyed and examined before a definition of the research problem is given. This means that the researcher must be well - conversant with relevant theories in the field, reports and records as also all other relevant literature. He must devote sufficient time in reviewing of research already undertaken on related problems. This is done to find out what data and other materials, if any, are available for operational purposes. "Knowing that data are available often serves to narrow the problem itself as well as the technique that might be used.".

This would also help a researcher to know if there are certain gaps in the theories, or whether the existing theories applicable to the problem under study are inconsistent with each other, or whether the findings of the different studies do not follow a pattern consistent with the theoretical expectations and so on. All this will enable a researcher to take new strides in the field for furtherance of knowledge i.e., he can move up starting from the existing premise. Studies on related problems are useful for indicating the type of difficulties that may be encountered in the present study as also the possible analytical shortcomings. At times such studies may also suggest useful and even new lines of approach to the present problem.

***(iv) Developing the ideas through discussions :*** Discussion concerning a problem often produces useful information. Various new ideas can be developed through such an exercise. Hence, a researcher must discuss his problem with his colleagues and others who have enough experience in the same area or in working on similar problems. This is quite often known as an *experience survey.* People with rich experience are in a position to earlier the researcher on different aspects of his proposed study and their advice and comments are usually invaluable to the researcher. They help him sharpen his focus of attention on specific aspects within the field. Discussions with such persons should not only be confined to the formulation of the specific problem at hand, but should also be concerned with the general approach to the given problem, techniques that might be used, possible solutions, etc.

***(v) Rephrasing the research problem :*** Finally, the researcher must sit to rephrase the research problem into a working proposition. Once the nature of the problem has been clearly understood, the environment within which the problem has got to be studied has been defined, discussions over the problem have taken place and the available literature has been surveyed and examined, rephrasing the problem into analytical or operational terms is not a difficult task. Through rephrasing, the researcher puts the research problem in as specific terms as possible so that it may become operationally viable and may help in the development of working hypotheses.

In addition to what has been stated above, the following points must also be observed while defining a research problem:

(a) Technical terms and works or phrases, with special meanings used in the statement of the problem, should be clearly defined.

(b) Basic assumptions or postulates (if any) relating to the research problem should be clearly stated.

(c) A straight forward statement of the value of the investigation (i.e., the criteria for the selection of the problem) should be provided.

(d) The suitability of the time-period and the sources of data available must also be considered by the researcher in defining the problem.

(e) The scope of the investigation or the limits within which the problem is to be studied must be mentioned explicitly in defining a research problem.

### *Criteria of a Good Research Problem*

A good problem selected by the researcher should be conforming to the following criteria.

***(a) Relation between variables expressed :*** The problem should express a relation between two or more variables.

***(b) Clarity and Unambiguity*** **:** The problem should be related clearly and unambiguously in question form.

(c) ***Empirical Verification :*** The problem should be such as to imply possibilities of empirical testing. This means not only that an actual relation as stated, but also that the variables of the relation can somehow be measured.

**(d)** ***Availability of Guidence*** **:** Every research activity needs the patronage of a guide and the approval and sanction of competent authority.

***(e)*** ***Level of research*** **:** It is another criteria to help in the selection of problem. The nature and the scope of a study will be determined in the light of levels like master's degree, M. Phil. and Ph.D. It may simply be an action research or a research to produce a research paper or an experimental project.

***(f)*** ***Experience and creativity :*** Good research problems stem from a clear understanding of the theoretical, empirical and practical aspects of the subject derived from personal experience and from a thorough review of the literature. Conversely, lack of familiarity with the subject is almost sure to result in a poor choice.

***(g)*** ***Novelty :*** It should be sufficiently original so that it does not involve objectionable duplication. Ignorance of prior studies may lead a researcher to spend time on a problem already investigated by some other worker. Moreover the study should employ the most recent data. While originality is an important consideration, the fact that a problem has been investigated in the past does not mean that it is no longer fit for study. There is constant need for verification of the findings of previous investigations, using newer and better devices and procedures. There is also need for testing of former findings under changed cultural conditions.

***(h)*** ***Interesting*** **:** The problem should be interesting for the investigator himself. If he is not interested in it, he will not be able to face and overcome the obstacles, which come at every step in research. He should have a strong inherent motivation for it otherwise there is little hope that he would do justice to it.

***(i)*** ***Importance :*** It should be significant enough and involve an important principle or practice. If it were not worthwhile, neither adds to knowledge nor leads to any improvement in the current practices; it would be in vain.

*(j)* ***Availability of data :*** The research worker should ensure the availability of valid and reliable data gathering devices and producers. In case the study demands confidential and sensitive and classified information, will it be possible for him to obtain it?

*(k)* ***Feasibility :*** The problem for a particular research worker is the matter of its feasibility. The original research does ensure and have practical values to meet society’s demand.

## 3.2.4 Research Proposal Steps

A good research design consists of the following components.

***(1) Title of the Study***

The title of any research project

- Should be as brief as possible.
- Should be as precise as possible.
- Should project the scope of the problem in generalized terms.

***(2) Statement of the Problem :*** The researcher should state his research problem in unambiguous and more precise terms. While stating the research problem the usage should be clear, simple and concise which are of most preferable. A problem well put is in half-selves.

***(3) Review of Literature :*** A brief survey of studies under-taken earlier has to be given. This is meant to fill up the gaps in the contribution of knowledge. Such survey would entail a study of a relevant literature. The review has to be done in such a manner that when the proposed study is completed, it would add its own contribution to the quantum of knowledge in the subject concerned. Earlier studies are not only reviewed but also critically appraised, then only it is more meaningful, useful and a correct approach in any field of investigation. In other words, this type of review would pinpoint the stage from where further research is called for.

(4) ***Area and Scope of the Study :*** Scientific projects differ widely in their geographical, temporal and functional dimensions. However, it will be better and purposeful if individual researchers do micro studies that may spotlight the dark and obscure lacunae in the data. As far as the time dimension is concerned benchmark surveys related to a particular data is undertaken. Generally, surveys cover a period of time with the object of studying the trend of some variable or other. For example, a study meant to assess the impact of credit on socio-economic improvement of farmers in Rayalaseema region in Andhra Pradesh could cover a minimum period of ten years.

The functional dimension refers to a research project concerning with a particular industry or section of the population covering certain functional variables. Besides the above, the scope of the investigation is depended on several aspects such as money available to the researcher and availability of sample, etc.

(5) ***Objectives of the Study :*** The main purpose of the research study is to lay down the objectives precisely. It may be mentioned in the form of questions to be answered, or it may be in the form of an explanation to a particular issue or phenomenon. In other words, the objectives should be worded in a lucid and precise form. For example, if the investigation is on interlocking or leadership in a village organizations, the objectives may be :

- Assessing the extent of interlocking leadership.
- Identifying the socio-economic characteristics of such interlocking leaders.
- Evaluating the effects of such interlocking leadership on the growth of institutions involved.

(6) ***Formulation of Hypothesis :*** Hypothesis is a tentative solution of the problem faced by the scientist. It is usually framed depending on one or more of the following sources :

- The facts established by previous investigations in the related areas.
- The history of science.
- The analogies.
- The findings of other studies.
- The body of theory.

In farming the hypothesis, the investigator has to resort to null hypothesis which means possible reasoning rejection of the variables under the study. A hypothesis should be empirical, conceptually clear, specific, close to things observable and related to the body of theory.

*(7)* ***Definition of Concepts :*** The researcher needs to know clearly the meaning and contends of every term that he uses. It clarifies the issue as well as explains the investigators, respondents and other readers of the research report specifically and pointedly. The term concept must be defined both (a) in abstract term giving general meaning it is intended to convey. This may be called the 'formal definition'. Formal definition is necessary in order to link the study with the body of knowledge using similar concept, (b) in operational term by which it will be represented in a particular study. This type of concept is known as 'operational definition' (some times, the workable definition). The operational definition helps in collection of data and in carrying out all research.

Care must be taken in defining concepts. Every definition is the product of the purpose for which and the environment in which the enquiry launched. This is generally accepted in all types of research projects. To illustrate the operational definition of the concept, size of the land holding operationally defined as that the land owned by a cultivator minus the land area leased out plus the land area leased in. All important terms and concepts, like the above, in a study should be operationally defined. New definitions and new concepts are put in all enquiries, precedents and past practices, which deserve respect.

*(8)* ***Methodology :*** The decision pertaining to method of investigation is to be done in this section. The worker has to decide whether the research project is going to be empirical in the sense of being based on primary data collected from the field, or analytical in the sense of being based on secondary data collected from published reports. In both cases research analysis will be crucial. Further, the researcher has to decide the method, either census method or sampling method to be adopted in the study with justification and the reasons for the adopted in the study with justification and the reasons for the same.

Census method involves comprehensive or complete enumeration. If a survey covers all the units of the universe, under investigation, it is called census. Example of census survey is livestock. This method is time consuming and expensive process. The method of sampling has gained increased popularity in recent times. The study of a few representative individuals or units is called a sample study. In social science research technique both the census and the sample survey methods have their relative merits, and the researchers can adopt the suitable one, according to nature and purpose of the given research project besides time and money factors.

***(9)*** ***Sampling Design :*** Complete coverage of a population on a scientific enquity is never possible nor advisable, even if it is possible. It may be possible when the population is very small. What is to be done is to take a sample from the population, which is numerically adequate and culturally representative. While selecting sample units, the worker should take utmost care in adopting particular criteria for stratification. The researcher should keep in mind the following facts while selecting sampling method :

- The definition of the universe or population.
- The size of the sample.
- The representativeness of the sample

***(10)*** ***Constructing of Schedule of Questionnaire :*** A questionnaire or schedule refers to the tool or instrument of data collection. The data collected thus, become the empirical material for testing the hypothesis/hypotheses of the study on hand.

The terms questionnaire and scheduled are treated as synonyms in social science research. Technically, a difference between them existed not in the format or contents but in the way of response obtained from the respondents with the help of them. In other words, questionnaire or schedule consists of a number of questions printed or typed in a order on a form. In case of questionnaire the responses to the questions are to be filled up by the interviewee himself. In case of schedule the responses to the questions are to be noted down by the interviewer. The questionnaire is usually associated with mail and therefore it is called mailed questionnaire. This method is used if the respondents are educated.

The form of questions may be either close or open. Open-ended questions are designed to permit a free response from the subject while the closed questions have the advantages of being standardizable or categorical. In other words, the closed questions may require the respondents himself to make a judgment about his attitude rather than leaving this to the interviewer. The questions to be asked should be direct bearing on the problem, avoiding the personal questions. The wording a questions is to be simple and should exact information that is needed avoiding multiple meaning and emotional connotations.

***(11) Collection of Data :*** The subject matter, the unit of enquiry and the scope of the study administer the choice of method for collecting data. A study of behaviour of a group would call for observational technique. For a simple enquiry among the class people a questionnaire is adequate. A survey of general population entailing many complicated questions would call for personal interviewing.

***(12) Analysis of Data :*** After collecting data, the classification and analysis are to be taken up. Depending on the nature of the data, and information required to fulfil the objective or hypothesis the researcher should analyse the data subject to the appropriate statistical analysis besides tabulation. Tabulation of results in a meaningful way is by itself a technique and an art. The data given in the tables must be in self-explanatory form. Tabulation may be done either by manual or electronic method.

***(13) Interpretation of Results :*** After organizing the data into a meaningful form the researcher should draw inferences. Drawing inferences or interpreting data will call for statistical inference based on usual test for significance. The interpretation of the results of an investigation is to be related :

(a) To the previous findings of the study.

(b) To a wider field of generalizations.

(c) To scientific objectivity.

To uncover any additional factors which would not be visualized by the investigator earlier.

***(14) Reporting the Findings :*** No research is an end in itself. The investigator must report of what he discovered or innovated based on the data collected to fulfill the need for which the study taken up and to ensure proper directions to other investigators in carrying out of similar researchers. Reporting the findings must be clear, specific, simple and directly relating to the objective of the study. The method of the report is given below for arriving at a clear picture of the report.

**Format of the Report**

(1) Preface.

(2) Introduction to the problem.

  (a) Statement of the problem.

  (b) Purpose, scope and method of investigation.

  (c) Objectives of the study.

  (d) Hypothesis/hypotheses.

  (e) Definition of concepts.

  (f) Limitations of the study.

(3) Review of literature.

(4) Analysis and interpretation of data.

  (a) Organization of data : Editing, Coding, Classification and Tabulation.

  (b) Statistical analysis.

(5) Discussion of results.

(6) Summary and conclusions.

(7) Recommendations.

(8) Reference.

  (a) Appendix.

  (b) Schedule or questionnaire.

  (c) Bibliography.

***(15) Time and Financial Budgeting :*** In carrying out the research, the time factor is very important, otherwise the work carried by the investigator and the findings of the study become a mere waste. That means research is meant to solve the problems of the society in the sense the researcher has to respond quickly and in time. Apart from time factor, the financial factor is also a governing one in carrying out the research successfully.

***Advantages of Research Design***

(1) The research design may result in the desired method of study with useful conclusions.

(2) It may lead to more accurate results or help to reduce inaccuracy.

(3) It may give optimum efficiency and reliability.

(4) It may minimize the wastage of time as well as money.

(5) It may be helpful for the collection of material and testing hypothesis/ hypotheses.

(6) It is a guide-post for carrying research in the right direction.

## 3.2.5 Illustration of Research Problem

The illustration of Integrated Pest Management verification project is given as under:

(1) Title of the Project : IPM Technology Verification Project on Cotton.

(2) Location : Village-Ashta, Taluka-Kinwat Dist. Nanded

(3) Name of Investigator : Mr. Vilas Sawant

(4) Area under project : IPM treatment : 100 ha.
Non-IPM : 30 ha.

(5) Year of Start : 2000

(6) Year of completion : 2003

(7) Treatments :

**IPM treatments**

(a) Clean-up campaign in the month of April.

(b) Seed treatment with Imidacloprid 76 WP @ 7.5 gms / kg of seed before sowing.

(c) Intercrop : Maize + Cowpea as border crop around the field. A row of *Sataria etallica* (Minor millet) in between each 10th and 11th row of Cotton.

(d) Release of Tricogramma Chilonis @ 1.5 lakhs / ha at 40 days after sowing.

(e) Spraying of 5 % Neem Seed Kernel Extract at 45 days after sowing.

(f) Release of Tricogramma Chilonis @ 1.5 lakhs per hectare at 50 days after sowing.

(g) Spraying of 5 % Neem Seed Kernel Extract at 55 days after sowing.

(h) Spraying of Heliothis armigera Nucleo Polyhydrosis Virus (HaNPV) 250 Larval equivalent (LE) per hectare at 65 days after sowing.

(i) Installation of Pheromone traps from 30 days after sowing.

(j) Hand collection of larvae.

Farmer's Practices ( Non-IPM )

Use of chemicals as per farmer's choice.

(8) Type of research : Survey Research

(9) Objectives :(a) To study impact of IPM technology on the productivity of cotton.

(b) To estimate research and technology development costs, quantify the benefits from the crop and transfer of technology.

(10) Practical Utility : To justify future financial support on sustainable basis, it is imperative to evaluate the impact of IPM technology of Cotton productivity at farmer's field.

(11) Review of Research on the topic : (a) Abroad

(b) In India

(12) Technical Programme : (a) Sampling

(b) Yearwise Research Programme

(c) Data Collection

(d) Analytical Framework

(13) Facilities Available : 

(14) Duration of the Project : 3 Years

(15) Estimates of cost : Enclosed

(16) References / Bibliography : Enclosed

## 3.3 Glossary

**Questionnaire**: The tool or instrument of data collection.

**Population**: The aggregate of all individuals is known as population or universe.

**Sample** : It is selected number of individuals each of which is the part of Population.

**Sampling** :The process of selecting individual from population is known as Sampling.

**Hypothesis** : A proposition which can be put to test to determine validity.

**Variable** : A variable is a quantity, which may take any one of a specified set of values. e.g. Sex, Height etc.

**Continuous Variable** : The variable, which takes all possible values in a given specified range, is termed as continuous variable.

**Discrete Variable**: Those variables which cannot take all possible values within a given specified range is called as discrete variable. Since marks can take only integer values from 0 to 100.

**Census** : Collection of information and data about each and every item of population.

## 3.4 Summary

Identifying the problem is a long drawn process for those who are new to research. After demarcating the broad area researcher must posses two considerations i.e. interest and competance. Once the main area of research is decided, the next step is to narrow it down to a suitable topic and it is a half a step forward. The selection of research problem is very important job for a researcher and it depends whether it is conducted as a requirement of degree or it is for academic interest. The research problem may be selected from the sources such as - (i) Theory of one's interest, (ii) Daily problems, (iii) Technological changes, (iv) Unexplored areas and (v) Discussions with supervisions. The research problem should be defined in a systematic manner giving the weightage to all relating points. The technique involves the steps such as, (i) Statement of the problem, (ii) Nature of the problem, (iii) Available literature, (iv) Developing idea and (vi) Rephrasing the research problem etc. It is generally said that the problem selected should be good which fulfill the following criteria i.e. relation between variables expressed, clarity and unambiguity, empirical verification, availability of guidance, experience and creativity, importance and feasibility.

The good research design should consist of the components i.e. Title of the study, The Problem, Review of Literature, Scope of the study, Objective of the study, Formulation of Hypothesis, Definition and Concepts, Methodology, Sampling Design, Constructing of Schedules, Collection of Data, Analysis of Data, Interpretation of Results and Reporting Findings.

## 3.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. Explain, how to select suitable research problem.
2. State and explain different criteria of good research problem.
3. Enlist and explain components of good research design.
4. What are the techniques involved in defining good research problems ?

# Unit 4 : Research Design

***Index***



## 4.1 Introduction

Now a day there is an increasing tendency on the part of students and teachers to udertake research projects. The selection of research topic is the first and foremost problem, which the prospective researcher notions on what he would research upon. Often in the initial stages, the canvas is very wide. It is necessary, in view of limited resources of an individual research worker to handle adequately and intensively large canvas, to delimit the area of investigation. This will facilitate intensive investigation into the problem so that well-founded conclusion can be drawn on the phenomenon under consideration. All this requires proper planning. But one of the most common defects noticed in the research after selection of the problem is the lack of smooth flow of research work. As soon as the problem is selected, one hastily leaps forward with the work of collection of data or writing the plan for the thesis or dissertation and even some times writing of initials, introductory chapters in the thesis or dissertation. This shows that the researcher is not devoting considerable time in thinking about his future work. Researcher should think about the way in which he should proceed in attaining his objectives in his research work. Without this research work will become futile and results is wastage of time and resources, which are very precious for a researcher.

**After the study of this unit, you will be able to know and understand :**

- Necessary steps in formulating a research design.
- Features of a good research design.
- Different types of research designs.
- Important Concepts Relating to Research Design
- Different Research Designs
- Basic Principles of Experimental Design

## 4.2 Content

### 4.2.1 Need for Research Designs

Research design is needed because it facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible yielding maximal information with minimal expenditure of effort, time and money. Just as for better, economical and attractive construction of a house, we need a blueprint or what is commonly called the map of the house well thought out and prepared by a expert architect, similarly we need a research design or a plan in advance of data collection and analysis for our research project. Research design stands for advance planning of the methods to be adopted for collecting the relevant data and the techniques to be used in their analysis, keeping in view the objective of the research and the availability of staff, time and money. Preparation of the research design should be done with great care as any error in it may upset the entire project. Research design, in fact, has a great bearing on the reliability of the results arrived at and as such constitutes the firm foundation of the entire edifice of the research work.

Even then the need for a well thought out research design is at times not realized by many. The importance, which this problem deserves, is not given to it. As a result many researches do not serve the purpose for which they are undertaken. In fact, they may even give misleading conclusions. Thoughtlessness in designing the research project may result in rendering the research exercise futile. It is, therefore, imperative that an efficient and appropriate design must be prepared before starting research operations. The design helps the researcher to organize his ideas in a form whereby it will be possible for him to look for flaws and inadequacies. Such a design can even be given to others for their comments and critical evaluation. In the absence of such a course of action, it will be difficult for the critic to provide a comprehensive review of the proposed study.

### 4.2.2 Features of a Good Research Design

A good research design is often characterized by adjectives like flexible, appropriate, efficient, economical and so on. Generally, the design, which minimizes bias and maximizes the reliability of the data collected and analysed is considered a good design. The design, which gives the smallest experimental error, is supposed to be the best design in many investigations. Similarly, a design, which yields maximal information and provides an opportunity for considering many different aspects of a problem is considered most appropriate and efficient design in respect of many research problems. Thus, the question of good design is related to the purpose or objective of the research problem and also with the nature of the problem to be studied. A design may be quite suitable in one case, but may be found wanting in one respect or the other in the context of some other research problem. One single design cannot serve the purpose of all types of research problems.

A research design appropriate for a particular research problem, usually involves the consideration of the following factors:

(i) The means of obtaining information;
(ii) The availability and skills of the researcher and his staff;
(iii) The objective of the problem to be studied;
(iv) The nature of the problem to be studied; and
(v) The availability of time and money for the research work.

If the research study happens to be an exploratory or a formulative one, wherein the major emphasis is on discovery of ideas and insights, the research design most appropriate must be flexible enough to permit the consideration of many different aspects of a phenomenon. But when the purpose of a study is accurate description of a situation or of an association between variables or in what are called the descriptive studies, accuracy becomes a major consideration and a research design which minimizes bias and maximizes the reliability of the evidence collected is considered a good design. Studies involving the testing of a hypothesis of a causal relationship between variables require a design, which will permit inferences about causality in addition to the minimization of bias and maximization of reliability.

But in practice it is the most difficult task to put a particular study in a particular group, for a given research may have in it elements of two or more functions of different studies. It is only on the basis of its primary function that a study can be categorized either as an exploratory or descriptive or hypothesis-testing study and accordingly the choice of a research design may be made in case of a particular study. Besides, the availability of time, money, skills of the research staff and the means of obtaining the information must be given due weightage while working out the relevant details of the research design such as experimental design, survey design, sample design and the like.

Every design has its own strengths and weaknesses and at the same time there is no such thing as a single correct design. A good research design should satisfy the following four conditions: objectivity, reliability, validity and generalisability of the findings.

**(a) Objectivity**

The objectivity of the findings pertains to the methods of collection of data and scoring of the responses or both may be judged by the degree of agreement between the final scores assigned to different individuals by more than one independent observation. The more subjective the observation, recording and evaluation of the responses, the less the diferent observers, recording and evaluation of the responses, the less the different observers agree.

The investigators who use closed ended questionnaires in which each item is supplied with certain alternatives which forcing the respondent to chose one among them are said to be collecting data with the help of objective tools because all the scorers can apply a scoring key and agree perfectly on the result. In contrast, the open ended questionnaires questionnaires in which the respondents are allowed to give free responses to each of the items allow room for great disagreement among the scores. However in certain instance the questionnaires consisting of both the types of item are used purposefully. In such instances free response items and observations can be made fairly objective by giving careful instructions and guidelines to the observer or scorer.

Thus any research design should permit the use of measuring instruments, which are fairly objective in which every observer or judge seeing a performance arrives at precisely the same report. This ensures the objectivity of the collected data that will be used for the analysis, inferences and generalizations.

**(b) Reliability**

Reliability refers to consistency throughout a series of measurements. That is to say, if a respondent gives out a response to a particular item, he is expected to give the same response to that item whenever he is asked subsequently. On the contrary, if the respondent keeps on changing his responses to see the item when he is asked repeatedly, then the investigator will be facing a difficulty in considering which one of these responses is the genuine response of the respondent. So, the investigator should frame his items in such a way that the respondent cannot but give only one genuine response. There are different methods in determining the reliability of the responses given out by a respondent. Some of these methods are using check items, administering the same test repeatedly; using a series of Parallel' forms etc.

**(c) Validity**

Any measuring instrument is said to be valid when it measures what it propose to measure. For example, an intelligence test, constructed for measuring intelligence should measure only intelligence and nothing else. As in the case of reliability there are a good number of procedures for establishing the validity of test. Some such procedures are validating the present data against a concurrent' criterion or a future criterion or a theory etc.

**(d) Generalization**

Once it is ensured that the measuring instruments used in a research investigation yield objective, reliable and valid data, the next important problem, a well planned research design has to answer is the generalisability of the findings of the present study. That is how best the data collected from a sample can be utilized for drawing certain generalizations. Applicable from a sample can be utilized for drawing certain generalizations; applicable to a larger group (population) from which the sample is drawn. In other words, with how much authority and confidence, an investigator can say that the same findings will be obtained even though the data is collected from the total population from which the sample is selected.

A research design thus helps an investigator in his attempt to generalize the findings, provided he has taken due care in defining the population, selecting the sample and using the appropriate statistical analysis while planning his research design. Thus a good research design should ensure that: (i) The measuring instrument can yield objective, reliable and valid data: (ii) the 'Population' is defined in unequivocal term, (iii) the requisite size of the sample is collected by using the most appropriate technique of sample selection, (iv) the appropriate statistical analysis has been employed, and (v) the findings of the present study can be 'generalized without being contaminated by the errors of measurement or sampling errors or any other interfering factors.

## 4.2.3 Important Concepts of Research Design

Before describing the different research designs, it will be appropriate to explain various concepts relating to designs so that these may be better and easily understood.

**(1) Dependent and independent variables**

A concept, which can take on different quantitative values, is called a variable.

As such the concepts like weight, height, income are all examples of variables. Qualitative phenomenon are also quantified on the basis of presence or absence of the concerning attribute(s) Phenomenon which can take on quantitatively different values even in decimal points are called ‘continuous variables’. But all variables are not continuous. If they can be only expressed in integer values, they are non-continuous variables or in statistical language ‘discrete variables’. Age is an example of continuous variable, but the number of children is an example of non-continuous variable. If one variable depends upon or is a consequence of the other variable, it is termed as a dependent variable, and the other variable that is antecedent to the dependent variable is termed as an independent variable. For instance, if we say that height depends upon age then height is an dependent variable and age is an independent variable. Further, if in addition to being dependent upon age, height also depends upon the individual’s sex, then height is a dependent variable and age and sex are independent variables. Similarly, readymade films and lectures are examples of independent variables, whereas behavioral changes, occurring as a result of the environmental manipulations, are examples of dependent variables.

**(2) Extraneous variable**

Independent variables that are not related to the purpose of the study, but may affect the dependent variable are termed as extraneous variables. Suppose the researcher wants to test the hypothesis that there is a relationship between children gains in social studies achievement and their self-concepts. In this case self-concept is an independent variable and social studies achievement is a dependent variable. Intelligence may as well affect the social studies achievement, but since it is not related to the purpose of the study undertaken by the researcher, it will be termed as an extraneous variable. Whatever effect is noticed on dependent variable as a result of extraneous variable(s) is technically described as an experimental error’. A study must always be so designed that the effect upon the dependent variable is attributed entirely to the independent variable(s), and not to some extraneous variable or variables.

**(3) Control**

One important characteristics of good research design is to minimize the influence or effect of extraneous variable(s). The technical term ‘control’ is used when we design the study minimizing the effects of extraneous independent variables. In experimental researches, the term ‘control’ is used to refer restrained experimental condition.

**(4) Confounded Relationship**

When the dependent variable is not free from the influences of extraneous variable(s), the relationship between dependent and independent variables is said to be confounded by an extraneous variable(s).

**(5) Research hypothesis**

When a prediction or hypothesized relationship is to be tested by scientific methods, it is termed as research hypothesis. The research hypothesis is a predictive statement that relates an independent variable to a dependent variable. Usually a research hypothesis must contain, at least, one independent and one dependent variable. Predictive statements, which are not to be objectively verified or the relationships that are assumed but not to be tested, are not termed as research hypothesis.

**(6) Experimental and Non-experimental Hypothesis Testing Research**

When the purpose of research is to test a research hypothesis, it is termed as hypothesis-testing research. It can be of the experimental design or the non-experimental design. Research in which the independent variable is manipulated is termed as ‘experimental hypothesis-testing research’. For instance, suppose a researcher wants to study whether intelligence affects reading ability for a group of students and for this purpose he randomly selects 50 students and tests their intelligence and reading ability by calculating the coefficient of correlation between the two set of scores. This is an example of non-experimental hypothesis-testing research because here in the independent variable, intelligence, is not manipulated. But, now suppose that our researcher randomly selects 50 students from a group of students who are to take course in statistics and then divides them into two groups by randomly assigning 25 in group A, the usual studies programme, and 25 in group B, the special studies programme. At the end of course, he administers a test to each group in order to judge the effectiveness of the training programme on the student's performance-level. This an example of experimental hypothesis-testing research because in this case the independent variable, viz, the type of training programme, is manipulated.

**(7) Experimental and Control Groups**

In an experimental hypothesis-testing research when group is exposed to usual conditions, it is termed a 'control group', but when the group is exposed to some novel or special condition, it is termed an 'experimental group'. In the above illustration, the group A can be called a control group and the group B are exposed to special studies programmes, then both the groups would be termed 'experimental groups'. It is possible to design studies, which include only experimental groups, or studies, which include both experimental and control groups.

**(8) Treatments**

The different conditions under which experimental and control groups are put, are usually referred as 'treatments '. In the illustration taken above, the two treatments are the usual studies programme and the special studies programme. Similarly, if we want to determine through an experiment the comparative impact of three doses of fertilizer on the yield of Sugarcane, in that case three doses of fertilizers will be treated as three treatments.

**(9) Experiment**

The process of examining the truth of a statistical hypothesis, relating to some research problem, is known as an experiment. For example, we can conduct an experiment to examine the usefulness of a certain newly developed drug. Experiments can be of two types, viz., absolute experiment and comparative experiment. If we want to determine the impact of a fertilizer on the yield of a crop, it is a case of absolute experiment. But if we want to determine the impact of one fertilizer as compared to the impact of some other fertilizer, our experiment then will be termed as a comparative experiment. Often, we undertake comparative experiment when we talk of designs of experiments.

**(10) Experimental Units**

The predetermined plots or the blocks, where different treatments are used, are known as experimental units. Such experimental unit must be selected very carefully.

## 4.2.4 Different Research Designs

The research design is a plan according to which observations are made and data assembled. It provides the empirical and logical basis for drawing conclusions and gaining knowledge.

There are different kinds of research designs. They vary from general and sketchy statements of intent of carefully detailed and highly complex investigations. The research designs can be divided into four types, viz., descriptive, diagnostic, exploratory and experimental.

Description is what we do when we try to answer the question; what is the state of affairs as it exists? Although description of an activity is essential it is not sufficient merely to explain the activity, event or phenomenon. Scientific explanation about the relationship among interactivity or isolating components of an activity or event is needed. This is accomplished through diagnostic research. In addition, we have experimental designs, which permit us to make fairly safe inferences about the relationships among the variables and to assess the contributions that each variable singly, or in combination with others, makes to the total experimental outcome. There are also exploratory designs.

Whichever type of research design the investigator selects, it is important that he should have a research design and the design should be made explicit before the research begins. The quality of results and the degree of confidence that the researcher can place in the casual inferences derived, depends upon the design according to which the data were collected. It is important for the social scientist to know the uses and the strengths and weaknesses of each type of research design, for it is vital to understanding and doing effective research.

The different research designs are:

(1) Descriptive design

(2) Diagnostic design

(3) Explanatory design and

(4) Experimental design

**Relation Between Problem Formulation and Research Design**:

The research problem may be formulated in different forms. It may be formulated with different purposes. The nature of the research design depends on the way in which the problem is formulated.

(a) If the problem is an explanatory one, it requires explanatory design.

(b) If the problem is to describe characteristics of groups or situations, a descriptive design is necessary.

(c) If the study aims at the solution of a particular problem a diagnostic design is necessary.

(d) If the researcher wants to test a hypothesis of casual relationship between variables, experimental design is necessary.

**(1) Descriptive Design**

The descriptive research designs enable researchers to describe or present picture of a phenomenon or phenomenon under investigation. The methodology involved in such designs is mostly qualitative in nature producing descriptive data, i.e., people's own written or spoken words and observable behavior. There are at least three such approaches which enable us to record or analyze the behavioral pattern : (i) Participant Observation; (ii) Personal Documents and (iii) Unstructured Interviewing.

**(i) Participant Observation :** Participant observation refers to research characterized by a period of intense social interaction between the researcher and the subjects, in the milieu of the subjects. During this period, data are systematically collected.

(ii) **Personal Documents :** The phrase "Personal documents" is used to refer to an individual's descriptive, first person account to the whole or a part of his or her life or an individual's reflection on a specific event or topic. All personal documents are valuable, however, once the researcher has taken the motivations into account. They may encompass a person's entire life or may focus on a single incidental time period or theme. Differences according to this dimension usually determine how the materials are used. Personal documents may be presented either on their original form or more likely in edited and recognized version. The author or subject of a personal document may reveal or conceal his or her name from the reading audience. For a number of reasons, it was left that pseudonyms and fictions names for researcher induced documents are best to use. More concretely such materials are available in the form of autobiographies, diaries, letters and long open-ended interviews that are recorded verbatim.

**(2) Diagnostic Design**

Diagnostic refers to scientific differentiation among various conditions or phenomenon for the purpose of accurately classifying these conditions. In its broadest sense diagnosis corresponds to the fact-finding aspect of clinical practice. Its objectives include screening and classification, personality description, prediction of outcome and attainment of insight by the client. With regard to both data gathering and interpretation, there are a number of ways in which the services of a skilled clinician are needed.

The diagnostic research paradigm represents the most typical and simple problem-solving strategy of the helper faced with problems and crises on the job. It consists of: (1) the emergence of a problem, (2) a diagnosis of its causes, (3) formulation of all the possible avenues of remediation, and (4) recommendations for a possible solution.

Data for diagnosis can be obtained in four major ways:(1) a case history or interview, (2) clinical observation, (3) informal testing, and (4) formal standardized testing. In general, these four methods are not often accomplished simultaneously. One procedure may suggest the others.

**Case Study Method**

Case study method is concerned with everything that is significant in the history or development of the case. The purpose is to understand the life cycle, or an important part of the life cycle of an institution, or an entire community. This method probes deeply and intensively, analyses interactions between the factors that produce change or growth. It emphasizes the longitudinal or genetic approach showing development, over a period of time. In each case the element of typical ness is the focus of attention, with emphasis upon many factors that characterize the type.

The characteristics of a good case study include an adequate data which is valid, continuous, carefully synthesized, and confidential and which should be useful for follow up. An adequate case history clarifies epistemological relationships, resolves controversial points, uncovers illogical thinking and misinformation and aids in the differentiation between functional and organic complaints.

*Interview*

The nature of the personal relationship between interviewer and subject requires an expertness and sensibility that might well be called an art. It requires a skilful interviewer to obtain a maximum amount of useful data. The interviewer must try to establish a feeling of mental trust with the persons being interviewed, being careful not to ask a question that might alarm them. His attitude should convey spirit of co-operation, acceptance, and empathy while maintaining a degree of professional objectivity to guard against excessive emotional involvement and consequent ineffectiveness.

*Clinical Observations*

Many attributes are inadequately identified through other standardized test instruments or through interview. The skilful diagnostician should be able to detect through observation or the behavior and through the proficient use of informal tests. Further informal tests and observation of behavior provide an opportunity to collaborate findings of the other two areas of assessment. Diagnostic tests provide a microscopic view of the component elements of some area of performance. Such tests enable the diagnostician to analyze the individual's functioning within specific sub-skill areas and supply direction for redemption. An assessment of the personal adjustment, co-ordination and development can be made by this method.

*Collaborative Diagnosis*

Many of the problems, which have been attached through field experiment, have in fact been problems related to rather highly developed professional skills in such areas as human relations training, therapy, community organization, etc. For such problems there are already available highly skilled professional people who may be called upon to conduct the manipulation. Once the basic role relationships are worked out, there should be a collaborative diagnosis of the situation by the researcher and at least some part of the client organization. The purpose of this diagnosis is to assess various factors that will be involved in executing the research design-the resistance that may be encountered, the dynamics of the situation, in regard to the problem of bringing about a change etc. The ways of going about such a collaborative diagnosis will vary tremendously, depending upon the problem to be studied, the setting, etc., all relevant experiences and information should be used; but whatever the source, a useful diagnosis will have to be formulated in theoretical terms. Because it is necessary to produce reactions and to bring about planned changes, the diagnosis must go beyond the mere statement of facts or the labeling of things as "good" or "bad". It must move toward more casual thinking, for it is only by manipulating its causes that we can vary dependent variable.

**(3) Exploratory Design**

The exploratory method according to Katz., "represents the earlier stage of science." This significant observation implies that all sciences must have at the beginning an approach, which was purely exploratory. Even as we approach a doctor with an ailment he starts with all sorts of questions, to begin with. From the point of view of the doctor, he is systematically exploring the complaints and is striving to categorise our symptoms. On the basis of this exploration he will arrive at a conclusion, at least tentatively, about the disease. Only then he will, if necessary, call for a pathological report. All these steps, starting with the exploratory questions will help him to arrive at a correct diagnosis. The example will, while making the meaning of exploration clear, also point out the inevitability and universality of the exploratory approach, which cuts across the barrier of the natural and social sciences. At this stage one may be tempted to conclude that the moment one envisages his study as exploratory, the investigator has absolute freedom for random and aimless activity. The social scientist has freedom to follow interesting leads and to utilize his own ingenuity in obtaining information. Yet the social scientist should exercise judicious temperance in this approach. Man's knowledge of the social sciences has progressed to a degree where social scientists do have some knowledge of "The types of things to look for in most social situations".

**(4) Experimental Design**

An experiment in the social sciences is not only different from but also more difficult than an experiment in the natural sciences. A student of natural sciences deals with things that are palpable, tangible and more often than not inert, and consequently more amenable to the rigorous of an experiment. The likelihood on an error creeping into an experiment in the natural sciences probably is more due to the person performing the experiment, rather than the substance, which is being subjected to the experimental verification. On the other hand, in social sciences, in addition to danger of an error on the part of the researcher, there is also the difficulty to understanding what is being measured.

The phenomenon in social sciences may themselves be intrinsic states of the organism, and may have to be measured as inferred properties. Further there is every likelihood that the substance is less amenable to the rigorous of an experiment. The substance in natural sciences remains remarkably alike while in social sciences the substance being human being differ from one another in their genetic make up, like histories, learned personality characteristics, values and abilities and immediate past experience. In spite of these difficulties as Selltiz *et al.*, observed, "When an experiment is possible it is the most effective method of testing a hypotheses, i.e., one variable 'X' causality influences another variable 'Y'.

### 4.2.5 Basic Principles of Experimental Designs

Professor Fisher has enumerated three principles of experimental designs:

(1) Principle of Replication;

(2) Principle of Randomization; and

(3) Principle of Local Control.

*The Principle of Replication :* In this design the experiment should be repeated more than once. Thus, each treatment is applied in many experimental units instead of one. By doing so the statistical accuracy of the experiment is increased. For example, suppose we are to examine the effect of two varieties of grape. For this purpose we may divide the field into two parts and grow one variety in one part and the other variety in the other part. We can then compare the yield of the two parts and draw conclusion on that basis. But if we are to apply the principle of replication to this experiment then we first divide the field into several parts, grow one variety in half of these parts and the other variety in the remaining parts. We can then collect the data of yield of the two varieties and draw conclusion by comparing the same. The result so obtained will be more reliable in comparison to the conclusion we draw without applying the principle of replication. The entire experiment can even be repeated several times for better results. Conceptually replication does not present any difficulty, but computationally it does. For example, if an experiment requiring a two-way analysis of variance is replicated, it will then require a three-way analysis of variance since replication itself may be a source of variation in the data. However, it should be remembered that replication is introduced in order to increase the precision of a study; that is to say, to increase the accuracy with which the main effects and interactions can be estimated.

*The Principle of Randomization :* It provides protection, when we conduct an experiment, against the effects of extraneous factors by randomization. In other words, this principle indicates that we should design or plan the experiment in such a way that the variations caused by extraneous factors can all be combined under the general heading of “chance.” For instance, if we grow one variety of grape, say, in the first half of the parts of a field and the other variety is grown in the other half, then it is just possible that the soil fertility may be different in the first half in comparison to the other half. If this is so, our results would not be realistic. In such a situation, we may assign the variety of grape to be grown in different parts of the field on the basis of some random sampling technique. We may apply randomization principle and protect ourselves against the effects of bias and other extraneous factors (soil fertility differences in the given case.) As such, through the application of the principle of randomization, we can have a better estimate of the experimental error.

*The Principle of Local Control :* It is another important principle of experimental designs. The local control means controlling variation in the experimental units. Under it the extraneous factor, the known source of variability, is made to vary deliberately over as wide a range as necessary and this needs to be done in such a way that the variability it causes can be measured and hence eliminated from the experimental error. This means that we should plan the experiment in a manner that we can perform a two-way analysis of variance, in which the total variability of the data is divided into three components attributed to treatments (varieties of grape in our case), the extraneous factor (soil fertility in our case) and experimental error. In other words, according to the principle of local control, we first divide the field into several homogeneous parts, known as blocks, and then each such block is divided into parts equal to the number of treatments. Then the treatments are randomly assigned to these parts of a block. Dividing the field into several homogenous parts is known as 'blocking'. In general, blocks are the levels at which we hold an extraneous factor fixed, so that we can measure its contribution to the total variability of the data by means of a two-way analysis of variance. In brief, through the principle of local control we can eliminate the variability due to extraneous factor(s) from the experimental error.

## Important Experimental Designs

Experimental design refers to the framework or structure of an experiment and as such there are several experimental designs. We can classify experimental designs into two broad categories, viz., informal experimental designs and formal experimental designs. Informal experimental designs are those designs that normally use a less sophisticated form of analysis based on differences in magnitudes, whereas formal experimental designs offer relatively more control and use precise statistical procedures for analysis. Important experimental designs are as follows:

(a) Informal experimental designs :

(i) Before-and-after without control design.

(ii) After-only with control design.

(iii) Before-and-after with control design.

(b) Formal experimental designs :

(i) Completely Randomized Design (C.R.D.)

(ii) Randomized Block Design (R.B.D.)

(iii) Latin Square Design (L.S.D.)

(iv) Factorial Designs.

We may briefly deal with each of the above stated informal as well as formal experimental designs.

### (1) Before-and-after without control design

In such a design a single test group or area is selected and the dependent variable is measured before the introduction of the treatment. The treatment is then introduced and the dependent variable is measured again after the treatment has been introduced. The effect of the treatment would be equal to the level of the phenomenon after the treatment minus the level of the phenomenon before the treatment. The design can be represented thus :

| | | | |
|---|---|---|---|
| Test area: | Level of phenomenon Before treatment (X) | $\xrightarrow{\text{Treatment}}$ | Level of phenomenon after treatment (Y) |

$$\therefore \textit{Treatment Effect} = Y - X$$

The main difficulty of such a design is that with the passage of time considerable extraneous variations may be there in its treatment effect.

**(2)After-only with control design**

In this design two groups or areas (test area and control area) are selected and the treatment is introduced into the test area only. The dependent variable is then measured in both the areas at the same time. Treatment impact is assessed by subtracting the value of the dependent variable in the control area from its value in the test area. This can be exhibited in the following form :

Test area :

Control area $\xrightarrow{\text{Treatment}}$ $\dfrac{\text{Level of phenomenon after treatment (Y)}}{\text{Level of phenomenon without Treatment (Z)}}$

$$\therefore \; \textit{Treatment Effect} = Y - Z$$

The basic assumption in such a design is that the two areas are identical with respect to their behaviour towards the phenomenon considered. If this assumption is not true, there is the possibility of extraneous variation entering into the treatment effect. However, data can be collected in such a design without the introduction of problems with the passage of time. In this respect this design is superior to before-and-after without control design.

**(3) Before-and-after with control design**

In this design two areas are selected and the dependent variable is measured in both the areas for an, identical time-period before the treatment. The treatment is then introduced into the test area only, and the dependent variable is measured in both for an identical time-period after the introduction of treatment. The treatment effect is determined by subtracting the change in the dependent variable in the control area from the change in the dependent variable in test area. This design can be shown in this way:

| | Time Period I | | Time Period II |
|---|---|---|---|
| Test area : | Level of phenomenon Before treatment (X) | Treatment → | Level of phenomenon after treatment (Y) |
| Test area : | Level of phenomenon Before treatment (A) | Treatment → | Level of phenomenon after treatment (Z) |

$\therefore$ *Treatment Effect = (Y-X) – (Z-A)*

This design is superior to the above two designs for the simple reason that it avoids extraneous variation resulting both from the passage of time and from non-comparability of the test and control areas. But at times, due to lack of historical data, time or a comparable control area, we should prefer to select one of the first two informal designs stated above.

**(4) Completely Randomized Design (C.R.D.)**

'Involves only two principles viz., the principle of replication and the principle of randomization of experimental designs. It is the simplest possible design and its procedure of analysis is also easier. The essential characteristic of this design is that subjects are randomly assigned to experimental treatments (or vice-versa). In C.R.D. treatments are completely randomly allotted to experimental units; hence experimental units have equal chance to receive any treatment. Therefore number of treatment per replications may not be same. In C.R.D. the experimental units are required to be homogeneous. If there is variation among experimental units, then C.R.D. is not suitable. In C.R.D. there is only one factor i.e. different treatments causing variation, hence one-way analysis of variance procedure is used to analyze the data. For instance, if we have 10 subjects and if we wish to test 5 under treatment A and 5 under treatment B, the randomization process gives every possible group of 5 subjects selected from a set of 10 an equal opportunity of being assigned to treatment A and treatment B. One-way analysis of variance or one-way ANOVA is used to analyse such a design. It provides maximum number of degrees of freedom to the error. Such a design is generally used when experimental areas happen to be homogeneous.

**(i) Two-group simple randomized design**

In a two-group simple randomized design, first of all the population is defined and then from the population a sample is selected randomly. Further, requirement of this design is that items, after being selected randomly from the population, be randomly assigned to the experimental and control groups (Such random assignment of items to two groups is technically described as principle of randomization). Thus, this design yields two groups as representatives of the population. In a diagram form this design can be shown in this way:

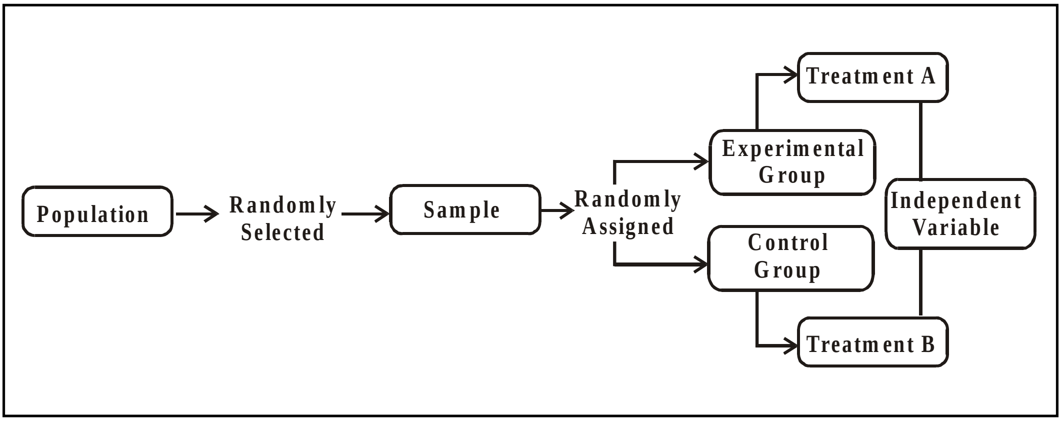



**Diagram 1 : Simple Randomised Design**

Since in the simple randomized design the elements constituting the sample are randomly drawn from the same population and randomly assigned to the experimental and control groups, it becomes possible to draw conclusions on the basis of samples applicable for the population. The two groups (experimental and control groups) of such a design are given different treatments of the independent variable. This design of experiment is quite common in research studies concerning behavioural sciences. The merit of such a design is that it is simple and randomizes the differences among the sample items. But the limitation of it is that the individual differences among those conducting the treatments are not eliminated, i.e., it does not control the extraneous variable and as such the result of the experiment may not depict a correct picture. This can be illustrated by taking an example. Suppose the researcher wants to compare two groups of students who have been randomly selected and randomly assigned, two different treatments viz., the usual training and the specialized training are being given to the two groups. The researcher hypothesis's greater gains for the group receiving specialized training. To determine this, he tests each group before and after the training, and then compares the amount of gain for the two groups to accept or reject his hypothesis. This is an illustration of the two-groups randomized design, wherein individual differences among students are being randomized. But this does not control the differential effects of the extraneous independent variables (in this case, the individual differences among those conducting the training programme).

**(ii) Random replications design**

The limitation of the two-group randomized design is usually eliminated within the random replications design. In the illustration just cited above, the *teacher differences* on the dependent variable were ignored, i.e., the extraneous variable was not controlled. But in a random replications design, the effect of such differences are minimized by providing a number of repetitions for each treatment. Each repetition is technically called a 'replication'. Random replication design serves two purposes viz., it provides controls for the differential effects of the extraneous independent variables and secondly, it randomizes any individual differences among those conducting the treatments. Diagrammatically we can illustrate the random replications design as shown in Diagram 2.

From the diagram it is clear that there are two populations in the replication design. The sample is taken randomly from the population available for study and is randomly assigned to, say, four experimental and four control groups. Similarly, sample is taken randomly from the population available to conduct experiments (because of the eight groups eight such individuals be selected) and the eight individuals so selected should be randomly assigned to the eight groups, Generally, equal number of items are put in each group so that the size of the group is not likely to affect the results of the study. Variables relating to both population characteristics are assumed to be randomly distributed among the two groups. Thus, this random replication design is, in fact, an extension of the two-group simple randomized design.

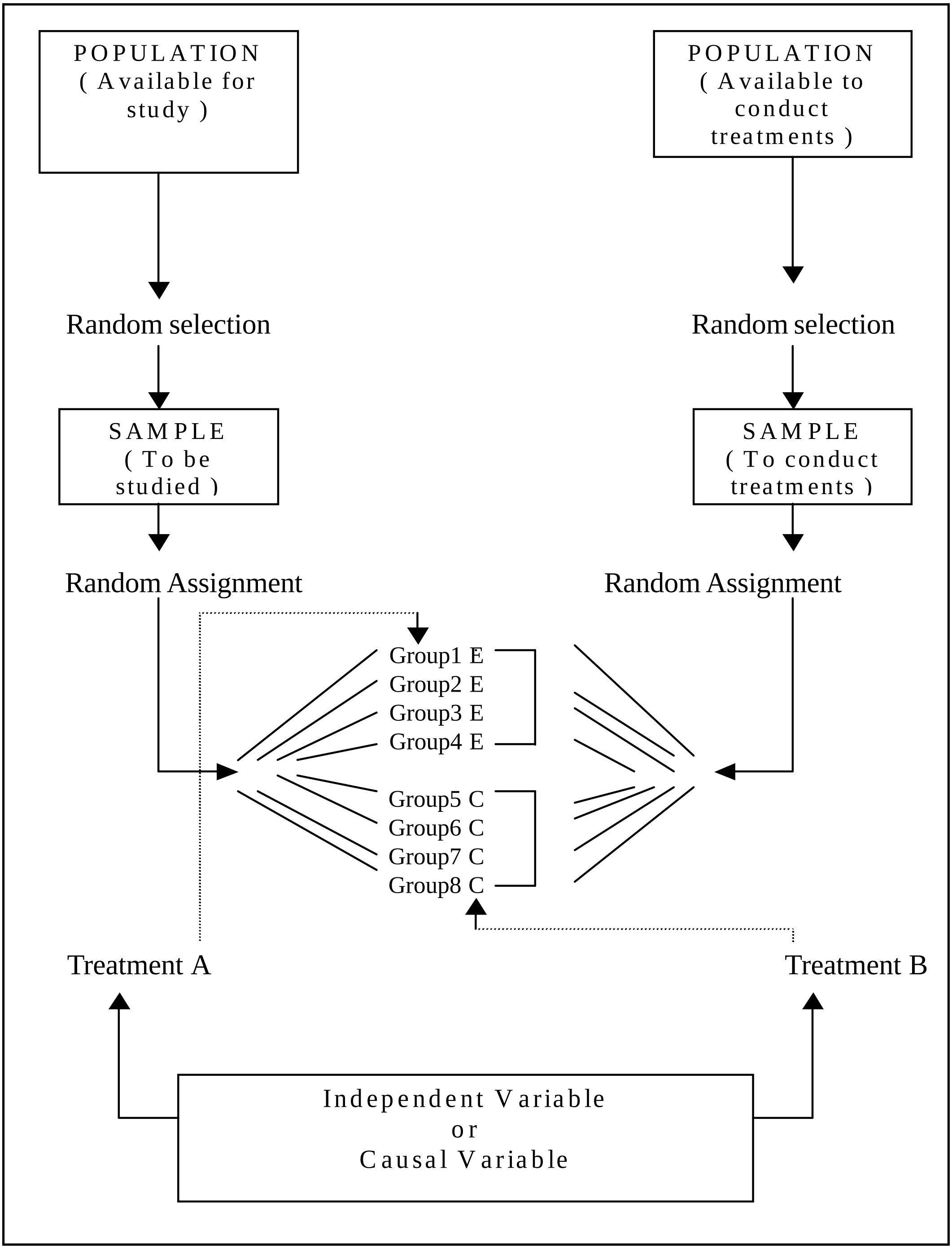



**Diagram 2 : Random Replication Design**

## (5) Randomized Block Design (R.B.D.)

Randomized block design is an improvement over the Completely Randomised Design. In the R.B.D. the principle of local control is applied along with the other two principles of Replication and Randomization of experimental designs. In the R.B. design, subjects are first divided into groups, known as blocks, such that within each group the subjects are relatively homogeneous in respect to some selected variable. The variable selected for grouping the subjects is one that is believed to be related to the measures to be obtained in respect of the dependent variable. The number of subjects in a given block would be equal to the number of treatment. In general, blocks are the levels at which we hold the extraneous factor fixed, so that its contribution to the total variability of data can be measured. The main feature of the R.B. design is that in this each treatment appears the same number of times in each block. The R.B. design is analysed by the two-way analysis of variance (two way ANOV(A) technique. Let us illustrate the R.B. design with the help of an example. Suppose four different forms of a standardized test in statistics were given to each of five students (selected one from each of the five I.Q. blocks) and following are the scores which they obtained.

| | Very low I.Q. | Low I.Q. | Average I.Q. | High I.Q. | Very high I.Q. |
|---|---|---|---|---|---|
| | Student A | Student B | Student C | Student D | Student E |
| Form 1 | 82 | 67 | 57 | 71 | 73 |
| Form 2 | 90 | 68 | 54 | 70 | 81 |
| Form 3 | 86 | 73 | 51 | 69 | 84 |
| Form 4 | 93 | 77 | 60 | 65 | 71 |

If each student separately randomized the order in which he or she took the four tests (by using random numbers or some similar device), refer to the design of this experiment as a R.B. design. The purpose of this randomization is to take care of such possible extraneous factors (say as fatigue) or perhaps the experience gained from repeatedly taking the test.

### (6) Latin Square Design (L.S.D.)

Latin square design is an experimental design very frequently used in agricultural research. The conditions under which agricultural investigations are carried out are different from those in other studies for nature plays an important role in agriculture. For instance, an experiment has to be made through which the effects of five different varieties of fertilizers on the yield of a certain crop, say wheat, is to be judged. In such a case the varying fertility of the soil in different blocks in which the experiment has to be performed must be taken into consideration; otherwise the results obtained may not be very dependable because the output happens to be the effect not only of fertilizers, but it may also be the effect of fertility of soil. Similarly, there may be the impact of seeds on the yield. The overcome such difficulties, the L.S. design is used when there are two major extraneous factors such as the varying soil fertility and varying seeds.

The Latin-square design is one wherein each fertilizer, in our example, appears five times but is used only once in each row and in each column of the design. In other words, the treatments in a L.S. design are so allocated among the plots that no treatment occurs more than once in any one row or and columns (one through rows and the other through columns). The following is a diagrammatic form of such a design in respect of, say, five types of fertilizers, viz., A, B, C, D and E and the two blocking factors, viz., the varying soil fertility and the varying seeds :

Fertility Level

Variety of Seeds

| | I | II | III | IV | V |
|---|---|---|---|---|---|
| $X_1$ | A | B | C | D | E |
| $X_2$ | B | C | D | E | A |
| $X_3$ | C | D | E | A | B |
| $X_4$ | D | E | A | B | C |
| $X_5$ | E | A | B | C | D |

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

The merit of this experimental design is that it enables differences in fertility gradients in the field to be eliminated in comparison to the effects of different varieties of fertilizers on the yield of the crop. But this design suffers from one limitation, and it is that although each row and each column represents equally all fertilizer varieties, there may be considerable difference in the row and column means both up and across the field. This, in other words, means that in L.S. design we must assume that there is no interaction between treatments and blocking factors. This defect can, however, be removed by taking the means of rows and columns equal to the field mean by adjusting the results. Another limitation of this design is that it requires number of rows, columns and treatments to be equal. This reduces the utility of this design. In case of ( 2 x 2 ) L. S. design, there are no degrees of freedom available for the mean square error and hence the design cannot be used. If treatments are 10 or more, then each row and each column will be larger in size so that rows and columns may not be homogeneous. This may make the application of the principle of local control ineffective. Therefore, L.S. design of orders (5 x 5) to (9 x 9) are generally used.

## (7) Factorial designs

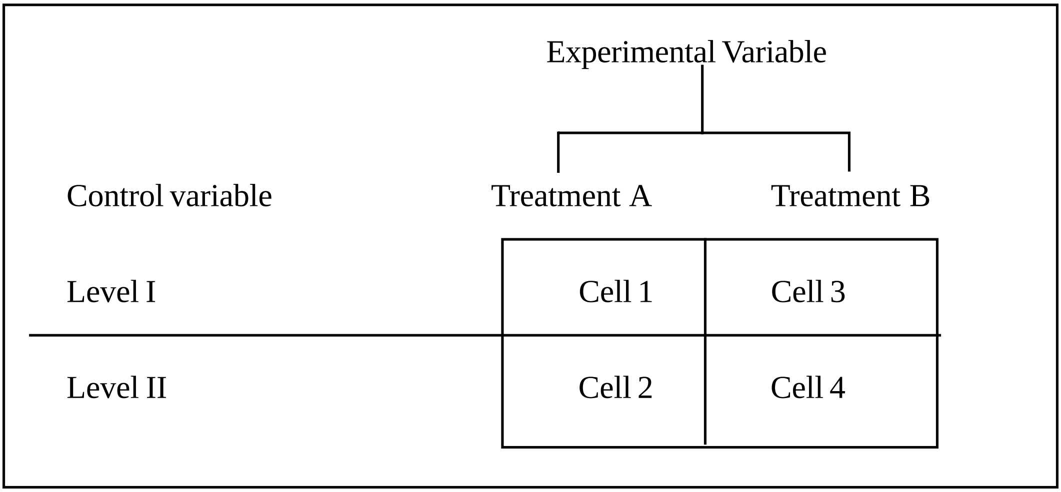
Factorial designs are used in experiments where the effects of varying more than one factor are to be determined. They are specially important in several economic and social phenomenon where usually a large number of factors affect a particular problem. Factorial designs can be of two types; (i) simple factorial designs and (ii) complex factorial designs. We take them separately.

### (i) Simple factorial designs

In case of simple factorial designs, we consider the effects of varying two factors on the dependent variable, but when an experiment is done with more than two factors, we use complex factorial designs. Simple factorial design is also termed as a 'two-factor-factorial design,' whereas complex factorial design is known as 'multi-factor-factorial design.' Simple factorial design may either be a 2 x 2 simple factorial design, or it may be, say, 3 x 4 or 5 x 3 or the like type of simple factorial design. We illustrate some simple factorial designs as under.

**Illustration 1 :** (2 x 2 simple factorial design).

A 2 x 2 simple factorial design can graphically be depicted as follows:




**2 x 2 Simple Factorial Design**

In this design the extraneous variable to be controlled by homogeneity is called the control variable and the independent variable, which is manipulated, is called the experimental variable. Then there are two treatments of the experimental variable and two levels of the control variable. As such there are four cells into which the sample is divided. Each of the four combinations would provide one treatment or experimental condition. Subjects are assigned at random to each treatment in the same manner as in a randomized group design. The means for different cells may be obtained along with the means for different rows and columns. Means of different cells represent the mean scores for the dependent variable and the column means in the given design are termed the main effect for treatments without taking into account any differential effect that is due to the level of the control variable. Similarly, the row means in the said design are termed the main effects for levels without regard to treatment. Thus, through this design we can study the main effects of treatment as well as the main effects of levels. An additional merit of this design is that one can examine the interaction between treatments and levels, through which one may say weather the treatment and levels are independent of each other or they are not so. The following examples make clear the interaction effect between treatments and levels. The data obtained in case of two (2 x 2) simple factorial studies may be given as under.

**Study I Data**

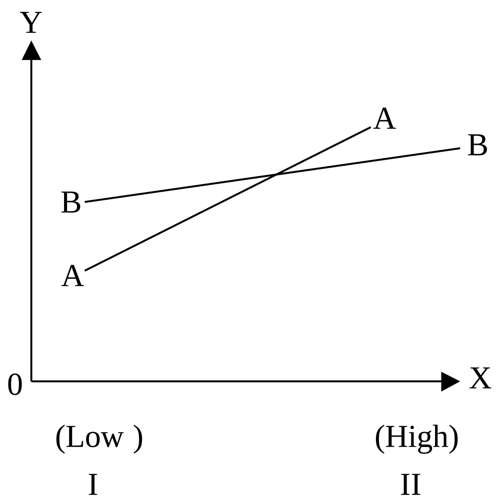
| | | Training | | Row mean |
|---|---|---|---|---|
| | | Treatment A | Treatment B | |
| Control (Intelligence) | Level I (Low) | 15.5 | 23.3 | 19.4 |
| | Level II (High) | 35.8 | 30.2 | 33.0 |
| | Column mean | 25.6 | 26.7 | |

**Study II Data**

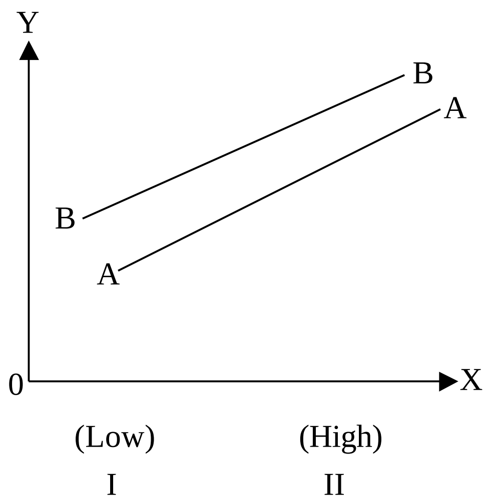
| | | Training | | Row mean |
|---|---|---|---|---|
| | | Treatment A | Treatment B | |
| Control (Intelligence) | Level I (Low) | 10.4 | 20.6 | 15.5 |
| | Level II (High) | 30.6 | 40.4 | 35.5 |
| | Column mean | 20.5 | 30.5 | |

All the above figures (the study I data and the study II data) represent the respective means. Graphically, these can be represented as shown below.

Study I




Study II




On X axis : Control level (say intelligence)

On Y axis : Mean score of Dependent variables (say ability)

In factorial experiment effects of several factors at different levels in all possible combinations are studied together. The graph relating to study I indicates that there is an interaction between the treatment and the level which, in other words, means that the treatment and the level are not independent of each other. The graph relating to Study II shows that there is no interaction effect which means that treatment and level in this study are relatively independent of each other.

The 2 x 2 design need not be restricted in the manner as explained above i.e., having one experimental variable and one control variable, but it may also be of the type having two experimental variables or two control variables. For example, a college teacher compared the effects of the class-size as well as the introduction of the new instruction technique on the learning of research methodology. For this purpose he conducted a study using a 2 x 2 simple factorial design. His design in the graphic form would be as follows :

| | | Experimental Variable I (Class Size) | |
|---|---|---|---|
| | | Small | Usual |
| Experimental Variable II (Instruction technique) | New | | |
| | Usual | | |

But if the teacher uses a design for comparing males and females and the senior and junior students in the college as they relate to the knowledge of research methodology, in that case we will have a 2 x 2 simple factorial design wherein both the variables are control variables as no manipulation is involved in respect of both the variables.

**Illustration 2 :** 4 x 3 simple factorial design.

The 4 x 3 simple factorial design will usually include four treatments of the experimental variable and three levels of the control variable. Graphically it may take the following form :

**4 x 3 Simple Factorial Design**

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

This model of a simple factorial design includes four treatments viz., A, B, C, and D of the experimental variable and three levels viz., I, II, and III of the control variable and has 12 different cells as shown above. This shows that a 2 x 2 simple factorial design can be generalized to any number of treatments and levels. Accordingly we can name it as such and such (-x-) design. In such a design the means for the columns provide the researcher with an estimate of the main effects for treatments and the means for rows provide an estimate of the main effects for the levels. Such a design also enables the researcher to determine the interaction between treatments and levels.

**(ii) Complex factorial designs**

Experiments with more than two factors at a time involve the use of complex factorial designs. A design which considers three or more independent variables simultaneously is called a complex factorial design. In case of three factors with one experimental variable having two treatments and two control variables, each one of which having two levels, the design used will be termed 2 x 2 x 2 complex factorial design which will contain a total of eight cells as shown below.

**2 x 2 x 2 Complex Factorial Design**

| | Experimental Variable | | | |
|---|---|---|---|---|
| | Treatment A | | Treatment B | |
| | Control Variable 2 Level I | Control Variable 2 Level II | Control Variable 2 Level I | Control Variable 2 Level II |
| Control Level I | Cell 1 | Cell 3 | Cell 5 | Cell 7 |
| Variable Level II | Cell 2 | Cell 4 | Cell 6 | Cell 8 |

A pictorial presentation is given in Diagram 3 for design. The dotted line cell in this diagram corresponds to Cell 1 of the above stated 2 x 2 x 2 design and is for Treatment A, level I of the control variable 1 and level I of the control variable 2. From this design it is possible to determine the main effects for three variables i.e., one experimental and two control variables. The researcher can also determine the interactions between each possible pair of variables such interactions are called 'First Order interactions' and interaction between variable taken in triplets such interactions are called Second Order interactions. In case of a 2 x 2 x 2 design, the further given first order interactions are possible:

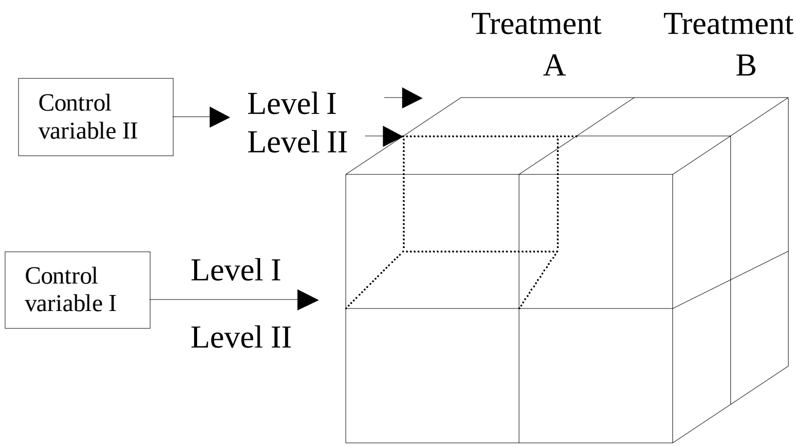



**Diagram 3 : 2 x 2 x Complex Factorial Design**

It may, however, be noted that the complex factorial design need not necessarily be of 2 x 2 x 2 type design, but can be generalized to any number and combination of experimental and control independent variables. The greater the number of independent variables included in a complex factorial design, the higher the order of the interaction analysis possible. But the overall task goes on becoming more and more complicated with the inclusion of more and more independent variables kin our design.

Factorial designs are used mainly because of the two advantages (i) They provide equivalent accuracy as happens in the case of experiments with only one factor with less labour and as such are a source of economy. Using factorial design, we can determine the main effects of two in simple factorial design or more in case of complex factorial design factors or variables in one single experiment. (ii) They permit various other comparisons of interest. For example, they give information about such effects which cannot be obtained by treating one single factor at a time. The determination of interaction effects is possible in case of factorial designs.

There are several research designs and the researcher must decide in advance of collection and analysis of data as to which design would prove to be more appropriate for his research project. He must give due weightage to various points such as the type of population and its nature, the objective of his study, the source list or the sampling frame, desired standard of accuracy and the like when taking a decision in respect of the design for his research project.

## 4.3 Glossary

**Research design**: A research design is a plan of proposed research work.

**Objectivity**: Pertains to the methods of collection of data and scoring of the responses.

**Reliability**: Refers to consistency throughout a series of measurements.

**Validity**: Any measuring instrument is said to be valid when it measures what it purports to measure.

**Variable**: Different quantitative values is called a variable.

**Independent and Dependent variable**: When we say that yield of any crop depends upon the weather parameters, viz., rainfall, humidity, temperature etc., then weather parameters are independent variables and yield of a crop is dependent variable.

**Extraneous Variables**: Independent variables that are not related to the purpose of the study, but may affect the dependent variable are termed as extraneous variables.

**Research hypothesis**: When a prediction or hypothesized relationship is to be tested by scientific methods, it is termed as research hypothesis.

**Experiment**: The process of examining the truth of a statistical hypothesis, relating to some research problem.

**Experimental Units**: The predetermined plots or the blocks, where different treatments.

## 4.4 Summary

It is essential for the researchers the way in which he should proceed in attaining his objectives for which proper design of research is important. Without thoughtful designing of research project may result in rendering the research futile. Every design has its own strengths and weaknesses and at the same time there is no such things as a single correct design. A good research design should satisfy the four conditions i.e. Objectivity, Reliability, Validity and Generarisability of the findings.

It is appropriate to explain various concepts relating to design so that these may be better and easily understanding, before describing the research design. The important concepts here are independent and extraneous variables, control, research hypothesis and testing, treatment, experiments and units. It is important for Social Scientist to know the uses, strength and weaknesses of each type of design. There are four important designs i.e. Descriptive, Dignostic, Explanatory and Experimental designs which need detail study.

The researcher must decide in advance of collection and analysis of data as to which design would prove to be more appropriate for his research project.

## 4.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. State and explain four conditions of a good research design.
2. Define, differentiate and explain different variables in research design.
3. Enlist and explain different research designs.
4. Explain principles of experimental designs.
5. Enlist and explain informal and formal experimental designs.
6. Define and differentiate among C.R.D., R.B.D. and L.S.D. experimental designs.

# Unit 5 : Plan of Study

***Index***



## 5.1 Introduction

The plan of study implies locating, studying and evaluating reports of relevant researches, study of published articles going through related portions of Encyclopaedias and research abstracts, study of pertinent pages out of comprehensive books on the subject and going through related manuscripts if any. With the help of such plan we can avoid the duplication of work and carryout our research successfully in stipulated time and correct methodology. Such type of plan of study helps in the assessment of research.

**After the study of this unit, you will be able to know and understand :**

- Delimitation of Scope of Investigation
- Choice of Research Topic
- Classification of Data
- Sequence of Study
- Source of Information Glossary

## 5.2 Content

### 5.2.1 Delimitation of Scope of Investigation

It is a statement of the limits or scope of the investigation. It will determine the boundaries of the project in hand. This delimitation will mention the geographical limits of the study i.e. whether the study will be covering a single town, a district, a region, a state or a country. Again it will be important to mention as to how many subjects will constitute the sample of the study and how they will be distributed over the institutions, geographical areas or time intervals.

**Justification of the Problem**

The urgency and worth whileness of the project has to be justified. It will convince the readers about the significance of the investigation. This step would prevent wastage of research effort on unimportant, trivial, superficial or insignificant problems. The problem should be broad-based enough to provide an investigation of real significance.

**Evaluating the Problem**

Before the proposed research problem can be considered appropriate, several searching questions should be raised. Only, when those questions are answered in the affirmative, can the problem be considered a good one.

(a) Is this type of problem that can be effectively solved through the process of research ? Can relevant data be gathered to test the theory or find the answer to the problem under consideration ?

(b) Is the problem significant ? Is an important principle involved ? Would the solution make any difference as far as theory or practice is concerned ? If not, there are undoubtedly more significant problems waiting to be investigated.

(c) Is the problem a new one ? Is the answer already available ? Ignorance of prior studies may lead a researcher to spend time needlessly on a problem already investigated by some other researcher.

Although novelty or originality is an important consideration, the fact that a problem has been investigated in the past does not mean that it is no longer worthy for study. There are times when it is appropriate to replicate or repeat a study to verify its conclusions or to extend the validity of its findings to different situations.

(d) Is research on the problem feasible ? After a research project has been evaluated, there remains the problem of suitability for a particular researcher.

   (i) Am I competent to explain and carry out a study of this type ? Do I know enough about this field to under-stand its significant aspects and to interpret my findings ?

   (ii) Are pertinent data accessible ? Are valid and reliable data-gathering devices and procedures available ?

   (iii) Will I have the necessary financial resources to carry on this study? What will be the expense involved in data-gathering, equipment, printing, test materials, travel, and clerical help ? If the project is an expensive one, what is the possibility of getting a grant from a research foundation ?

   (iv) Will I have enough time to complete the project ? Will there be time to devise the procedures, select the data-gathering devices? Gather and analyze the data, and complete the research report ?

   (v) Will I have the courage and determination to pursue the study in spite of the difficulties and social hazards that may be involved ? Will I be willing to work aggressively when data are difficult to gather and when other are reluctant to co-operate ?

## 5.2.2 Choice of Research Topic

The choice of a topic for research is a commitment of one's time and efforts in a particular direction. There should not be any haste in deciding on the topic, nor in defining its scope.

The choice of a suitable topic for research is in many ways the most difficult task. There are many pitfalls to be avoided and there are many students who have failed to complete their research, not because they were lazy or badly organized, but because their topics were not suitable for research. Probably the most common mistake is to choose a topic that is too large at the level at which the student is studying/ working; in their enthusiasm undergraduates / postgraduates often consider undertaking research that would overawe a M. Phil., or M. Litt, or even Ph.D., student. A topic, which may be suitable for a doctoral thesis, may not be so for a dissertation for a graduate or postgraduate degree, which has to be completed in less than two semesters. Obviously, a single undergraduate cannot usually undertake extensive social research, involving large samples and presenting considerable problems in terms of analysis. Some topics are simply not researchable at any level. This may be because the sources do not exist or the difficulty may arise because no means has yet been devised for investigating that particular field.

*Time* is another important factor. If for one reason or another, gathering the information takes many months or even years, then the topic is obviously not suitable for an undergraduate / post-graduate student with only a few months to complete his project. It is advisable to the undergraduate student to confine himself to research based on local sources.

**Topic Suggested by Research Guide / Choice Oneself**

There are basically two ways in which one arrives at a topic. Whether the research adviser / guide suggests it or one arrives at the choice oneself. In the former case, the chances are that a topic suggested by one's guide will be suitable for research and that its scope will be sufficiently limited. On the other hand, the main pleasure to be derived from research is to immerse oneself in a topic in which one is really interested; if the student finds the subject boring then the special study will become millstone around his neck. For this reason, therefore, it is probably better to choose the topic oneself. Hence, it is important for the student to understand the ultimate responsibility for the choice of topic is his and that, although the research adviser may offer advice, he will expect the student to make the final decision.

## 5.2.3 Classification of Data

Once the data is collected and edited, the first task of the statistician is the organization of the figures in such a form that their significance, for the purpose in hand, may be appreciated that comparison with masses of similar data may be facilitated, and that further analysis may be possible. This is done through classification and tabulation. But before tabulating the data into different homogeneous classes, it is necessary to sort out the relevant and significant features from the irrelevant and insignificant ones. The process of arranging the data into groups or classes according to resemblance and similarities is technically called classification.

It is of interest to give below the following definitions of classification.

Classification is the process of arranging data into sequences and groups according to their common characteristics, or separating them into different but related parts.

A Classification is a scheme for breaking a category into a set of parts, called classes, according to some precisely defined differing characteristics possessed by all the elements of the category.

Thus classification means the 'arrangement of the data into different classes which are to be determined depending upon the nature, objective and scope of the investigation. For instance, the number of students registered in Delhi University during the academic year 1983 may be classified on the basis of any of the following criterion :

- Sex.
- Age.
- The state to which they belong.
- Religion.
- Different faculties, like Arts, Science, Humanities, Law, Commerce, etc.
- Heights or weights.
- Institutions (Colleges) and so on.
- Large, Marginal and small farmers
- Types of soil
- Different cultivation practices

Thus the same set of data can be classified into different groups or classes in a number of ways based on any recognizable physical, social or mental characteristic which exhibits variation among the different elements of the given data. The facts in one class will differ from those of another class with respect to some characteristic called the basis or criterion of classification.

As an illustration, the data relating to socio-economic enquiry, e.g., the family budget data relating to nature, quality and quantity of the commodities consumed by the group of people together with expenditure on different items of consumption may be classified under the following heads :

- Food
- Clothing
- Fuel and Lighting
- House Rent
- Miscellaneous (including items like education, recreation, medical expenses, gifts, newspaper, etc.).

Each of the above groups or classes may further be divided into sub-groups or sub-classes. For example, 'Food' may be sub-divided into cereals (rice, wheat, maize, pulses etc.); vegetables; milk and milk products; oil and ghee; fruits and miscellaneous. Thus it may be understood that to analyze any statistical data classification may not be limited to one criterion or basis only. We might classify the given data with respect to two or more criteria or bases simultaneously. This technique of dividing the given data into different classes with respect to more than one basis simultaneously is called cross-classification and this process of further classification may be carried on as long as there are possible bases for classification. For instance, the students in the university may be simultaneously classified with respect to sex and faculty or with respect to age, sex and religion, three criteria simultaneously and so on.

**Objectives of Classification**

The chief objectives of classification are :

*(i) To present the facts in a simple form* : Classification process eliminates unnecessary details and makes the mass of complex data, simple, brief, logical and understandable. For example, the data collected in a population census is so huge and fragmented that it is not possible to draw any conclusion from them. When these massive figures are classified according to sex, education, marital status, occupation, etc. then the structure and nature of the population can easily be understood.

*(ii)* *To bring out clearly points of similarity and dissimilarity* : Classification brings out clearly the points of similarity and dissimilarity of the data to that they can be easily grasped. Facts having similar characteristics are placed in a class, such as educated, uneducated, employed, unemployed etc.

*(iii)* *To facilitate comparison* : Classification of data enables one to make comparison, draw inference and locate facts. This is not possible in an unorganized and unclassified data. If marks obtained by B. Com. Students in two colleges are given, no comparison can be made of their intelligence level. But classification of students into first, second, third and failure classes on the basis of marks obtained by them, will make such comparison easy.

*(iv)* *To bring out relationship* : Classification helps in finding out cause effect relationship, if there is any in the data. For example, data of small-pox patients can help in finding out whether small-pox cases occurred more in vaccinated or unvaccinated population.

*(v)* *To present a mental picture* : The process of classification enables one to form a mental picture of objects of perception and conception. Summarized data can easily be understood and remembered.

*(vi)* *To prepare the basis for tabulation* : Classification prepares the basis for tabulation and statistical analysis of the data. Unclassified data cannot be presented in tables.

Though many advantages are derived from classification, some amount of details is lost in the process of summarization. The greater is the extent of summarization, the more is the loss of details. The statistician will have to weight and balance the advantages he would derive from summarization the loss of details he will have to suffer, and decide on the extent of summarization he will have.

**Characteristics of Classification**

When we make a classification, we break up the subject-matter into a number of classes. It is important that the classification should possess following characteristics:

*(i)* *Exhaustive* : The classification system must be exhaustive. There must be no item which cannot find a class. There must be a class for each item of data in one of the class. If classification is made exhaustive, there will be no place for ambiguity. For example, a classification of persons by conjugal condition having two classes – 'married' and 'single' is not exhaustive. There will arise doubt in several classes where persons are widowed or divorced. Therefore, in order to remove doubt, classification should be made exhaustive.

*(ii)* *Mutually exclusive* : The classes must not overlap. That, is, each item of data must find its place in one class and one class only. There must be no item which can find its way into more than one class.

*(iii)* *Stability* : Classification must proceed at every stage in accordance with one principle, and that principle should be maintained throughout. If a classification is not stable and is changed for every enquiry, then data would not be fit for comparison.

*(iv)* *Flexibility* : A good classification should be flexible and should have the capacity of adjustment to new situations and circumstances.

*(v)* *Homogeneity* : The items included in one class should be homogeneous.

*(vi)* *Suitability* : The classification should be conform to the objects of enquiry. If an investigation is carried on the enquiry into the economic conditions of laborers, then it will be useless to classify them on the basis of their religion.

*(vii)* *Arithmetical Accuracy* : The total of the items included in different classes, should tally with the total the universe population.

## Types of Classification

The raw data are not digestible, therefore, it must be classified for understanding the salient features of the data. There are limitless ways of classifying the data. In this section, we discuss five main types of classification :

- Classification based upon differences in kind.
- Classification based upon differences of degree of a given characteristic.
- Geographical Classification.

- Chronological Classification.
- Alphabetical Classification.
- Qualitative Classification.
- Quntitative Classification

Below we discuss these types in details :

### (I) Classification Based on Differences in Kind

It is also called qualitative classification and classes are set up on the basis of qualitative differences. The given data are classified on the basis of some quality or attribute such as religion, sex or intelligence. Here the attribute under study cannot be expressed numerically, therefore, one has to say whether the attribute is present in a population or not. Suppose, we study the attribute unemployment then we can count the number of persons at a given place or in a given community who are unemployed.

The number of classes to be formed depends upon the number of attributes studied. This type of classification may further be divided into two categories

#### (i) Simple Classification

When the data are classified on the basis on one attribute, it is named simple classification. In this type of classification, two classes are formed, one possessing an attribute and the second without an attribute. It is also called two-fold or dichotomous classification. Suppose, the attribute unemployment is studied, we get the following two classes :

Table 5.1 : Showing the data Classified According to One Attribute (Unemployment)

| *Employed persons* | *Unemployed Persons* | *Grand* |
|---|---|---|
| Total Number (500) | Total Number (150) | Total 650 |

### (ii) Manifold Classification

When more than one attribute are studied, the data are put into several classes. Suppose, attribute's unemployment and sex are studied together, the above two classes are further divided into sub-classes as depicted below :

Table 5.2 : Showing the Data Classified According to Unemployment and Sex

| *Employed Persons* | | | *Unemployed Persons* | | | *Grand Total* |
|---|---|---|---|---|---|---|
| *Males* | *Females* | *Total* | *Male* | *Females* | *Total* | |
| Number of employed males (380) | Number of employed females (120) | 500 | Number of unemployed males (105) | Number of unemployed females (45) | 150 | 500 +150 = 650 |

The above classification may further be divided into sub-classes if three attributes, namely, unemployment, sex and place are studied.

### (II) Classification Based on Differences of Degree of a Given Characteristic

The classification of statistical data based on differences of degree of a given characteristic is also called quantitative classification. The most common method of expressing the difference of degree is in quantitative terms. The differences of degree of a given characteristic are given numerical expressions such as height is measured in inches, weight in kilograms, income in rupees and so on. For illustration, the following table shows the number of persons classified according to income :

Table 5.3 : Showing the Number of Persons According to Income per month

| *Income Rs.* | *Number of Persons* |
|---|---|
| 3000-3999 | 15 |
| 4000-4999 | 18 |
| 5000-5999 | 23 |
| 6000-6999 | 30 |
| 7000-7999 | 16 |
| 8000-8999 | 8 |
| Total | 110 |

In this type of classification, we generally, encounter two terms, namely :

(i) Variable

(ii) Frequency

In the above illustration, income is variable and the number of persons frequency. Below, we explain these two terms :

**(i) Variable**

A characteristic which is expressed in numbers is called a variable. When a characteristic varies from one individual to another, it is known as variable or variate. In statistical terminology, varying sizes are called variate values and only variables have variate values. Examples of variables are height of students, earnings of factory workers, production of food grains, prices of commodities; or weights of babies born on a particular day. The list of such examples is endless. In brief, variable is any characteristic which can be quantitatively measured and its measurements are variate values. Variables are of two types.

***(1) Discrete Variables***

A discrete variable is one which assumes only discrete values i.e isolated, particularly integral values. There is a definite gap between two values. Examples of discrete variables are the number of children in a family, the number of rooms in a house, the number of books in a library or the number of students in a class. Counting gives rise to discrete variables. The data, which are described by discrete variables, are called discrete data.

***(2) Continuous Variables***

A continuous variable may be defined as one, which assumes any value within a certain range such as height, weight or income. For example, the height of a person may be measured upto a fraction of an inch or weight upto grams. Measurement gives rise to continuous variables. The data, which are described by a continuous variable, are known as continuous data.

### *(ii) Frequency*

Frequency may be defined as the number of times a value appears in a series. Suppose, 30 workers of a particular firm are paid Rs. 3000 each as monthly wage, the frequency of this wage is 30. A frequency distribution may consist of many classes and each class has a number of items that fall within the range of its interval and this number of items is called frequency of that class. For example in Table 5.3, 23 is the frequency of the class 5000-5999 or 8 is the frequency of the class 8000-8999.

## (III) Geographical Classification

In this type of classification, the data are classified according to the geographical location such as continents, countries, states, districts or other sub-divisions. For example, per acre yield of wheat in major wheat-producing countries depicted below:

Table 5.4 : Showing the Per-Acre yield of Wheat in Major Wheat Producing Countries

| *Name of the country* | *Yield of Wheat (Quintals/Acre)* |
|---|---|
| America | 45 |
| Canada | 42 |
| Australia | 39 |
| India | 25 |

## (IV) Chronological Classification

When the given data are classified on the basis of time, it is named Chronological classification. In this type of classification, the data may be classifed on the basis of time, i.e., years, months, weeks, days or hours. For illustration, we present the population along with density of population in India from 1931 to 1971 below :

Table 5.5: Showing the Population and Density of Population in India from 1931-71

| *Years* | *Population (Crores )* | *Density of Population (per sq. km)* |
|---|---|---|
| 1931 | 27.89 | 88 |
| 1941 | 31.85 | 100 |
| 1951 | 36.09 | 113 |
| 1961 | 43.90 | 138 |
| 1971 | 54.73 | 182 |

**(V) Alphabetical Classification**

When the data are arranged according to alphabetical order, it is called alphabetical classification. This type of classification is mostly adopted for data of general use because it aids in locating the items easily. For illustration, state-wise density of population in India is depicted in an alphabetical order.

Table 5.6 : Showing the Density of Population in an Alphabetical Order

| Name of States | Density of Population (per sq. km) |
|---|---|
| Andhra Pradesh | 157 |
| Assam | 150 |
| Bihar | 324 |
| Gujarat | 136 |
| Haryana | 225 |
| Himchal Pradesh | 62 |
| Kerala | 548 |

**(VI) Qualitative Classification**

When the data are classified on the basis of qualitative Phenomenon which are not applicable to quantitative measurement then it is termed as the qualitative classification.

<u>Remark</u> : The qualitative data can be observed, seen, judged but cannot be measured.

e.g. Intelligence
→ Genius
→ Highly intelligent
→ Average intelligent
→ Below Average
→ Dull

### (VI) Quantitative Classification

When the data are classified on the basis of quantitative measurement like age, weight, price, income, production etc. is termed as quantitative classification.

e.g. Daily earning in Rs. In a office.

| Daily earning | No. of persons |
|---|---|
| 101 – 200 | 6 |
| 201 – 300 | 14 |
| 301 – 400 | 18 |
| 501 – 600 | 3 |

### Rules for Determining the Class Intervals

No hard and fast rules can be given but the following few facts must be taken into account in determining the class intervals -

- The range is the difference between the largest and smallest observation in the given data. The range is divided into a suitable number of classes by means of class intervals.
- The choice of the number of class intervals basically depends upon the number of items to be classified, the magnitude of items and the accuracy desired. It also depends upon the case of calculation for further processing of data.
- The actual number of class intervals would also depend on the size of class intervals because the number of classes and size of class intervals are inversely related. If one is increased, the other is automatically squeezed. Therefore, both must be kept in mind to form a suitable distribution.
- According to F.C. Hills ‘In deciding upon the size of class intervals which is deciding the number of classes one fundamental consideration should be born in mind viz., the classes should be so arranged that there will be no material departure from an even distribution of cases within each class. This is necessary because interpreting the frequency table and in subsequent calculations based upon it the mid-value of each class is taken to represent the value of all cases falling within the class.

- Number of classes should be so determined that an orderly and regular sequence of frequencies is secured.
- It is better of the class interval of all the classes are equal.
- Classes with zero frequency should neither be added with other classes nor they should be omitted.
- A general rule for determining the classes is to have 5 to 15 classes. The choice of actual number of classes will depend on the number of observations and the size of class interval required.

**Tabulation**

The systematic presentation of information contained in the data in rows & columns according to characteristics is known as tabulation.

**Role of Tabulation**

(1) It simplifies complex data.

(2) It facilitates comparison.

(3) It gives identity to data.

**Parts of Table**

1. Table number
2. Title of table
3. Caption
4. Body of the table
5. Foot-note
6. Source note

**Types of Tabulation**

Type of Tables
- On the basis of objectives
  - General Purpose tables
  - Special purpose tables
- On the basis of nature of of enquiry
  - Primary tables
  - Derived tables
- On the basis of coverage
  - Simple table
  - Complex table

**(1) General Purpose Tables**

General-purpose tables are also called as reference tables. They provide information for general use. e.g. tables published by the government.

**(2) Special Purpose Tables**

Special purpose tables are of analytical nature & prepared for studying relationship & comparison.

e.g. tables like ratios, percentages etc.

**(3) Primary Tables**

In primary tables facts are presented in original form. It contains actual & absolute figures.

e.g. weight of students (Kg.) : 72.3, 60, 58.8, 32.9

**(4) Derived Tables**

Derived tables contain figures & results derived from original data.

e.g. percentage, ratio, average, dispersion.

**(5) Simple Tables**

In simple tables data are classified with respect to single characteristics.

e.g. Paddy yield/ ha in southern states.

| *States* | *Yield (q/ha)* |
|---|---|
| Andhra Pradesh | 18.7 |
| Tamilnadu | 14.6 |
| Kerala | 13.2 |
| Karnataka | 15.3 |

**(6) Complex Tables**

In complex table data are classified or grouped into two or more characteristics simultaneously.

Table 5.7 :Area, production and productivity of cotton in Marathwada Region.

| *District* | *Area (ha)* | *Production (quintals)* | *Productivity (Kg/ha)* |
|---|---|---|---|
| Nanded | 224000 | 761600 | 340 |
| Parbhani | 242000 | 871200 | 360 |
| Beed | 58000 | 226200 | 390 |
| Jalna | 104000 | 270400 | 260 |
| Aurangabad | 90000 | 180000 | 200 |

## 5.2.4 Sequence of Study

Sequence of research problem is an essential step in preparing a model.

(1) *Review of Earlier Literature* : Reviewing of the literature on the area of research is a preliminary step before attempting to plan the study. It is essential to review all the relevant material connected with the problem chosen. It is necessary to show how the problem under study relates to previous research studies. It is also equally important to show how his work differs with the existing literature. After going through the concerned works, one will get an insight in to the problem and thus will be able to formulate, a correct plan for his future investigation. The following steps are generally considered while preparing any type of research design.

(2) *Sources of information to be tapped :* The sources of information to be tapped vary with the interest of the researcher and the type of his study. The sources are divided into *documentary* and *field sources*. The latter include living persons who have a fund of knowledge about or have been in intimate contact with social conditions and changes over a considerable period of time. These persons are regarded as personal sources or direct sources. Documentary sources of information are those which are contained in the published and unpublished documents, reports, statistics, manuscripts, letters, diaries and so on. These sources are either primary or secondary. Primary source of information include data gathered at first hand (i.e. the responsibility for their compilation and pro-mulgation remaining under the

same authority) and summaries gathered from primary sources. Census reports may be cited as an example for the former type and statistics based on Census reports may be the example for the latter type. Participant observation, personal interview, correspondence, conference, questionnaire, and other devices may gather data from primary sources.

(3) *Development of Bibliography* : As soon as the consultation of available source is begun, the *development of bibliography* - preferably with annotations should be undertaken. Each reference should appear on a separate card or sheet with last name of the author first, his initials or given name following and then the title of the reference, publisher's name and the data of publication.

(4) *Nature of study* : The next step in formulating research design is to ascertain the nature of study, whether it is a statistical study, case study or a comparative study or an experimental study or a combination of these and other types, should be decided. Since research study is complex, the specific nature of study should be determined early and precisely.

(5) *Objectives of Study* : The objectives of the research study should be compiled in clear cut terms. The objectives, of course, differ with the nature of studies and goals to be attained. Some research studies aim to gather descriptive data, or explanatory data or data from which theoretical constructs could be deduced or data which promote administrative changes or comparisons. Sometimes hypotheses may be formulated and tested. These should be closely and clearly stated and the relationship to previous research made clear.

(6) *Socio-cultural Context of Study* : If the problem under investigation relates to human-beings, then it is necessary to ascertain the socio-cultural behaviour pattern of the persons. It is necessary to understand whether the persons adhere, deviate or even withdraw from them completely. Unless this is ascertained, it is not possible to draw useful conclusions.

(7) *Geographical areas to be covered* : It is essential to determine the geographical area to be covered in connection with the research study.

Therefore, the physical foundries of the area are to be specified in the research design.

(8) *Periods of Time to be covered or Time Dimension of the Study* : In the case of historical studies, it is necessary to determine the period to be encompassed so that exploration of the problem will be made easier and clear. This will help the researcher to attribute the conclusions to the particular period in question.

(9) *Dimensions of the Study* : It is necessary to make certain assumptions in every study, wherever these are made, they should be stated clearly. Sometimes certain limitations may have to be imposed while making the study. The limits of the proposed investigation should be clearly mentioned in the research design. *New concepts should be defined.* Precision in research can be obtained with greater care by defining even the common words in usage. It was also realized that dictionary definitions do not suffice for scientific studies. The interpretation of the findings of a study depends in part of the way the terms were originally defined. For example, the statement that the cost of living of the people in India increased by 10% during 1986, depends upon its meaning on the definition of "cost of living".

(10) *The basis for selecting the Date* : The factors of time and cost are usually important considerations in social research. It is more economical and efficient to base *studies on samples* rather than to *study the universe*. Instead of studying every case which might be included in an investigation logically, only a small portion is selected for analysis. From this analysis, useful conclusions, which are equally applicable to the universe, can be drawn. Great care is to be taken in drawing the sample from the universe. The sample should be closely representative of the universe. The size of the sample is no guarantee of its representative. Relatively small samples properly selected may be much more reliable than large samples improperly selected. *statistician Margaret Hagood suggests the following criteria as a guide in the choice of a sample* : (i) "the sample must represent the universe. (that is, it must be unbiased). (ii) the sample must be of adequate

size to produce reliable results (that is as measured in terms of specific range of error); (iii) the sample must be designed in such a way as to be efficient (that is, in comparison with alternate designed)".

(11) *Techniques of Study* : The next stage in the preparation of a research design is the determination of suitable techniques for collecting the necessary data. Here also the techniques that are normally used will differ on the basis of the nature of study. If the study requires close attention of the researcher, the observation method will be used. If the subject matter for the study is scattered over a wider area, then "questionnaires" may be used for collecting such information. If the subject matter to be collected is spread in a limited area, then the "schedule and interview" method may be used for collecting the data. Sometimes these techniques may be used collectively in the collection of data relating to the study.

(12) *The Control of Error* : The control of error applies principally to experimental studies. The experimenter has to consider what variables are operating in a given situation. In the laboratory it is usually possible to control all variables or sources or error. For field studies it is usually only possible to control key variables and to randomize others. In both cases the variables and their control need to be described.

(13) *Establish the reliability and validity of test instruments* : It is necessary in empirical studies to establish the reliability and validity of test instruments. This is, to tests provide consistent measurements and to tests in fact measure what is claimed for them? It is of crucial importance in the analytical/literary study, too, to evaluate the data collected. Hence, the desirability of consulting primary sources rather than secondary sources of information. The fewer translations or transformations of material, the less the possibility of distortion.

(14) *Chapter Scheme* : The preparation of a chapter outline is the last step in planning the thesis/dissertation and it is a useful first step in writing the rough drafts. Some of the headings may need to be changed as the investigation progresses. The final form is determined by the nature of the

study itself and by the conventions. The number of chapters that the study contains and name of each chapter must be mentioned. The basis for chapter scheme is the objectives of the research problem. The chapter arrangement must include main findings prominently in individual chapters. The first chapter is usually an introductory one and the last chapter is for the findings, conclusions and suggestions. In between is the body of the research report, which vary according to the study. Normally an optimum number of five to six will be sufficient.

### 5.2.5 Sources of Information

The sources of information that a researcher should tap vary with his interest and the type of study he has undertaken. Generally, sources of information are divided into documentary and field sources. The latter include living persons who have a fund of knowledge about social conditions and changes that take place over a considerable period of time. These people are sources of information regarding the existing state of affairs and also the observable trends in a social process. These persons are regarded as personal sources or direct sources of information. Participant observation, personal interview, conference, correspondence, questionnaire and other devices may collect data from these sources. A detailed discussion of these techniques appears in chapter five. In this section we will be focusing our attention on documentary sources.

A considerable amount of statistical data on the behaviour of its members is available in every literate community. Although these data have been accumulated primarily for purposes of administration and historical description, social science research can make good use of them. To neglect their existence often involves either a disregard of a relevant information or, if the investigator laboriously collects data that already exist, a waste of effort. These data are contained in the published and unpublished documents, reports, statistics, manuscripts, letters, diaries and so on.

These sources are further sub-divided into primary sources and secondary sources. The former provide data gathered at first hand, the responsibility for their compilation and promulgation remaining under the same authority that originally collected them. The secondary sources provide data that have been compiled from original sources and data issuing authority is different from that which controlled the collection of data at first hand. Before making use of the secondary data, it is important for the researcher to scrutinize the secondary sources very closely. "Their reliability for research work can be determined only by reference to the primary source which should be cited in notes or bibliography. This will enable anyone who so desires to make himself responsible for the facts by reference to the original source. Discrepancies appear in different secondary sources which must be settled from the original source.

**Merits of using Documentary Sources**

Following merits are usually claimed for using documentary sources.

(1) The first purpose of the use of available materials is to explore the nature of the data and the subjects to get an insight into the total situation. While looking for the data required by the researcher he may uncover many more available data than are often assumed to exist and hence contributes significantly to the unfolding of hidden information.

(2) Secondly, the use of documentary sources sometimes helps in the formulation of research hypotheses. While an investigator may have one or two hypotheses which he might have deduced from theory, the study of available materials may suggest further hypotheses. If a research idea or hypothesis can be formulated in such a manner that the available recorded material bears on the question, the use of such material becomes possible.

(3) Thirdly, the available records may also help in testing the hypothesis. For example, Beale in his study of freedom in public schools and teachers colleges, tested his hypothesis on the relation between lack of freedom and other variables and much of his data came from newspapers, periodicals, books, public documents, court decisions and so on.

(4) Fourthly, available documents may be used to supplement or to check information gathered specifically for the purposes of a given investigation. For example, if one has drawn a random sample of a small group in order

to interview individuals, the accuracy of one's sample could be checked by comparing socio-economic data of the sample, like income, education standard, caste, family size etc. with the same data of the most recent census or with available data in local Government offices.

Thus we find that "as a preliminary to field research, or in conjunction with it, a sustained and high quality search for data in the library is a most pressing need in the social sciences". To quote **Thomas Edison, "when I want to discover something, I begin by reading everything that has been done in the past. I see what has been accomplished at great labour and expense in the past. I gather the data of many thousands of experiments as a starting point, and then make thousands more".**

Clyde Kluckhohn, an eminent field researcher mentioned that, "unless a greater proportion of available source materials are collected and synthesized, field research will suffer materially, for the right questions will not be asked". According to Young, the pertinent, materials are able to : (1) provide general orientation essential in creating insight and asking the right questions; (2) They suggest the use of certain techniques of study, not thought of at the outset (3) They aid in conceptual thinking and in ways of testing tentative assumptions. (4) They help avoid necessary duplication. In short, a review of pertinent works and thinking by others help to enlarge, enrich and clarify one's own work and thinking.

A practical difficulty may, however, arise in using old statistical records. The definitions of categories used in available statistical records frequently do not coincide with those used by modern researcher. In view of such differences, use of available records may be more misleading than enlightening unless the precise definition on which the statistics are based is known. Although this restricts their usefulness, such data remain very important. They provide unique access to historical social situations and to some current social situations which are otherwise difficult or expensive to observe.

**Sources of Information in Social Science :** The main sources of research material which are of immense use to the social science researcher are given below with particular reference to the documentation services available in India.

**Types of Sources.**

1. Bibliographies
2. Indexes
3. Abstracts
4. Statistical sources
5. Directories and year books.
6. Encyclopedias
7. Other sources.

**1. Bibliographies**

Bibliographies contain information about the most important sources of research material. A bibliography mostly contains particulars of books as well as articles excluding those articles published in news papers published in a particular discipline in the span of a specified period. Following are the important bibliographies useful to social science researchers in India.

*(i)* *International Bibliography of Social Sciences* : This bibliography is published since 1952 by UNESCO with collaboration of the International Committee for Social Science Documentation. It is an annual publication and appears in four parts – Sociology, Political Science, Economics and Anthropology. This bibliography lists the important publications in the four disciplines. The bibliography is published in bilingual form – English and French – and includes both author and subject indexes and a list of periodicals consulted with their official abbreviations.

*(ii)* *Asia Social Science Bibliography:* This bibliography is a continuing project of the Institute of Economic Growth and covers (1) Social Science, Social Data, (2) Education and Communication, (3) Political Science, (4) Economics, (5) Sociology, Social Anthropology. A unique feature of this bibliography is that more than 90 per cent of the entries have been provided with necessary annotations and abstracts. This bibliography contains all significant contributions in English language pertaining to the social sciences disciplines referred to above. Author, subject and geographical indexes are provided.

(iii) *Documentation on Asia* : This document is published under the auspices of the Indian Council of World Affairs The literature surveyed in this document mainly belongs to political, economic and social developments and international relations of the countries of Asia. The series have been mainly designed for the students of International Relations and Area Studies.

(iv) *Indian National Bibliography* : (National Library, Calcutta). Indian National bibliography is an authoritative record of current publications in the major languages of India, received by the National Library, Calcutta. It consists of two parts. In the first part general publications are dealt with while second part contains information about Indian official publications.

## 2. Index

Index, though can hardly be distinguished from bibliographies, usually concentrate more on periodical literature. Important Indexes useful to social science researcher in India are :

(i) *Social Sciences Citation Index* : This international multi-dimensional index to literature of social sciences is published by the Institute of Scientific Information, 325 Chestnut Street, Philadelphia, U.S.A. This document indexes about 70,000 new articles and significant editorials from every issue of over 1,000 of the world's most important social sciences journals. These journals belong to 26 disciplines which include : Agriculture, Anthropology, Archaeology, Area Studies, Business and Finance, Communication, Community Health, Criminology and Penology, Demography, Economics, Educational Research, Ethnic Group Studies, Geography, History, Information and Library Science, International Relations, Law, Linguistics, Management, Marketing, Political Sciences, Psychiatry, Psychology, Sociology, Statistics and Urban Planning and Development.

The most important and notable feature of this indexing service is that it also cites the references cited in a particular article while it was written by its author. Generally, the indexed items consist of the author, title, journal, its volume and number, year of publication and pagination of a particular article. Hence it is a very useful source for searching periodical articles in the field of social sciences.

*(ii)* *Index to Scientific Review* : This document is also published by the Institute for Scientific Information, Philadelphia, U.S.A. Every year, it indexes review articles selected from more than 2700 of the world's most important journals. Among different areas of sciences, it also includes "Social and Behavioural Sciences."

*(iii)* *Social Sciences Index :* This index is published by H.W. Wilson and Co. New York and appears in two parts (i) Social Sciences Index and (ii) Humanities Index.

The important indexing services in India are,

(a) Index India

(b) Guide to Indian Periodical Literature.

(a) *Index India* : This document is being published quarterly by the *Rajasthan University, Jaipur* since, 1967. It indexes articles appearing in Indian news papers, Indian periodicals, foreign periodicals, composite publications, biographical profiles, book reviews, index to these and dissertations.

(b) *Guide to Indian Periodical Literature* : It is also being published quarterly since 1964 by Indian Documentation Science, Gurgaon. It contains author-subject index to articles, research papers, notes, conference proceedings and book reviews from about 300 Indian Journals in social sciences and humanities. It also covers the daily *Times of India* for news, signed articles and significant editorials. It is, thus a basic reference tool on social sciences in India.

### 3. Abstracts

Abstracts give a gist of the content of an article, book, dissertation etc. Important abstracts useful to a social science researcher are as follows :

(i) *World Agricultural Economics and Rural Sociology Abstracts (1959)* : This abstract is published quarterly by North Holland Publishing Co. Amsterdam. It is an international abstract journal covering the literature of agricultural economics and rural sociology in the broadest sense, including agricultural policy, land reforms, farm labour, marketing agrarian reforms etc. as well as agricultural education, legislation, geography and history.

(ii) *Dissertation Abstracts International* : It is published monthly by Xerox University Microfilms, Ann Arbor, Michigan, U.S.A. It is a compilation of abstracts of doctoral-dissertations submitted to Xerox University Microfilms by more than 350 co-operating institutions in United States and Canada. Noro, dissertations from European Universities are also included in it.

(iii) *Indian Behavioural Science Abstracts* : This document is published quarterly by Behavioural Science Centre, 32 Subhash Marg, Delhi. It contains abstracts of behavioral science literature produced in India or abroad. It covers such field as psychology, sociology, Social anthropology and behavioral aspects of economics, education management, political science, public administration, extension education, social work, history and human geography. It is very useful document to the social science researchers as it provides detailed abstracts of articles, books and research publications as well as unpublished doctoral and master's theses.

(iv) *Indian Dissertation Abstract* : It is a quarterly publication made by Indian Council of Social Science research, New Delhi. It publishes abstracts of doctoral dissertations accepted by the Indian Universities in social sciences. The disciplines included within the scope of this journal are economics, education, management, political science, psychology, public administration, sociology and social science aspect of anthropology, demography, geography, history, law and linguistics.

**4. Statistical Sources**

Governments or other agencies in most of the countries of the world regularly publish a large variety of data on social problems. The statistical data thus released are of immense use to social science researcher and are very much valued for their coverage, regularity and reliability. The most frequently consulted statistical publications are :

*(i)* *United Nations Statistical Year Book* : This is published annually by the U. N. Department of Economic and Social Affairs, New York and contains country-wise statistics on population, national income, agricultural and industrial production, energy, external trade and transport.

*(ii)* *Demographic Year Book* : This year book is published by United Nations, Department of Economic and Social Affairs, New York and contains statistics of area, population, mortality etc. from every country of the world. It also contains census data on the geographical distribution of the population, on house-hold numbers, size and on personal and economic characteristics of the population.

*(iii)* *Production Year Book* : It is an annual publications of Food and Agriculture Organization (F.A.O.), Rome. It contains annual data on all important aspects of food and agriculture, including population, prices, freight rates and wages.

*(iv)* *Census Publications* : Census publications in India remain the largest single source of data in respect of the land and its people. First census in India was undertaken between 1865-1872. Second census was undertaken in 1881 and thereafter every ten years without a break. The latest census in India was taken in 1991. The All-India tables and the analytical report volumes as well as of each state and Union territory are divided into parts covering different subjects.

*(v)* *Statistical Abstracts, India* : It is another annual publication made by the Central Statistical Organisation, Department of Statistics, Ministry of Planning, Government of India. It is a very useful document and provides data on different sectors of the Indian economy.

| *Part* | *Subject* |
|---|---|
| I | General Report |
| II | All India census table on 1% sample basis |
| II A | General population tables |
| II B | Economic tables |
| II C | Social and cultural tables |
| II D | Migration tables |
| III A | Establishment report |
| III B | Establishment tables |
| IV A | Housing Report |
| IV B | Housing tables |
| V | Special tables and ethnographic notes on SC+ST |
| VI A | Town Directory |
| VI B | Special survey reports on selected towns |
| VI C | Survey reports on selected villages |
| VII | Special tables on degree holders and technical persons |
| VIII | Administration report on enumeration and tabulation |
| IX | Census Atlas |
| X | District census hand books and village and town directories |

*(vi)* *Indian Agriculture in Brief :* It is an annual publication by the Union Ministry of Agriculture and Irrigation, Directorate of Economics and Statistics. It mainly contains data regarding Indian agriculture.

*(vii)* *Statistical Outline of India* : It is another statistical annual publication published by the Tata Industries Pvt. Ltd., Bombay.

*(viii)* *Labour Statistics* : This document is published annually by the Ministry of Labour, Government of India and contains mainly Indian Labour Statistics.

**5. Directories and Year Books**

Year books and directories provide recent information and the main trends in subjects of common interest, and are useful sources of general information. Important Directories and year books available to a researcher are :

*(i)* ***Year Book of International Organizations.*** This year book is being published since 1948 by the Union of International Associations, Brussels. It contains information about international organizations. Information for each entry includes : name, address, very brief history, membership, offers, finances, current activities and publications in brief.

*(ii)* ***Directory of Social Science Research Institutions in India.*** This directory has been compiled by Indian Council of Social Science Research, New Delhi. The directory covers research institutions in the field of social sciences which are outside the university system.

**6. Encyclopedias**

Encyclopedias contain informational articles on subjects in every field of knowledge usually arranged in alphabetical order. Two important encyclopaedias useful to a researcher in social sciences are: (i) International Encyclopaedia of the Social Sciences, edited by David Sils, New York, Macmillan Co. and Free Press and (ii) Encyclopedia of the Social Sciences, edited by R.A. Edwin, New York, Machmillan Co. (1930-1957) 16 volumes.

## 5.3 Glossary

**Classification :** Classification is the process of arranging data in groups or classes according to their similarities and resemblance.

**Geographical Classification** : The basis of classification is geographical or location difference in various items, viz., States, Cities, Regions, Zones etc.

**Chronological Classification** : Classification in which data are classified on the basis of time is called as chronological classification.

**Qualitative Classification** : When the data are classified on the basis of qualitative Phenomenon which are not applicable to quantitative measurement then it is termed as the qualitative classification.

**Qualitative Classification** : When the data are classified on the basis of quantitative measurement like age, weight, price, income, production etc. is termed as quantitative classification.

**Tabulation** : The systematic presentation of information contained in the data in rows and columns according to characteristics is known as tabulation.

**Variable** :The name given to the particular object is called as variable. e.g. Sex, height etc.

**Variate** : A value given by particular variable is called as variate. e.g. Weight of student is 70 kg. Weight is variable and 70kg. is variate.

**Continuous variable** : The variable, which takes all possible values in a given specified range and contains fractional values are termed as continuous variable. e.g. height, weight, yield etc.

**Discrete variable** : Those variable which we can count are called as discrete variable. e.g. No. of students, no. of insects, no. of plants etc.

**Direct Sources :** The sources, which provide direct description of the study conducted are called as direct sources. e.g. Books, Journals, Monographs, dissertations, thesis, Government Publications etc.

**Indirect sources :** The sources, which include publications written by authors and which can not observe directly are called as indirect sources. e.g. Abstracts, Indexes, Directories, Bibliographies, Encyclopaedias etc.

**Quotations :** The reader reproduces exact words of the author and copies the material accurately is called as quotation.

**Summary :** The reader records content of the material read by him in condensed form is called as summary.

**Evaluation :** The reader interprets the points of view of the author, records his own impression and reactions and indicates agreement or disagreement with the author.

## 5.4 Summary

The proper planning of the research study, researchers can avoid the duplication of work and carry the research in stipulated time. It is important for the researcher to understand the ultimate responsibility for the choice of topic. After finalisation he has to collect the appropriate data and make classification into different classes so as to meet the objectives and scope of the enquiry. The objectives of classification are to present the facts in a simple form, facilitate comparison, relationship and basis for tabulation. The classification should possess the characteristics such as exhaustive, stability, flexibility, homogeneity, suitability and accuracy.

There are several types of classification of data however five important classifications are discussed which are based on differences of degree, kind, geographical, chronological, alphabetical and qualitative classifications. Sequences of research problem is an essential step in preparing a model. The different steps are Review of Literature, Sources of Information, Development of Bibliography, Nature of Study, Objectives, Socio-cultural Context, Geographical Areas, Time Dimensions, Selecting data and Techniques, Control of errors and so on.

The sources of information depends on interest and type of study undertaken, which is mainly devided into documentary and field sources. The main sources of research material is of immerse use to the social sciences researcher which can be available from Bibliographies, Indexes, Abstracts, Statistical Sources, Directory and Year Books, Encyclopedias, etc.

## 5.4 Exercises for Practice

**Answer the following questions in 200 words each**.

1. Define classification of data and explain objectives of classification of data.
2. State and explain characteristics of good classification system.
3. Enlist and explain types of classification.
4. What are the sequential steps involved in preparing model for research problem?
5. What are the sources of information and their significance in research project ?

# Unit 6 : Sampling Methods

***Index***



## 1.1 Introduction

In Agricultural research collection of data is an integral and essential part, results are based on the data collected. Population Data can be collected either by census enquiry or by sample enquiry. In the census enquiry all units of population are included while in the sample inquiry only representative units are observed and information is recorded. Sampling saves labor money and time. Sampling enquiry is more adaptable than census enquiry. Sample enquiry can be reliable as it is conducted in a scientific way and results are applicable to the universe from which samples are drawn. Sample enquiry or testing is essential in many more cases wherein census enquiry is not possible: In a lot of grains or fruits decisions are taken on the basis of the samples drawn and not on whole produce.

## 6.2 Content

### 6.2.1 Complete Enumeration

It is also known as census, data are collected for each and every unit. If there are 10000 plants in a field and height is to be measured, plant height is recorded for each and every plant. The advantage of this type of survey is that no unit is left out and hence greater accuracy may be achieved. The efforts, money and the time for complete enumeration will be extremely large. Therefore, the data is collected by adopting sampling techniques.

**Sampling :** In sampling technique instead of every unit of the population (Universe) a selected part is studied and conclusions are drawn for the Universe. The idea of sampling is old, a handful of grains are sufficient to ascertain about a lot. Two three grains of boiling rice are sufficient to know about complete pot of rice is cooked or not. One or two cobs can give information whether the crop is ready for harvesting.

**Purpose of sampling :** The basic object of sampling is to study about population. Sampling is a tool to know the characteristics of population.

*Population : A total collections of units possesing certain characteristics is called as population.*

*Sample : A part of the population selected for study.*

**Sampling Unit :** A sampling unit is the unit in the sample on which actual measurement is made. A sample must necessarily be smaller than the population. In field experiments some of the commonly used sampling units are leaf, plant, a group of plants per unit area. An appropriate sampling unit will differ among crops, among characters to be measured and among cultural practices. Thus in the development of a sampling technique, choice of an appropriate sampling unit should be made to fit the requirements and specific conditions.

**The important features of an appropriate sampling unit are**

***Easy to identify:*** *A sampling unit should be easy to identify*

***Easy for measurement :*** *The measurements of the characters should be made easily.*

## 6.2.2 Types of Sampling Designs

**Sampling design :** A sampling design specifies the manner in which the “n” sampling units are to be selected from the population. The commonly used sampling designs are :

**1. Simple Random sampling**

It a population contains finite number of of units, it is a finite population. If the finite population contains N distinct units, then samples of n ($n \leq N$) units can be drawn from the population. A simple random sampling is a method of selecting a sample of "n" units such that every one of all possible samples has an equal chance being selected or that the chance of selecting every unit is same.

The random samples can be selected by (i) Random number method, (ii) Lottery method (iii) Random pair method

***(i) Random number method :*** Samples can be selected using random numbers. There are three random number tables available i.e. (i) Tippett's table of random numbers (ii) Fisher & Yate's table of random number and (iii) Kendall & Babington Smith's table of random number. Tippett's tables are popular. Select the starting point in the table of random numbers in some random manner so that every unit has an equal chance of being selected. Selection of random number table is made on the basis of digits in population. If total number of units are less than 100 select a two digit random number table to select population units as sampling units.

Suppose a sample of 10 is to be selected from 40 units. Two digit random number table is used and 10 units are selected.

| | |
|---|---|
| 29 | 30 |
| 12 | 05 |
| 01 | 20 |
| 27 | 27 |
| 23 | 14 |

The units with selected serial numbers are included in the sample. Same unit may be repeated in the sample if the sampling is done without replacement. If the sampling is done with replacement, unit once selected is replaced by another unit.

***(ii) Lottery Method :*** Lottery method is popular method of taking a random sample. In this method all items of the population are numbered on small slips of papers of identical size. Slips are folded and mixed up. A blind fold selection is made of the number of slips required to constitute a desire size of sample. The selection of items depends entirely on chance.

To select 10 unit from a population of 40. All 40 population unit are numbered from 1 to 40 on small paper slips. Slips are folded and mixed. 10 slips are selected from the 40 slips. If selected numbers are

| | |
|---|---|
| 03 | 08 |
| 09 | 14 |
| 12 | 22 |
| 28 | 29 |
| 34 | 39 |

Above 10 units from a population of 40 are selected as samples. Lottery method is popular. However, while drawing a sample a care needs to be taken that slips are of identical shape and size and there are no possibilities of prejudice and bias mind.

***(iii) Random pair technique*** : It is applicable whether or not the plot can be divided uniquely in to "n" sampling units. Mostly this method is used where clear division is not possible. Determine width (W) and length (L) of the plot in terms of sampling unit specified, select "n" pairs of random numbers with 1st number of each pair ranging from 1 to W and 2nd number ranging from 1 to L. Thus selected random pairs are

6,7 5,8 3,2 9,3 1,6 *6,7 number indicates 6th row and 7th column*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |
| * | * | * | * | * | * | * | * | * | * |

The selection of five randomly selected sample plants using random pair technique for a plot consisting 10 rows and 10 plants per row.

## 2. Multistage random sampling

In simple random sampling only one type of sampling unit is involved, multistage random sampling design is characterized by a series of sampling stages. Each stage has its own unique sampling unit. This design is suited for cases where the best sampling unit is not the same as the measurement unit. In case of leaf area is to be measured, three stage sampling design, with individual plant as the primary sampling unit, individual branch as secondary unit and individual leaves as the tertiary sampling unit.

The selection of sample is done separately and independently at each stage of sampling, starting with the first stage sampling then the second stage sampling and so on in the proper sequence. At each stage the random selection procedure is adopted. A sample of "n" units is selected as

$$n = n_1 \times n_2 \times n_3$$

$n_1$ = Number of plants as primary unit

$n_2$ = Number of branches as secondary unit

$n_3$ = Number of leaves as tertiary unit

## 3. Stratified random sampling

The population of N units is divided into K groups or strata, $4^{th}$ strata containing Nn units. Therefore N=ΣNn, n=1,2, ...... k, then simple random samples are drawn from k strata, independently, such that simple random sample of size nn is drawn from hth stratum of Nh units therefore the total number of units selected in the sampling is n and d n=Σnh. This procedure is the stratified random sampling.

The stratified random sampling is useful where there is large variation between sampling units and sources of variation follow consistent pattern. The precision of the sample estimate can be improved by first grouping the units within strata variation is as large as possible.

**Examples**

In insecticide trial blocking is based on the direction of insect migration.

Soil fertility causes substantial variability. Stratification is based on fertility pattern. If plant height within field of observations is not uniform stratification of sampling units can have taller, medium and shorter plant. Administrative convenience is also a reason for the use of stratification. Collecting information from various category of farmers, farmers are classified as small, medium, large and marginal farmers.

### 4. Systematic Sampling

It is popularly used in those cases where a complete list of population from which sampling is to be drawn is available.

In this method every $K^{th}$ item from the list is selected where 'K' refers to the sampling interval.

**Example**

If a sample of 100 units is to be selected from a population of 500. First decide K as :

$$K = \frac{\text{Size of population}}{\text{Size of sample}} = \frac{500}{100} = 5$$

K = 5

Select one item between 1 and 5 randomly say 2. Now go on adding 5 and obtain sample numbers as:

2, 7, 12, 17, 19, 24, 29, 34 , 39, 44 - - - - - -.

This method is more convenient to adopt than the random sampling.

### 5. Stratified Multistage Sampling

The stratification technique is combined with the multistage sampling technique the resulting sampling design is known as stratified multistage random sampling.

**Examples**

Farmers are grouped into four groups or strata e.g. marginal, small, medium and large to estimate crop production. Stages of selecting samples are:

| | |
|---|---|
| First stage | Selection of village |
| Second stage | Selection of farmers |
| Third stage | Unit area for crop cutting experiment. |

**6. Sampling with Probability Proportional to Size**

In the probability proportional sampling samples are drawn from each stratum with certain fixed proportion or sampling ratio. In the following example, sample size in different strata is $1/7^{th}$ of the respective strata size.

**Examples**

A group of six villages is formulated for sample selection.

| Village | No. of farmers | Sampled farmers |
|---|---|---|
| Babhulgaon | 140 | 20 |
| Rampur | 194 | 28 |
| Raipur | 110 | 16 |
| Chandur | 87 | 12 |
| Waghalgaon | 69 | 10 |
| Umri | 100 | 14 |
| Total | 700 | 100 |

From a population of 700 farmers a sample of 100 is selected by probability proportional to size.

**Sample size**

The number of sampling units taken in the sample from the population is called as sample size. Different opinions have been expressed by experts about the sample size. Sample size depends on various factors relating to the subject under investigation. However, two considerations may be kept in mind while determining appropriate sample size, i.e. cost of sampling and variation in the population.

(1) Sample size should increase with the variation in the individual items of the population.

(2) Sample size should be large for desired degree of precision for the character under study.

The optimal sample size should provide maximum precision with minimum cost.

Advantages of sampling

(i) Saves time, (ii) Minimizes the cost of data collection, (iii) Reliability of Results, (iv) Collects detailed information

$$\text{Sampling size } n = \frac{Z^2.V}{d^2.(X)^2}$$

| | | |
|---|---|---|
| n | = | Sample size |
| Z | = | Value of standard normal variate |
| V | = | Sampling varience |
| d | = | Margin of error |
| X | = | Sampling mean |

**7. Quota Sampling**

In quota sampling, the population is divided into some groups according to certain characteristics and fixed number of units selected from each group. Within quota selection of individual sample depends upon personal judgment.

To study the impact of some broadcasted information through radio in a particular area interviewers are to be told about how many are to be selected, from large farmer group, medium and small farmer group. This method is used in public opinion studies.

**8. Purposive Sampling**

In purposive sampling, units are selected in the sample with some consideration. In, Extension and Communication surveys this method is used to incorporate the known sampling units so that the proper information can be collected. Selection of particular area, village or house hold to whom researcher is familiar and the information can be collected to the extent of reliability, the variability of the known quantitative characteristics is least is made in purposive sampling.

## 6.2.3 Sampling and Non Sampling Errors

The errors arising due to drawing inferences about the population on the basis of sample is termed as sampling error. Sampling error does not exist in complete enumeration. In complete enumeration whole population is studied and hence there exist no error. The errors arising at the stages of ascertainment and processing of data are non sampling errors, these errors are common both in complete enumeration and sample surveys.

## Sampling Errors

Sampling errors are due to the fact that samples are used and to the particular method used in selecting the items from the population. In sampling the basic assumption is that all units are homogeneous (identical) and representative of the population. Some times selection of a particular sampling unit gives rise to bias or sampling error, while the variation between sampled units gives rise to unbiased sampling error. Sources of sampling errors are:

(i) An in appropriate method of selection : Method of selection should be appropriate, biased selection may give rise to error.

(ii) Errors in collection of information

(iii) Improper method of analysis

Untrained investigators, poor response of respondents, faulty editing, coding of responses are some of the errors in data collection.

Proper method of analysis and interpretation of data be used to avoid error.

## Reducing Sampling Errors

The simplest way of increasing accuracy of a sampling is to increase its size. Sampling error usually decreases with increase in sample size.

## Non Sampling Errors

In complete enumeration of units in the universe it is expected that data is free from errors. However in practice it is not true. Errors may be committed in recording observations or tabulating data. Thus data obtained may be free from sampling errors would be subjected non sampling errors, where as results of sample survey would be subject to sampling error as well as non sampling error. Non sampling errors can occur at every stage of planning and execution of the census or survey. Such errors can arise due to a number of causes such as defective methods of data collection and tabulation, faulty definition, incomplete coverage of the population or sample, non sampling errors may arise from one or more of the following

(i) Inadequate and inconsistent data specification in respect of the objectives
(ii) Inappropriate statistical unit.
(iii) Inappropriate method of interview and/or ambiguous schedules.
(iv) Lack of trained and experienced staff.
(v) Non-response of respondents.
(vi) Errors in data processing.
(vii) Errors committed during presentation and printing of tabulated results.

**Non Response in sampling**

In data collection major difficulty faced by the research workers in non response of the selected respondents. Most of the times respondents are reluctant or not willing to provide the desired information and avoid it by some supplementary reasons. In mailed questions this problem is more serious. To over come it the following steps be followed

(i) Send mailed questionnaire at least for three times.
(ii) Collect information by sending investigator.
(iii) If investigator is not able to collect information ask him to collect information from local correspondents.
(iv) If steps above (i) to (iii) failed, select another representative from the population as a sampling unit.

## 6.3 Glossary

**Sampling :** It is the method of selecting representatives of the Universe

**Sample :** Representatives of universe on which observations are recorded

**PPS Sampling**: Method of selecting samples with probability proportional to size

**Research :** Activity of solving problem

**Survey :** Collecting data on units by personal visit or through questionnaire

**Census :** Complete enumeration

## 6.4 Summary

In sampling technique instead of every unit of the population (Universe) a part is studied and the conclusions are drawn for the Universe. The idea of sampling is pretty old, a handful of grains are sufficient to ascertain about lot. Two three grains of boiling rice are sufficient to know about complete pot of rice is cooked or not. One or two cobs can give information whether the crop is ready for harvesting. The basic object of sampling is to study about population. A sampling unit is the unit on which actual measurement is made. A Sample must necessarily be smaller than population. In field experiments some of the commonly used sampling units are leaf, plant, a group of plants per unit area. The appropriate sampling unit will differ among crops, among characters to be measured and among cultural practices. Thus in the development of a sampling technique, choice of an appropriate sampling unit should be made to fit the requirements and specific conditions. A sampling unit is easy to identify if its boundary with the surrounding units can be recognized.

The measurement of the characters should be made by the choice of sampling unit. A sampling design specifies the manner in which the "n" sampling units are to be selected from the population. The commonly used sampling designs are : (1) Simple random sampling, (2) Stratified random sampling, (3) Multistage sampling

## 6.5 Exercises for practice

**Answer the following Questions in 200 words each.**

1. Define sampling and explain significance of sampling in large scale survey research.
2. Enlist and explain types of sampling designs
3. What is random sampling ? How it differ from purposive sampling.
4. What are the sampling and non-sampling errors ?
5. What are the reasons for sampling errors ?

# Unit 7 : Methods of Field Observation

***Index***



## 7.1 Introduction

The term field observation is used for facts gathered by direct survey in the field or through experimentation. Field data by and large, relate to the information of the former kind. Field observations lead to the collection of currently unknown or unavailable facts; or may relate to facts of a problem which have not been adequately enquired into or often they might be merely in the nature of verification of known facts and conclusions. The field observation involves contact between the researcher and the investigated 'Subject' or 'field' either through a schedule or a mailed questionnaire or in other schedule or a mailed questionnaire or in other ways described below. The direct contact results in the raw and mixed data, In spite of the most carefully formulated schedule or questionnaire or experiment, a lot of extraneous though perhaps valuable, material would be gathered.

## 7.2 Content

### 7.2.1 Formulation Hypothesis and Observations

The researcher has to decide whether the problem or phenomenon, which has been selected, can be studied through observation method. It means that the problem should be such that outside persons may observe it. This is to be done on the basis of hypothesis.

**Hypothesis formulation :** Through observations, researcher is able to know about the sequences and the cause and effect relationship. It is these sequences and the cause and effect relationship that forms the bases of the hypothesis. Researcher can observe and formulate hypothesis while making observation. The validity of the hypothesis can be judged on the basis of the sequences and cause and effect relationship that is observed in the phenomenon.

### 7.2.2 Field Observation

To conduct research on the basis of field observation, first thing to be decided is whether the phenomenon on can be studied by observational technique and whether the subject will allow it to be observed by an outsider. When this has been decided a plan of observation is made. The plan is based on the topic of study or hypothesis formulated. On this basis we have to decide what information would be required for proper test of hypothesis or valid generalization on the topic.

Next step is to decide the time, place and people to be observed. The process of observation may be of short duration or may be continued over long periods. If latter is the case a special arrangement may be needed for the purpose. The observation may be carried on in natural surroundings on non-controlled basis, or on a laboratory type experimental basis. If non-controlled observation has to take place, the observer has to be on the look out for the incident to be studied. If guided laboratory type experiment is to be held, the arrangement for the same must be made. In both the cases prior determination of people to be observed and the place of observation is necessary.

Observations may be carried on by the researcher himself or team of observers. If a team of observers is to be used, right type of field workers to be selected and training be arranged for them. In order to give greater validity and reliability to the collected data, use of field charts, maps, photo films, tape recorders or socio economic charts are to be made; these should be explained to the field workers.

When observation begins the field worker must be very cautious and alert so that no significant factor should be left unobserved. While recording observations observer should note the points. No bias should be introduced due to faulty preparation. For this purpose self observation is necessary. The plan of study and different aspects should be kept in mind to avoid any pit-falls but the observer should at the same time be vigilant to note any significant event outside the plan also.

**Limitations of observation method**

Observation method can study only the on going events neither the past nor the future events can be incorporated. It is often impossible to anticipate the occurrence of an event precisely enough to be able to be present to observe it. Secondly we can not have any control on the duration of the event, if it is in natural setting. Thirdly, we can not have access to every event. We can not reach every where in the social world. Fourthly, we also see that most of the observational data are qualitative in nature thus it is difficult to quantify them. Qualitative data have limited scientific validity. However, many researchers refuse this limitation and claim scientific validity.

Capacity of sense organs can not be over estimated, sense organs have their own limitations. The observers have little effect on the situations they observe. Individuals and groups adapt rather quickly to an observer's presence and to act as they usually act.

**Types of observation**

The observation as a method of data collection has become a standardized tool. In science observation involves some experiment, it may involve some kind of rather controlled observation at the same time, it may be a simple observation. If the involvement of the observer with the community is very close it may be called participant observation opposite to this mode of observation if the observer is maintaining a physical and interactional distance from the object of observation, it may be called non participant observation.

The observation are classified as :

(i) The simple observation

(ii) The non-controlled observation

(iii) The participant observation

(iv) The non participant observation

The knowledge about social relations is derived from uncontrolled observation, whether participant or non participant.

**Simple or non controlled observation**

Quite a lot of information is collected with the help of non controlled observation, no matter whether that is participant or non participant observation. Control means standardization of observational techniques or in some cases, controls over the variables in an experimental situation. In this method observation is made in natural course without extra influences, controls or guidance or instructions from external agencies and factors. He observes such conditions which are responsible for the creation of problem. In this subject matter to be observed is left free and there is no interference in that there is careful scrutiny of real life situations and no use of instruments of precision is made.

The most important advantage of uncontrolled observation is made that it becomes possible to have first hand knowledge of the group to be studied without extra influences and pressures. The difficulty with this method is the solution of complex facts can not be obtained.

**Non controlled or non participant observation**

Uncontrolled observation is of two types: Non-participant and participant observation. In non controlled non participant observation the observer does not actually participate but watches every thing from distance. The observer is very passive and does not try to influence the activity in any manner. The observation keeps penetrating the activity in any manner. The observer keeps penetrating eye on the events "Non-participant observation is then usually quasi participant observation."

The advantages of this system is that the observer maintains his status as an observer and can remain impartial. He remains away from group rivalries. Information collected is quite dependable and reliable.

**Non controlled participant observation**

Non controlled participant situation is one when the observer participates in the activities of the group whose behaviour is to be observed. It is not necessary that observer must identify himself with the group. The degree of participant depends largely on the nature of the study and the practical demands of the situation.

### 7.2.3 Aids in Field Observation

The data collected should be valid and reliable, and for this purpose, the field worker or the observer must make use of aids like charts, maps, photo films tape recorders, socio-metric charts etc. If they have to be used, their use must be explained to the field workers and practical training should be given. Schedules are most commonly used for this purpose. A great care should be taken in drafting of schedules. Before putting them in operation the schedules must be tested and properly amended in the light of difficulties experienced. Even if formal schedules are not used for the purpose of observation, some sort of log books in the form of diaries or daily record charts must be maintained. These diaries must be filled up every day at the end of observation period. They must contain the heading, the subheading and a narration of the observation made.

### 7.2.4 Reporting Observation

Information about the observation must be collected before the observation begins. At the second stage observation tools be prepared. Actual observation begins at third stage. When data is collected it should be processed to give shape of an observation report or a field observation report. If we want to maintain a higher level or reliability which using observation method for data collection the following points are taken into consideration.

(i) Limitations of our sense or sense organs.

(ii) Artificially and complex nature of human behaviour

(iii) Limitations involved in observation and interpretation of data

(iv) Personal bias of the observer

(v) Incompleteness of the social phenomenon and their transistories

It would be the first step to consider above points. There is no dispute of about the scientific way of collecting the data this should be first consideration.

It should be noted that utmost care is required at level of interpretation of data wrong and distorted interpretation of data diminishes the quality of data out of proportion. Thus, real identification of data is necessary. Natural conduct of inquiry is a fundamental contract between the researcher and the research problem. The observer must know that an event is linked with duration of time to be successful one should be aware of all these issues.

### 7.2.5 Errors in Observation

The observation method as scientific method requires special skill and training which can not be claimed by every ordinary observer, this method is more subjective and less objective. This method is least useful in studying the attitude and opinion. The following short comings are observed in observation method.

(i) An observer must be aware of the superficial phenomenon being observed. Observers misunderstanding might attach unnecessary importance. Therefore it is useful on the part of the observer to find out the reality and it is only possible if he goes deeper in the course of observation.

(ii) Observer should not apply his own stand while studying some aspects. However, it should be other way round. The observer should apply the yardsticks of cultural evaluation as prevalent in the social group under observation.

(iii) Observed phenomenon should not be examined in isolation, this approach may lead to misleading interpretation. It should be seen in totality.

(iv) In course of observation, the observer should not complete his task in hurry. He must review the phenomenon to his logical satisfaction.

(v) Observer must be aware of uniformity and regularity as evident in the events. He must examine the facts in the light of established knowledge. It is significant to examine as to what extent the observed phenomenon are helping in modifying rejecting or in establishing a new area of knowledge.

(vi) Observer must be ready to correct the errors if identified. Errors may creep in due to sense defect, it may be due to misconception or due to wrong definition or due to subjective orientation of the observer himself. These defects may not be noticed at the initial stage of observation, but a serious observer may be successful in finding out the defects at the time of closure of the observation.

Observation method is considered to be significant and very useful method, never the less, its negative effects can not be over looked as it limits the range of experience of observer. The success of observation method depends upon skill of the observer himself. Thus observer remains a variable.

## 7.3 Glossary

**Hypothesis :** Hypothesis is a statement about some assumptions, truthness of which can be tested on the basis of observations.

**Field observation** : Information collected from individual, field, plot, as a source of information which is subjected to environmental changes

**Non participant observation** : observer does not participate but watches from distance.

**Personal bias:** Bias in recording information by an investigator

## 7.4 Summary

The field observation involves contact between the researcher and the investigated 'Subject' or 'field' either through a schedule or a mailed questionnaire or in other schedule or a mailed questionnaire.. The direct contact results in the data being raw and mixed, in spite of the most carefully formulated schedule or questionnaire or experiment, a lot of extraneous though perhaps valuable, material will be gathered. Through observations, researcher is able to know about the sequences and the cause and effect relationship that forms the bases of the hypothesis. Researcher can observe and formulate hypothesis while making observation. The validity of the hypothesis can be judged on the basis of the sequences and cause and effect relationship that is observed in the phenomenon. Observations may be carried on by the researcher himself or team of observers. If team of observers is to be used, right type of field workers be selected and training be arranged to them. In order to give greater validity and reliability to the collected data, use of field charts maps, photo films, tape records or socio- econometric charts is to be made.

The observation are classified as :

(i) The simple observation

(ii) The un controlled observation

(iii) The participant observation

(iv) The non participant observation

The observation method as scientific method requires special skill and training which can not be claimed by every ordinary observer, this method is more subjective and less objective. This method is least useful in studying the attitude and opinion. The following short comings are seen in observation method.

(i) An observer must be aware of the superficial phenomenon being observed. Observers misunderstanding might attach unnecessary importance. Therefore it is useful on the part of the observer to find out the reality and it is only possible if he goes deeper in the course of observation.

(ii) Observer should not apply his own stand while studying some aspects. However, it should be other way round. The observer should apply the yardsticks of cultural evaluation as prevalent in the social group under observatin.

(iii) Observed phenomenon should not be examined in isolation, this approach may lead to misleading interpretation. It should be seen in totality.

(iv) In course of observation, the observer should not complete his task in hurry. He must review the phenomenon to his logical satisfaction.

(v) Observer must be aware of uniformity and regularity as evident in the events. He must examine the facts in the light of established knowledge. It is significant to examine as to what extent the observed phenomenon are helping in modifying, rejecting or in establishing a new area of knowledge.

(vi) Observer must be ready to correct the errors if identified. Errors may creep in due to sense defect, it may be due to misconception or due to wrong definition or due to subjective orientation of the observer himself. These defects may not be noticed at the initial stage of observation, but a serious observer may be successful in finding out the defects at the time of closure of the observation.

## 7.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. Define observation, discuss its utility as a technique of data collection.
2 Discuss the merits and demerits of participant observation.
3. Describe the steps tobe taken for organizing field observation study.
4. Assess the importance of observation as a data collection method.
5. What are the methods of field observation ? Differentiate between controlled observation and non controlled participant observation.

# Unit 8 : Methods of Data Collection

***Index***



## 8.1 Introduction

Statistical data are either primary or secondary. Data collected for the first time and are original in character are the primary data, where as the data which have been collected by some other person and passed through statistical procedures are called as secondary data. After statistical treatment primary data becomes secondary data. Methods of collecting primary and secondary data may not be identical. The data can be collected by various methods. The method selected should be such that it suits to the type of enquiry to be conducted. Statistical data refers to numerical description in the form of counts or measurements. Statistical methods or techniques are applicable only when some data has been collected. The data is some times of qualitative nature also, which can be quantified by adopting techniques like ranking, scoring, scaling or coding. The data are collected by experiments or by survey methods and they are tabulated and analysed statistically. Proper and correct inferences are drawn from these statistically analysed numerical values. These inferences lead to a final decision.

Once the objective of study is decided, it becomes necessary to collect information about the study. Mostly in the form of data for this information has to be collected from certain individuals directly or indirectly. Such a techniques is called as survey method. These are commonly used in social sciences. In survey, the required information is supplied by the individual under study or is based on measurements of certain units. Respondents are selected from population by standard sampling techniques. Another way is of collecting details by experimentation i. e. an actual experiment is conducted on certain individuals or units about which the inference is to be drawn, observations are taken on the subject under study.

**After the study of this unit, you will be able to know and understand :**

- Different types of data and their collection methdos
- Problems in data communication
- Aspects in design of questionnaire
- Organization of data from questionnaire

## 8.2 Content

### 8.2.1 Types of Data

Data are classified into (i) Primary data and (ii) Secondary data.

**(i) Primary Data:** The data which are collected from the units or individual respondents directly for the purpose of certain study or information, are known as primary data,. For instance enquiry made from each farmer in a village to know about availability of seed and fertilizer. Data obtained in census study are also termed as primary data. If an experiment is conducted to know the effect of certain fertilizer doses on the yield of crop, yields recorded in different plots of the field constitute a primary data.

**(ii) Secondary Data** : The data, which had been collected by somebody else and statistically treated and the information contained in it is used again to draw some information for other purpose is termed as secondary data.

Hence to define, Primary data which are collected for first time as original data. Secondary data are the data, which have been originally collected by someone than the investigator for some other object. The distinction between primary and secondary data in many cases in one of degree only. Data which are secondary in the hands of one man may be primary for others.

**Internal and External data**

Internal data are the data on the organizations own production or operations. While the external data are the data obtained from the reports or publications of Government, Statistical agencies or official reports.

## 8.2.2 Data Collection Methods

There are several methods of data collection.

(a) Direct personal observation

(b) Indirect oral investigation.

(c) Investigation through enumerators.

(d) Local correspondents.

**(a) Direct Personal Observation**

In this method investigator has to collect the information personally from the concerned sources. Investigator has to go to field for conducting the investigation. The method of direct personal observation is useful for intensive enquiries.

The advantages of this method are

(i) This method is very useful when the scope of the investigation is limited and the investigator has sufficient time to deal with individual items

(ii) Data collected is reliable because investigator himself collects the information

(iii) In this method investigator can always keep in his mind the object, scope and nature of the enquiry .

**Limitations**

(i) It limits the scope of investigation and if the field of investigation is large, it is difficult to cover.

(ii) Investigator may have personal bias.

**(b) Indirect oral investigation**

People having first level knowledge about the problem under study are interrogated and the desired data is collected. In this method enquiry is made through enumerators specially appointed for the purpose. In such enquires a small list of questions relating to the problem is prepared and these questions are put to different people known as witnesses and their replies are recorded.

This method has following advantages :

(i) The data is fairly accurate

(ii) Study completed with least time

(iii) More response from informants

(iv) Saves time and cost

(v) Gives not only clear picture of the problem but also suggestions.

**(c) Investigation through Enumerators**

In this method enumerators go to the informants along with the schedule and assist them, numerators explain the aims and objects of the enquiry to the informants and emphasize the usefulness of correct replies. This method is expensive.

Forms, Schedules or Questionnaires sent through Post. This method is cheap and with it information can be collected from a wide area investigation is properly conducted the method can ensure a reasonable standard of accuracy. Success of this method depends on the cooperation that the informants are prepared to give. Generally it is found that the informants are indifferent towards such enquires and do not return the questionnaire.

The merits of this method are :

(i) Large territory is covered

(ii) Information is more authentic

(iii) Direct link between the investigator and the informant

**(d) Local Correspondents**

In this method local correspondents or agents are asked to send in estimates using their own judgment as to the best way of obtaining them. Obviously such data cannot be very reliable, as such this method is used in cases where the purpose of investigation can be served with rough estimates and where a high degree of precision is not necessary while deciding the choice of method following points should be kept in mind.

(i) Nature of the enquiry

(ii) Scope of enquiry

(iii) Degree of accuracy desired

(iv) Availability of funds and time

(v) Aim of the enquiry

**Sources of Secondary Data**

The main sources of secondary data are;

(1) Official publications : Publications of the central and State governments. International bodies. Governments of Foreign countries

(2) Semi-Official Publications of semi government bodies municipal corporation/District boards etc.

(3) Publications relating to Trade : Publications of the trade association chamber of commerce bank, co-operative society, stock exchanges, trade unions etc.

(4) Journals/News papers : Journal, News papers publishing data

(5) Data colleted by Research agencies such as Universities and Research establishments also collect useful data

## 8.2.3 Major aspects in the Design of Questionnaire

The structured Questionnaire can be classified into following parts :

(i) General information

(ii) Nature of Establishment of farm inventory
(iii) Live stock inventory
(iv) Cropping pattern of the farm
(v) Management
(vi) Production
(vii) Miscellaneous

Following points should be noted while drafting questionnaire

(i) Clarity : Questions should be as simple as possible
(ii) Avoid certain types of questions : The questions which offend informants or to be answered in prejudice such questions may be avoided
(iii) Objectivity of the answers : When the enquiry is aimed at making a factual study the questions should be such as result is in objective answers
(iv) Definiteness of answers : The question should be so framed that the answers to them are perfectly definite.
(v) Number of questions : The number of questions should be consistent with the scope of the investigation
(vi) Unit of measurement : The unit in terms of which information is to be given must be clearly mentioned.

## 8.2.4 Problems in Communication

Data collection is a major and important aspect in research. Success of research is based on quality of data collected. Data collection is totally dependent upon investigators who are collecting information. Investigator should have sufficient knowledge of the problem and the skill of data collection. The important problems faced are :

(1) Scientific training in methodology of Research is a great impediment for researcher. There is paucity of competent researchers. A systematic study of research methodology is necessary. Researchers should be well equipped with all the methodological aspects.

(2) Efforts should be made to develop liaison among all concerned for better and realistic researches. There is need for developing mechanism of a interaction programme between Academicians and research worker.

(3) Confidence should be given to respondents that the information collected from them in the form of data will not be misused which will help in receiving realistic information.

(4) Over lapping research studies are under taken, it results in duplication of work and the resources of information. This problem can be solved by compilation and revision of lists of subjects, and place where the research is going on.

(5) There is need for developing a code of conduct for researchers.

(6) Adequate and timely secretarial assistance to avoid delay in completion of research

(7) Investigators must be conversant to location of the study and language spoken by respondents which will help him in to be familiar with the respondents.

(8) Investigator should have skill to collect information from respondents

(9) Researcher should be able to get recent copies of the Acts/Rules and regulations, reports and other government publications in time.

(10) Government and other publishing agencies publish data, it should be published timely

(11) Investigator and Researchers must be very clear about the scope and limitations of the study for which data is collected.

### 8.2.5 Organization of Data from Questionnaire

The data after collection, has to be organized, processed and analysed in accordance with the of developing the research plan. This is essential we have all relevant data for making contemplated comparisons and analysis. Technically speaking processing implies editing, coding, classification and tabulation of collected data so that they are amenable to analysis. Thus in the process of analysis relationships or differences supporting or conflicting with original or new hypothesis should be subjected to statistical tests of significance to determine with what validity data can be said to indicate any conclusions.

**Editing :** Editing of data is a process of examining the collected raw data to detect errors and omissions and to correct these when possible. As a matter of fact, editing involves a careful scrutiny of the completed questionnaires and or schedule. Editing is done to assure that the data are accurate, consistent with other facts gathered, uniformly entered, as complete as possible and have been well arranged to facilitate organization and tabulation.

**Central editing :** Central editing should take place when all forms or schedules have been completed and returned to central place of data collection. This type of editing implies that all forms should get a thorough editing by a single editor or a team of editor(s). Editor (s) may correct errors.

**Coding :** Coding is the process of assigning numbers or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. Coding is necessary for efficient analysis and through if the several replies may be reduced to a small number of classes, which contain the critical information required for analysis. Coding decision should be taken at the designing stage of the questionnaire. This makes it possible to pre-code the questionnaire choices and which in turn is helpful for tabulation. While coding care needs to be taken to minimize coding errors.

**Classification**: Large volume of raw data needs to be reduced into homogenous groups to get meaningful relationships. This necessitates classification of data which happens to be the process of arranging data in groups or classes on the basis of common characteristics. Data having common characteristic are placed in one class and in this way the entire data is divided into number of groups or classes. Classification can be one of the following two types. Classification according to attributes: As stated above, data are classified on the basis of common characteristics, which can either, be descriptive or numerical. Descriptive characteristics refer to qualitative phenomenon which cannot be measured quantitatively. Data obtained this way on the basis of certain attributes are known as statistics of attributes and their classification is said to be classification according to attributes. Such classification can be simple classification or manifold classification. The numerical characteristics refer to quantitative phenomenon which can be measured through some statistical units.

## 8.3 Glossary

**Data :** Numerical Description.

**Primary Data :** Data collected from individuals.

**Secondary Data :** Data collected by somebody else and statistically treated.

**Tools :** Methods of collection.

**Internal Data :** Data on individuals, institutions or organizations.

**Enumerator :** A person appointed for collecting data.

**Editing :** Process of examining row data for errors & omissions.

## 8.4 Summary

Statistical data refers to numerical descriptions form of counts or measurements. Statistical methods or techniques are applicable only when some data collection. The data is some times of qualitative nature also, they are quantified by adopting techniques like ranking, scoring, scaling or coding. The data are collected by experiments or by survey methods and they are tabulated and analysed statistically. Proper and correct inferences are drawn from these statistically analysed numerical values. These inferences lead to a final decision. Once it is decided what type of study is to be made, it becomes necessary to collect information about the study. Mostly in the form of data for this information has to be collected from certain individuals directly or indirectly. Such techniques are called as survey method. These are commonly used in social sciences. In survey, the required information is supplied by the individual under study or is based on measurements of certain units. Respondents are selected from population by standard sampling techniques. Another way of collecting details by experimentation i. e. an actual experiment is conducted on certain individuals or units about which the inference is to be drawn observations are taken on the subject understudy. The data which are collected from the units or individual respondents directly for the purpose of certain study or information, are known as primary data.

Primary data are collected for the first time as original data. Secondary data are the data which have been originally collected by some one other than the investigator for some other object.

There are several methods of primary data collection.

(1) Direct personal observation

(2) Personal interviews or indirect oral investigation.

(3) Investigation through enumerators.

(4) Forms, schedules or questionnaires sent by post.

(5) By local correspondents.

## 8.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1 Define data and explain types and nature of data.

2 State and describe methods of data collection.

3. Describe the steps involved in organization of data collected through questionnaire.

4. Distinguish between primary and secondary data and explain clearly the precautions to be taken in making use of secondary data.

# Unit 9 : The Interview

*Index*



## 9.1 Introduction

Various devices are used in the collection of primary data a schedule, questionnaire, participants observation, interview, case study and so on. Although these are not mutually exclusive and independent of one another the interview is the most important and all pervasive tool, particularly in an under developed country both because of the types of problems tackled and the environment from which data are to be collected. Illiteracy hampers comprehension of the questions and the correct and relevant responses. It demands skill and experience in the investigators in developing a rapport, in leading the initial questions, in probing, in recording and so on, if the interview is to succeed in gathering adequate, all sided and reliable data. For reliability of data it depends not only on the design-wording, construction and sequence of the questionnaire but also on the administration and technique of asking questions and eliciting answers in interviewing.

## 9.2 Content

### 9.2.1 Major Purpose of Research Interview

Interview is a device for collecting data required to test hypothesis. The principles which govern questionnaire design, interviewing and the training of interviewers are, however, relevant to most situations in which information is desired from a respondent. The fact that the interview is used very widely does not imply that it is the best device for collecting social data.

The interview is conversation with a purpose and therefore is more than a mere oral exchange of information. Its importance arises from the necessity to come into contact with individuals to get access to facts and opinions and to receive them directly from the persons. Where the source is accessible to the investigator, the interview is the device to tap it; and if it is not easily reachable, the questionnaire is the means. Apart from accessibility the controlling factor in the success of the interview is the reaction of the personalities involved – the investigator and the respondent.

The interview, as a research tool, has certain advantages, which become more important in an under developed country. Being always direct, it is the only way of which certain types of information can be obtained. It is as trustworthy a means of getting facts as participant observation is highly flexible, it permits maximum variation in directing the inquiry and here it scores over the questionnaire. People unable or unwilling to answer even the shortest questionnaire will talk, and freely too, owing to the informality of the occasion. As Eigelberner remarks "The contact of minds, a hint here and a suggestion there, leads to the birth of ideas which might not have been conceived but for the stimulation of the discussion". The flexibility and other advantages of the interview depend largely on the approach and experience of the person conducting it. If he can control the interview tactfully, lots of information and attitudes can be drawn out, specially if the direction of enquiry can be imperceptibly changed.

**Purpose of Interview :** Interview method has certain objectives, in view. The aim of interview may be finding the views of others or finding the attitude' of a group of persons. The purpose can be studied under following heads :

(1) Collecting information about unknown facts through face to face contact : This is a direct method of collecting data. In this method the interviewer tries to collect unknown facts and ideas through direct method. This is direct method for knowing the ideas or inner feelings of other persons.

(2) Formulation of Hypothesis : Through interview we get an opportunity to formulate thesis. For formulation of thesis, data or information about the problem is needed. Through interview it is possible to acquire valuable information for formulation of a hypothesis.

(3) Collecting information about qualitative facts : Social facts are basically qualitative and they are found in form of ideas, feelings, views, faith, convictions etc. These facts may be individual as well as collective. It is difficult to acquire qualitative facts easily and interview is the best method for acquiring the qualitative facts.

(4) Collecting additional information : It is a method of collecting information from normal respondents through schedule or questionnaire but there are certain person who suggest or are capable of providing additional information and it can be done only through interview method; through interview views and reactions from different persons in different situations can be known.

(5) Observational method provides an opportunity to observe things. Through interview, researcher is able to observe things, when researcher reaches a person for interview it is not a mere interview; an interviewer observes events of the surroundings, the background and other things. Through all these observations he is able to make assessment of the respondent. This type of observation adds to interview or an interview adds to observation. The combined approach through observation and interview method considered a very good method of research. Benjamin B. Paul stated "To a certain extent interview and observations are alternative techniques; one substitutes for another according to the chronological sequence".

### 9.2.2 Advantages and Limitations

Interview method is quite popular in the field of social research particularly when the study deals with the personal life of the respondents, feelings, perceptions and behavior in work situations. Interview method has following advantages.

(1) Study of the events that are not open to observation : Many events that are not known to any body except the respondent. events cannot be studied through observation method or any other method. Social activities and behaviors are of this type. Through interview it is possible.

(2) Reliable information : Information collected through interview is quite reliable. Its reliability depends upon the way of collecting information. The information like feelings, emotions, sentiments. Description of their depth can be given only by the person who is affected. Interview method helps the interviewer to contact that man and collect information. Thus infact, interviewing situation offers a better opportunity to appraise validity of reports.

(3) Studies phenomenon in Historical Background : Interview methods helps to ascertain many events in historical back ground. The man involved is real person to give information on all these feelings, emotions and sentiments these can be properly be studied only when the circumstances and the factors when they actually occurred are known and respondent can only tell things about them.

(4) Possible to study abstract factors like attitudes, feelings, emotions, reactions etc. This method helps to study the abstract and intangible personal factors like feelings emotions, attitudes can be studied. Social phenomenon influences all these factors as well and only the respondent knows the influence of these factors. Through this method it is possible to know the influence of the social phenomenon on these factors.

(5) Past events can be studied : The events may have taken place in the past and the interviewer can know about it from the person who was affected by it or who was witness to it. Through this method the difficulty that

arises in seeing things with one’s own eyes is avoided and objective study made possible.

(6) Useful for all segments of the population : Interview method is useful for all segments of the population. Unlike questionnaire, interview is very useful for illiterate population.

**Disadvantages of the interview method**

The interview method in spite of its advantages has certain weaknesses in it. The disadvantages or limitations of this method are :

(1) Bias data : The data collected through this method is considered to be biased and prejudiced. The checks applied to remove bias are not successful.

(2) Dependence on memory : In this method it is not possible for the interviewer to record things immediately. Recording has to be done after the interview. The interviewer before he has put the information gathered into black and white has to keep it in memory. This dependence on memory is likely to disturb the objective collection of the data. The memory bias renders the respondent unable to provide accurate information. The recording is also likely to be influenced by the personal bias and prejudices of the interviewer.

(3) Difference in the social background : It is not possible to have both the interviewer and the interviewee from the same social background. Because of the difference in the social backgrounds these two may have different values. They may not agree on many of the things. Their norms may not agree on many of the things. Their norms and the values are also different. This difference and discrimination vitiates the objective collection of data.

(4) Specialized knowledge is not always possible : Interview method requires specialized knowledge. The interviewer has to expert in human psychology and behavior he should be a personality of high caliber and possess many other qualities such as keen observation and contact initiative. These qualities are not always available with every interviewer and that is why the information collected suffers from draw backs.

(5) Irrelevant information : Interviewee narrates the things and interviewer has to listen and record them. If interviewer is not able to remove irrelevant details, the collected data may have irrelevant information.

(6) Subjectivity and individual feelings : The interviewee may have variation in his/her statements for the same question. This variation is on account of approach of the interviewer. This is governed by difference in social background and outlook of the interviewee and the interviewer. This subjectivity and individual feeling make data doubtful.

(7) Inferiority complex : Some times inferiority complex influences the results very much. If the interviewer feels that certain questions that are being asked, put him at a discount, he suffers from inferiority complex and that colours the answer. On the other hand when the interviewer is discarded by many respondents, he suffers from inferiority complex and his enthusiasm for the study gets damped. In that event study is vitiated.

(8) Too much importance to respondent : In this method the researcher is fully at the mercy of the respondent. He can not attain information by observation or verification. He has to behave in what the respondent says. Due to this at many occasions, answers are invalid and the data collected is unreliable.

**Limitations :** The useful ness of the device depends on quality of the interviewer, other limiting factors are time, space and expense. In average investigations, especially those conducted in under developed areas there is time only for a limited use of the interview and so the opportunity for errors is increased. The interview is limited in use to a particular locality. The limitations are :

(i) Associated with the respondent

(ii) Associated with the interviewer

(iii) Associated with the respondent : Inability or unwilling-ness to provide certain information

(iv) Associated with the interviewer : Bias of interviewer himself method of recording, follow up questions soon.

### 9.2.3 Types of Interview

Interview has been categorized on the basis of various characteristics and qualities.

Interview may be classified according to the following :

**(1) Classification of interview on the basis of object :** The interview is done with certain objects in view. From the point of view of the object, the interview may be further classified under following heads :

*(a) Clinical interview :* In this method of interview causes of certain abnormalities are ascertained to sought remedy on it.

*(b) Selection interview :* This type of interview is done to select person on the basis of certain traits and qualities. The interviewer through interview ascertains the qualities that are required and makes selection.

*(c) Diagnostic interview :* When object of the interview has been found out the serious causes of some social events or problems it is called diagnostic interview. This interview is confined to finding out the causes.

*(d) Research interview :* This is infact a kind of diagnostic interview in which we try to find out the causes of problem. In this method, a comprehensive study is made of the problem so that the causes will be found in detail.

**(2) Classification on the basis of functions and Methodology :** The interview are also classified on the basis of methodology, under following heads :

*(a) Non directed interview :* In this type of interview no classification plan is drawn about questions to be asked. This is uncontrolled interview. The informant is encouraged to exhibit and express his knowledge and views.

*(b) Focused interview :* The main objective of this type of interview is to test a particular hypothesis. The questions in this interview are pre determined and pre planned. The questions are framed on the basis of explanations about the behavior of man about which study has already

been made and hypothesis formulated. Such interviews are based on predetermined or pre studied situations.

(c) *Repeated interview :* This type of interviews are carried out after certain interviews in a repeated manner. The main task of this interview is to study those dynamic functions and attitudes to study those dynamic functions and attitudes that influence guide and determine the behavior of certain individuals. They are helpful from point of view of study of human behavior.

**(3) Classification on the Basis of number of respondents :** This interview may be further classified as :

(a) *Group Interview :* If a group of persons is interviewed for ascertaining their views or to collect information from them it is known as group interview.

(b) *Individual interview :* Interview confined on an individual to know information from him/her is called as individual interview.

**(4) Classification the basis of forms :** The interview may be classified on the basis of form in following two heads.

(a) *Structured interview or formal interview :* This is a type of interview in which the form is already determined. There is everything written about the material to be collected. The interviewer or the field worker has only to carryout the instructions. This type of schedule has following characteristics :

(i) Interviewer has no liberty or freedom, he has to act according to instructions written in the schedule.

(ii) Stimulus is specified in these types of studies and so standard studies are more fulfilled and satisfied.

(iii) In these studies schedules are standardized. Because of their structured and formal form, the objectivity is more reliable and dependable.

(iv) Field worker has to use language that has been prescribed and suggested to him.

(b) *Unstructured or informal interview :* This is an even informal type of schedule in which interviewer has full freedom. He can use language or the words that he likes. This schedule has the following characteristics :

(i) Interviewer has full freedom. He has to keep in objects of the study. This is very natural type of interview. So far as collection of data in natural form, this method is reliable. Standardization is the main point of stress.

(ii) Data collected through this method is reliable. Since there is no binding on it, as interviewer and respondent are free to ask questions and give the answers.

## 9.2.4 Some Techniques in Interviewing

Collection of reliable data is solely dependent on the techniques of the interview. Interviewer should conform the following points otherwise the purpose of collection of reliable data is defeated.

1. **Establish contact with the informants :** Interviewer must establish contact with the interviewee, he should give clear-cut idea of the scope of study nature of the data collected, he should have patience to listen and to repetition. Interviewer should not give any occasion to hesitation or announce. He should prepare interviewee to extend cooperation to him. Generally printed letter on letter head of the institution is given to know about research

2. **Start interview in an attractive manner :** Interviewer should start interview in scientific, accurate and pleasing manner. He should also try to warm up respondent and also try to make him feel at home. If needed he may discuss the problem with the respondent. He should convince the respondent that the information given shall be kept confidential. Interviewer should begin with a talk of general nature. Personal Inquiries and controversial points should be avoided.

3. **Establish rapport :** In order to get valid data interviewer has to establish rapport with the interviewee. In beginning respondent is very cautions and comes out

only with formal information. People do not like to discuss personal problems. Researcher has to be very tactful and get support and create rapport with the respondent, he should convince him that what ever information he has total is normal and not particular.

4. **Help the respondent to recall things correctly :** Respondent after narrating a particular thing becomes silent. The researcher like a patient nearer should keep quiet for some time and let the respondent begin again. Researcher may help him to refresh his memory friendly manner. He should create confidence in the respondent and encourage him to come out with the required information.

5. **Questions to find realities :** Some times respondent tells things that are not very correct and which do not fulfil the object of research. Interviewer has to be very alert to find out reality. The questions should be put in a familiar manner to know the facts.

6. **Encouragement and guiding the interviewee :** Sometimes respondent is to be encouraged while interview is going so that he continues to give information. When respondent feels that what he is telling is of interest to the interviewer he goes on giving necessary information. The researcher has to seek advice, take notes from what the interviewee is telling very valuable. Appreciation and seeking of advice is very helpful. The interviewer should not be passive listener.

7. **Recording the interview :** Interview recording is an important past of interview for data. Interview should make record of the interview. If necessary certain things should be taken but not at the cost of interview. Recording be done afterwards, interviewer should chart down certain points as important.

8. **Closing interview and submitting report :** Beginning of the interview is quite important and so is closing. The interviewer should not close the interview abruptly or in a manner that the respondent should feel that because the job has been done, therefore researcher does not bother about him. The interviewer should leave the respondent as a friend with the hope of meeting again. The closing of interview should be natural. But it should end only when job is done. At the close of interview respondent should feel that he has helped his friend.

### 9.2.5 Interview Guide

The interview guide is used as suggestive reference. According to P.V. Young "A guide may aid in"

(i) Focusing attention on salient points in the study.

(ii) Securing comparable data in different interviews by the same or by various.

(iii) Gathering the same range of items essential in the analysis of data or in testing the hypothesis formulated.

(iv) Accumulating specific concrete details as a basis for quantitative studies of life history.

### 9.2.6 References

The Respondent's Frame of reference are :

(1) Questionnaire should be casted in the language of respondent .

(2) The questionnaire must introduce each topic in a form which ties in to the perceptions of the respondent and is consistent with the respondents nations of what is and is not salient to the topic under discussion.

(3) Development of a topic from one question to another must not only meet the researcher's criteria for reasonableness and logic, it must also meet those of the respondent.

(4) Researcher must begin at the point which is respondent oriented.

## 9.3 Glossary

**Interview :** Method of collecting information

**Phenomenon :** Events

**Past events :** Events taken place in past

**Diagnostic Interview :** Interview conducted for finding serious causes

**Research Interview :** Interview conducted for comprehensive study

**Group Interview :** Interview on group of persons to know their views

**Unstructured Interview :** No specific directives to interviewer.

**Closing Interview :** Ending Interview.

## 9.4 Summary

Interview is a device for collecting data required to test hypothesis. The principles which govern questionnaire design, interviewing and the training of interviewers are, however, relevant to most situations in which information is desired from a respondent. The fact that the interview is used very widely does not imply that it is the best device for collecting social data.

The interview, as a research tool, has certain advantages, which become more important in an under developed country. Being always direct, it is under developed country. Being always direct, it is the only way of which certain types of information can be obtained. It is as trustworthy a means of getting facts as participant observation is. Being highly flexible, it permits of maximum variation in directing the inquiry and here in scores over the questionnaire. People unable or unwilling to answer even the shortest questionnaire will talk, and freely too, owing to the informality of the occasion. The flexibility and other advantages of the interview depend largely on the approach and experience of the person conducting it.

## 9.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. Define the term Interview and state and explain types of interviews.
2. Explain the importance of interview method in social research and state merits and demerits of interview method.
3. Describe various techniques in interviewing.
4. Explain the Major purpose of Interview as a method of data collection.

# Unit 10 : Project Assessment and Evaluation

***Index***



## 10.1 Introduction

Evaluation is a word which is very commonly used and which indicates assessment or appraisal of something. A significant characteristic of the last two or more decades has been the increasing number of planned projects mostly funded by public money and are planned and implemented by an organisation, which is a department of the government. In order to justify the appropriation of public funds and continuing support from the people, it is necessary that their management as well as impact be properly and adequately evaluated from time to time.

The word 'evaluation' has its origin in the Latin, word Valerie meaning to be stranger valiant. Its dictionary meanings are the determination of the value the strength or worth of something an appraisal, an estimates of the force of or making a judgement of some thing.

Evaluation may be defined as a process of systematic appraisal by which we determine the value, worth or meaning of an activity or an the value, worth or meaning of an activity or an enterprise. It is a method for determining how far an activity has progressed and how much further it should be carried to accomplish objectives.

More specific definations of evaluation are given by persons involved in development programmes. While most of these definations refer to the assessment of the results of programmes. Some definations of evaluation are :

(1) Evaluation is a co-ordinated process carried on by the total system and its individual sub system. It consists of making judgements about a planned programme based on established criteria and known observable evidence (Boone, 1985).

(2) Programme evaluation is the determination of the extent to which the desired objectives have been attained or the amount of movement that has been made in the desired direction (Boyle and Johns, 1970).

(3) It is a process which enables the administrator to describe effects of his programme and there by make progressive adjustments in order to reach is goal more effectively (Jahada and Barnit, 1955).

(4) Evaluation is a comparison of the situation before and after a development programme has operated within it for a predermined period (Mathew, 1954)

**After the study of this unit, you will be able to know and understand :**

- Basis of assessment of project
- Purpose and goals of evaluation
- Evaluation models and approaches
- Reporting the assignment

## 10.2 Content

### 10.2.1 Basis of Assessment

Evaluation is not measurement, it is an integral part of extension education. All aspects of extension work need evaluation. Evaluation does not mean mere measuring of achievements it aims to evaluate management of the programme and methods used, achievements accomplished in line with the objectives and also to determine the reasons for success or failure.

Evaluation is not exactly scientific research, it is process of collecting information as a basis for making decisions, forming judgements and drawing conclusions, we realise it has much in common with scientific research. There is difference between every day evaluation and scientific research. However, the difference is a matter of degree rather than kind. There are following five categories of evaluation which do not have sharp lines of distinction.

(i) Every day Evaluation
(ii) Self checking Evaluation
(iii) Do it yourself Evaluation
(iv) Extension Evaluation Studies
(v) Scientific Research

**Types of Evaluation**

**(1) Self evaluation** : This is to be carried out by every worker as a matter of routine. This requires the self critical attitude which is so essential for extension worker.

**(2) Internal evaluation :** Evaluation carried to by agency responsible for the planning and implementation of the porgramme.

**(3) External evaluation :** Evaluation conducted by a person or a committee outside the area of operation.

Evaluation can also be classified into

(i) Concurrent Evaluation and (ii) Ex post Facto evaluation.

**Objectives**

Objectives are expressions of the ends towards which efforts are directed objectives are defined as directions of movement.

**Level of objectives :** Objectives have following three levels.

**(1) Fundamental levels :** These are all inclusive objectives of a society, better prosperity, better citizenship, democracy.

**(2) General levels :** These objectives are general but more specific. Having better home and helping farmers to raise their farm production.

**(3) Working objectives :** These objectives are specific they may be stated in two ways.

(i) Point of view of extension worker

(ii) Point of view of people.

**Components of an objective**

There are at least three essential components of an educational objective:

(i) Learners to be affected.

(ii) Behavioral changes to be affected

(iii) Content or subject matter to be taught

**Qualities of objectives**

(1) To be effective, objectives must be based on research and on an analysis of the situation.

(2) To be valid, objectives must be based on representative thinking of the people.

(3) Objectives must be attainable

(i) Through extension education process

(ii) within the time limitations

(iii) within the physical resources of the learners

(iv) within the learner's learning ability

(4) Objectives must be significant to a relatively large number of potential participants.

(5) Objectives must be clearly stated.

(6) Objectives must be stated in terms of behavioural changes in a particular subject-matter area.

(7) Objectives must be dynamic

(8) Objectives need to have flexibility

(9) People must understand the objectives

(10) Objectives must be socially desirable

(11) Objectives must be such as can be measured or evaluated.

## 10.2.2 Purpose and Goals of Evaluation

The primary purpose of evaluation is to ascertain how to improve effectiveness. Total programme or major phases are evaluated to determine progress towards specific objectives.

1. **Programme improvement :** Evaluation is an integral part of educational process. It is focussed on improvement of this process. Taking a critical outlook, we can discover ways and means for improving educational efforts. Thus evaluation gives direction to continued improvement in programming. It also provides fresh data regarding situations essential for improving programming functions.
2. **Programme accomplishments :** Evaluation helps to determine progress with any activity. It also allows to assess results of efforts. Through evaluation strength can be assessed, weakness and value of programmes. It serves as a periodic check on the effectiveness of extension activity and teaching methods used. Thus evaluation helps to determine the degree to which specific objectives are being achieved and in the process, helps to clarify these objectives.
3. **Public relations :** Evaluation provides realistic information to report to the public, parliament and legislative bodies needs to be informed about programmes.
4. **Professional growth :** Evaluation enhances our knowledge, gives index of growth which is helpful to rectify shortcomings.
5. **Professional security :** Evaluation provides us with information, that gives us satisfaction. A feeling of accomplishment, confidence gives satisfaction.
6. **Effective workmanship :** Evaluation gives us the opportunity to work together as an extension staff to determine the effectiveness of our educational programme.

7. **Impact of extension programmes :** Evaluation helps to determine the short term and long term impact of the extension programmes in terms of social and economic dimensions.
8. **Content of the programmes :** Evaluation enables determination of whether the content is contributing to the overall objectives or not.
9. **Methods of Teaching** : Evaluation provides information as to whether the teaching methods are used effectively or not.

Thus the purpose of evaluation is to discover which programme objectives are being achieved to determine the reasons for specific success or failures to uncover principles underlying a successful, programme, to redefine the means to be used for attaining goals and to obtain continuous support, satisfaction and improvement.

Goal may be defined as the distance in any given direction one expects to go during a given period of time. Objectives are developed from an analysis of the situation range from the general to the specific. The general objectives are termed as aim. A goal is small part of the objective, it designates the distance to be traveled during a given period.

## 10.2.3 Theoretical Basis of Comprehensive Evaluation

Evaluation is not merely a part of all phases of extension teaching but concurrently an analysis of the activities leading towards the results. The following needs to be evaluated.

**(1) Evaluate programme planning** : As a result of experience, theory, research and experimentation, much information has been accumulated. Progress in science and technology have made the scientific planning more important. Some of the steps needed to evaluate programming.

(i) Identify the evidence needed to form a judgement about each criteria

(ii) Specify the method that will be used to obtain the evidence such as personal observation personal interview or through a systematic survey.

(iii) On the basis of the evidence gathered, judge whether or not each criterion is being adequately satisfied in the programme planning activities.

(2) **Evaluate Programme management :** The following list of criteria has been identified from the extension education literature and can be used to measure effectiveness in programme implementation.

(i) Appropriate groups and organisations are involved in carrying out the programme.

(ii) Workers and volunteers be given training.

(iii) The subject matter is current and appropriate to meet the programme objectives.

(iv) Materials and Methods used are varied and stimulating

(v) A variety of coordinated and integrated methods are developed to accomplish each programme objective.

(vi) Appropriate human and material resources are used.

The important steps to be followed are ;

(i) Identify the evidence you need together about criteria.

(ii) Work the method for collecting evidence

(iii) List the procedures for analysing and using the evidence collected.

(3) **Evaluate programme results :** The following steps can be used as a guide in carrying out evaluation of results:

(i) State the specific objectives to be evaluated in operational terms so they are measurable

(ii) Collect evidence from the specific groups.

(iii) Obtain valid and reliable evidence

(iv) Select appropriate method of collecting evidence.

(v) Sample selected should be adequate and representative.

(vi) Draw only logical conclusions based on collected evidences.

## 10.2.4 Evaluation Models and Approaches

### (a) Need for a model

Planning a good programme is primarily an intellectual activity since it usually involves study and use of facts, skill and principles, it requires knowledge, imagination and reasoning ability often it requires a mastery over special skills and techniques tested by empirical evidence. It is basically a process of making decisions that will be carried out in future. In sorting problems, fixing priorities, designing has been recognized world over.

All these involves several complex problems, a workable procedure needs to be formulated and understood.

**(b) Evaluation of the model**

Planning properly is an investment of time. It is a positive dynamic, useful and effective term when the concepts involved are understood and applied. A critical analysis has given rise to the model. Keeping in view the Indian conditions, following model is developed by Sandhu (1965)

***(1)*** ***Organization for planning***

***(2)*** ***Planing process :*** (i) Reach understanding regarding principles, procedures, roles and time schedule, (ii) Analyse situation, (iii) Determine objectives, (iv) Select problems with due regard to priorities, (v) Find solutions.

***(3)*** ***Planned programme :*** Prepare written statement of : (i) Situation, (ii) Objectives (iii) Problems and (iv) Solutions

***(4)*** ***Plan of work :*** Prepare a plan of work containing information regarding; (i) People to be reached, (ii) Goals dates and places, (iii) Teaching procedures to be followed, (iv) Duties, training and recognition of leaders, (v) Role of extension personnel, (vi) Role of other agencies.

***(5)*** ***Execution of plan work :*** (i) Make advance arrangement for inputs and teaching aids, (ii) Interpret the approved programme to the staff and people, (iii) Carry planned programme, phase by phase in a coordinated manner.

***(6)*** ***Evaluation of accomplishment :*** (i) Concurrent evaluation, (ii) Ex Post Facto evaluation

**(c) Models available in programme planning and Evaluation**

*(i)* *Warner* (1955) suggested following steps for programme planing.

(1) Analysis of local situation

(2) Determining objectives

(3) Plan of action.

(4) Calendar of work

(5) Evaluation of result

*(ii)* *Maunder* (1956) proposed working model of programme planning process:

(1) Collection and analysis of data.

(2) Determination of needs and objectives

(3) Defining problems

(4) Finding solution

(5) Selecting problems and determining priorities

(6) Preparing a plan of work

(7) Carrying out the plan

(8) Checking and evaluating results.

(9) Review of progress and projections of plan

*(iii) Raudabaugh* (1957)suggested five stages for planning extension programmes

(1) Identification of the problem

(2) Determination of the objectives

(3) Development of a plan of work

(4) Follow through on the plan of work established

(5) Determination of progress.

*(iv) Peason* (1966) identified eight steps in the programme planning process.

(1) Collect facts

(2) Analyse situation

(3) Identify problems

(4) Decide on objectives

(5) Develop plan of work

(6) Execute plan

(7) Determine progress

(8) Reconsideration with evaluation at each step

(v) Bereton (1972) listed five steps for developing a programme.

(1) Need determination

(2) Setting programme objectives

(3) Programme design

(4) Programme implementation

(5) Programme evaluation

### 10.2.5 Reporting the Assessment

The last phase in survey is writing report. After the collected data has been analyzed and interpreted and various generalization have been drawn. The report has to be prepared. The report of a survey is thus, the statement that contains in brief the procedure adopted and the findings arrived.

**(a) Purpose of a Report**

The purpose of a report is to convey to the interested persons the whole result of the study in sufficient detail and so arranged as to enable each comprehen the data and determine for himself the validity of conclusions. The perfection can be achieved by inviting comments from people. Research is essentially a co-operative venture and it is essential that every investigator should know what others have found about the phenomenon under study.

Report creates ground for hypothesis and leads to further research on same or allied problems. Various small segments can be coordinated and consolidated in single theory.

**(b) Report contents**

Report should contain following information

**(i)** **Purpose of study :** The report should start with the problem under taken in the investigation. It should give the background of problem need for research and hypothesis formed, scope of investigation. Importance of the problem for knowledge advancement.

**(ii)** **Methodology :** It should give various aspects of the problems. It should give exact meaning of various units of measurement. Method of collecting data. Generally questionnaires or schedules are used for the purpose of study. A copy of all such questionnaires, schedules, forms, statements be attached with the report forms statements be attached with the report to have an idea about meaning validity and reliability of various terms used in report. Besides, if some other technique of measurement is adopted viz. scaling techniques, sociological indices etc. must be mentioned in the report.

**(iii) Organisation of survey :** It gives information on method of selecting field workers, their training, procedure of interview, collecting schedules and questionnaires, method of data tabulation and classification.

**(iv) Analysis and inferences :** The data collected is analysed and various inferences are drawn based on statistical and logical reasoning. Basis for inferences all sort of proofs in support of theory advanced.

**(v) Appendices :** The report generally contains the statements or original documents on the basis of which generalizations have been formed. The questionnaire of schedule or other forms and statements are included in appendix. In short all those facts that media special elaboration, but can not be given in the main report be given in appendix.

**(vi) Foot notes :** Foot notes serve the same purpose as appendices. They are generally used as reference guide. Foot notes be given where ever necessary but their too frequent use be avoided.

**(vii) Suggestions** : Suggestions are given when investigation is not for research purpose. Suggestions are given at the end of report and must be practicable and based on logical reasoning, must be exhaustive and should take into consideration all the difficulties that may arise in their implementation.

## 10.3 Glossary

**Evaluation:** Process of systematic appraisal.

**Internal evaluation :** Evaluation carried by implementing agency.

**Concurrent evaluation:** Evaluation carried during the project period

**Ex Post Facto Evaluation:** Evaluation carried out after completion of the project

**Reporting Assessment:** Brief contents of procedure and findings

## 10.4 Summary

Evaluation may be defined as a process of systematic appraisal by which we determine the value, worth or meaning of an activity or an enterprise. It is a method for determining how far an activity has progressed and how much further it should be carried to accomplish objectives.

More specific definations of evaluation are given by persons involved in development programmes. While most of these definations refer to the assessment of the results of programmes.

Internal evaluation is carried to by agency responsible for the planning and implementation of the porgramme. Evaluation conducted by a person or a committee outside the area of operation is called as external evaluation.The primary purpose of evaluation is to ascertain how to improve effectiveness. Total programme or major phases are evaluated to determine progress towards specific objectives.Planning a good programe is primarily an intellectual activity since it usually involves study and use of facts, skill and principles, it requires knowledge, imagination and reasoning ability often it requires a mastery over special skills and techniques tested by empirical evidence. It is basically a process of making decisions that will be carried out in future.

In sorting problems, fixing priorities, designing has been recognized world over. Planning properly is an investment of time. It is a positive dynamic, useful and effective term when the concepts involved are understood and applied. A critical analysis has given rise to the model, keeping in view the Indian conditions, evaluation models are developed by Sandhu (1965).

The last phase in survey is writing report. After the collected data has been analyzed and interpreted and various generalization have been drawn. The report has to be prepared. The report of a survey is thus, the statement that contains in brief the procedure adopted and the findings arrived.

## 10.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. Define Evaluation and explain various categories of evaluation.
2. Define Evaluation given by Mathew and explain the objectives of evaluation.
3. State and explain purpose and Goals of evaluation.
4. Describe various steps involved in programme planning, management and evaluation of results.

# Unit 11 : Processing and Analysis of Data

***Index***



## 11.1 Introduction

After studying how to collect the data and information; it is essential to know how to process the data. The processing of data is the first step prior to analysis of the data to draw meaningful conclusions.

The data collected by the investigator is known as raw data; which is voluminous, huge, unwieldy, and incomprehensible. Hence raw data required editing, the edited data is then organized properly. The organization of the data means arranging the data in a proper format in a condensed form. The condensed data, which highlights the important characteristics of the data facilitates comparisons and render it for suitable statistical analysis and interpretation.

**After the study of this unit, you will be able to know and understand :**

- The classification of data
- The variables of data
- Frequency distribution
- Tabulation of data for statistical analysis

## 11.2 Content

### 11.2.1 Processing, Operations and Data Classification

The raw data is the data presented in the form in which it is collected prior making any arrangement. A statistical table is an orderly and logical arrangement of the data into rows and columns. It is the process to present the voluminous heterogeneous data in a condensed and homogeneous form.

**1. Classification**

The process of arranging data into groups or classes according to similarities and resemblances is known as classification. It is the first step of tabulation of data. The classification is the arrangement of the data depending upon the nature, objectives, and scope of the investigation. An illustration of the classification of data can be given for family budget information obtained from a socio-economic survey. The items of expenditure can be classified according to the heads such as food, clothing, fuel, lighting, house rent and miscellaneous charges

Each of the above group can be subdivided into subgroups or sub-classes e.g. food may be sub divided into cereals, pulses, vegetables, milks and milk products. Thus the criteria of sub classification can vary according to the enquiry desired and criteria decided.

**2. Functions of Classification**

The functions of classification are as given below :

- ***(i) It condenses the data :*** The classification condenses the huge unwieldy raw data which is readily comprehensible to draw significant features and highlight the enquiry.
- ***(ii) It facilitates comparison :*** The data classification facilitates meaningful comparison depending upon the criteria and basic for comparison
- ***(iii) The study of relationship is possible :*** The classification can be made with respect to two or more than two characteristics; which enables us to find out the relationship between two or more than two variables.

(iv) ***The statistical treatment of the data is possible :*** The heterogeneous data is arranged into homogeneous classes according to points of similarities. Hence this uniformity makes the data amenable for further processing and statistical treatment.

## 3. Basis of Classification of Data

The criteria of classification depends upon the purposes of enquiry and objectives. Generally the data can be classified according to the following points.

(i) ***Geographical classification :*** When the data is classified according to the criteria of geography, region or area, it is known as geographical classification. e.g. the information of land can be classified according to the criteria of soil type such as shallow, medium, black soil etc.

(ii) ***Chronological classification :*** When the data is a function of time. The data can be arranged according to the different time factors. When this arrangements is done in ascending or descending order of time it is called as chronological classification. e.g. Production of sorghum in Maharashtra state for 10 years. The change in population of bacteria in food over a period of time etc.

(iii) ***Qualitative classification :*** When the data is measured in Qualitative terms, it is required to label with certain symbols to distinguish one type of item from another. This is known as qualitative type of classification. e.g. the population can be classified into male and female. Then male and female can be classified as workers and non-workers and further into literate and illiterates. These may further be classified according to religion or caste etc.

(iv) ***Quantitative classification :*** There are certain measurements which are capable of assigning numerical values such as age, income, production, yield etc. when the quantitative data is classified according to a variable and its frequency it is known as quantitative classification. Quantitative classification is of two types such as (i) Inclusive type classification and (ii) Exclusive type of classification.

### 11.2.2 Variables and Attributes

**Population :** Total collection of individuals on which certain characteristics are observed. Numerical data raised when certain characterizes are observed for the population. The characteristics are of two types. (1) Qualitative characteristics and (2) Quantitative Characteristics. The qualitative characteristics or attributes represent certain qualities of individuals, such as colour, regions etc.

**Variable :** Variable is quantitative in character, which is assigned with different values in quantities, is known as variable or variate. Generally, variate is a fixed value which is assigned to the particular individual. Thus height, weight, yield, production, area are the examples of variable. The height is a variable. A particular person has 150 cm. height i.e. a variable takes value 150 cm.

**1. Types of Variables**

There are two types of variables such as :

***(i)*** ***Continuous Variable :*** The variable which can assume any value in a given range is known as continuous variable. Continuous variable is a function of rational numbers. It can take any value, integrals or fractional in the defined range. e.g. the height is an example of continuous variable, similarly age, weight, yield, area are the examples of continuous variable. There is no definite gap between two values of continuous variable.

***(ii)*** ***Discrete Variable :*** When the variable assumes only isolated, particularly integral values is known as discrete variables or discontinuous variable. There is a definite gap or jump between any two values of the discrete variable. Discrete variable is a function of counting numbers of natural numbers. The number of flowers of roses to a plant is an example of discrete variable. Similarly number of pods per plant, number of branches per plant, number of students in the class, are the examples of discrete variables.

**2. The Data**

The organization of the quantitative data can be done in three ways.

(i) The series of individual observations or unorganized raw data, (ii) Discrete grouped data (iii) Continuous grouped data

## 11.2.3 Frequency Distribution

The organization of quantitative data according to values and their frequencies in a systematic order is known as frequency distribution. The frequency value is the number of times of occurrence of an individual value in the data. The organization of data involves four stages.

### 1. Array

The raw data can be of items. The arrangement of the data in ascending or descending order of magnitude is known as array. The array allows the raw data to be written in a systematic order but does not reduce the volume of data. e.g. 10, 15, 11, 12, 2, 1 is a raw data and its arrangement is ascending order such as 1, 2, 10, 11, 12, 15 or descending order as 15, 12, 11, 10, 2, 1 is called as array

### 2. Discrete Frequency Distribution

In this frequency distribution we count the repetition of an individual value and its occurrence in the series of observations (frequency) e.g. 11, 9, 13, 8, 6, 12, 9, 8, 11, 11, 9, 8, 12, 13, 9, 8, 17, 19, 8, 25 are the marks in a class test out of 30. Arrange the data in an array and ungrouped frequency distribution. In this series there are 20 values where the largest value is 25 and the smallest is 6. If we write these values in ascending or descending order. e.g. 6, 8, 8, 8, 8, 8, 9, 9, 9, 9, 11, 11, 11, 12, 12, 13, 13, 19, 19, 25. It is called as an array. If we write individual values and respective frequencies as tally marks for occurrence the values can be written as given below.

<table>
<tr><th>Item</th><th>Tally marks for occurrence</th><th>Frequency</th></tr>
<tr><td>6</td><td>|</td><td>1</td></tr>
<tr><td>8</td><td>卌</td><td>5</td></tr>
<tr><td>9</td><td>||||</td><td>4</td></tr>
<tr><td>11</td><td>|||</td><td>3</td></tr>
<tr><td>12</td><td>||</td><td>2</td></tr>
<tr><td>13</td><td>||</td><td>2</td></tr>
<tr><td>19</td><td>||</td><td>2</td></tr>
<tr><td>25</td><td>|</td><td>1</td></tr>
<tr><td>Total</td><td>20</td><td>20</td></tr>
</table>

In this manner the data of 20 items is reduced and presented as a meaningful.

### 3. Grouped Frequency Distribution

The data classified as above still can be grouped into different small groups or classes. The various groups into which the values of the variable are classified are known as classes. The class or class interval is a group or a small range within which the values of the variable occur.

The smallest value of a group is known as lower limit of the class thus lower limit is starting point of the class. The highest value of a class is known as upper limit of the class. The distance between lower limit and upper limit is width of the class internal and it is difference between upper limit of a class and lower limit of the class.

When the variables are discrete in nature, then it is convenient to classify the values according to inclusive classification. the inclusive type of classification is a classification in which both the lower limit and upper limit of the class are included in the same class or interval. e.g. in the above case the marks are distributed from 6 to 25 then the total range will be 25-6 =19 . if we wanted to have four classes then the frequency distribution is presented in the given table.

| Class interval | Frequency |
|---|---|
| 6-10 | 𝍸 𝍸 = 10 |
| 11-15 | 𝍸 𝍷𝍷 = 07 |
| 16-20 | 𝍷𝍷 = 02 |
| 21-25 | 𝍷 = 01 |
| **Total** | **20** |

In a discrete frequency distribution only whole numbers are the observations for frequency distribution for examples if a number in between 10 and 11 say 10.5 occurs then this type of classification is not useful.

### 4. Continuous Frequency Distribution

When the variable under consideration is a continuous variable i.e. when the items are in integral number (whole number) and or is fractions number (rational numbers) then the discrete classification is not useful because there is definite gap between two class intervals. To avoid this type of situation we can have exclusive type of classification for continuous variable. The exclusive type of classification is a type of classification in which only lower limit of the class is included in the current class interval and upper limit of the class is excluded from that class interval and it is included in the following class interval as lower limit of next class. Above example is classified according to exclusive classification and have the classes as given below.

<table>
<tr><th>Class interval</th><th>Frequency</th><th>Remarks</th></tr>
<tr><td>6-11</td><td>$\cancel{||||}$ $\cancel{||||}$ = 10</td><td>11 is not included in this class</td></tr>
<tr><td>11-16</td><td>$\cancel{||||}$ || = 07</td><td></td></tr>
<tr><td>16-21</td><td>|| = 02</td><td></td></tr>
<tr><td>21-26</td><td>| = 01</td><td></td></tr>
<tr><td>Total</td><td>20</td><td></td></tr>
</table>

Thus this type of classification accommodates all the fractional or integral in the range of 6 to 26.

## Basic Principles of Grouped Frequency Distribution

A statistician can use his discretion while classifying the data according to his wisdom, skill and sound experience in data management. However some guideline are always helpful for good and sound classification given as below.

**(i)** **Types of Class :** The classes should be well defined and should not be ambiguous. They should be mutually exclusives (disjoint) so that a particular value should correspond to one and only class interval. There should be one to one correspondence for the value of the variable and class interval.

**(ii)** **Number of Classes :** There is no hard and fast rule to decide the number of classes. However the same depends on total frequency, nature of data, accuracy and easiness for computation.

The practical consideration to decide the number of classes which should not be too small nor too large. If the size of the class is too large, so many observation will be there for a class. This will obscure the important information while estimating mean and variance because of crowding observation around a single value, similarly the accuracy will decrease because of small number of classes. Ordinarily the number of classes should not be more than 20 and should not be less than 5. A thumb rule proposed by prof. Sturge known as **Sturge's rule for deciding the number of classes,** which is as follows.

$K= 1+3.322 \text{ Log } N$

where K= number of classes and N= total frequency

From this formula, if N= 10, the value of K= 4 approximately, for N = 100 then K= 8, for N=500 then K= 10, for N=1000 then K= 11 and for N = 10,000 then K= 14

**(iii) Size of the Class :** The size of the class interval is also subjective and depends upon the statistician according to the total frequency, computational ease and accuracy. If number of classes are determined by the formula of Sturge; then the size of a class determined by the number of class intervals.

$$\therefore \text{ Magnitude of class } = \frac{\text{range}}{\text{number of class}}$$

The thumb rule is that the size of class interval should not be greater then $\frac{1}{4}^{th}$ of the estimated population standard deviation

$$I \leq \frac{\sigma}{4} \qquad (\sigma = \text{Sigma})$$

Where I = size of class and $\sigma$ population, standard deviation.

Generally, size of class interval is taken as 5 or some multiple of 5 viz. 10, 15, 20 etc. for facilitating computation. Class intervals are fixed in such way that the mid point is convenient for calculation, assuming that the class frequency is concentrated at mid point of the class. The size of class intervals should be same for all class intervals through out the frequency distribution, so that the comparison will be meaningful.

**(iv) Types of Class Intervals :** There are two types of class intervals, when the data under consideration is in whole numbers or belongs to discrete variable, one can use inclusive type of class interval when the data is of whole number and fractions or the variable under consideration is continuous variable, we may use exclusive type of classification to take care of inclusion of all values in the given series. The average of upper and lower limit of class is known as mid value of the class interval.

$$\text{Midpoint} = \frac{U + L}{2}$$

Where U = upper limit and L= lower limit

The class limits are chosen is such a manner that the observations in any class are evenly distributed throughout the class interval so that the actual average of any class is close to mid value of class interval. In other words the observations should be concentrated at the mid points of the classes.

**(a) Open end class :** The classification is termed as open end classification, if the lower limit of the first class or the upper limit of the last class or both are not specified and such classes in which one of the limits is missing are called open end classes. Thus in classification the total frequency i.e. total observations are distributed over different classes, hence the name frequency distribution.

**(b) Cummulative Frequency Distribution :** There are following two types of cummulative frequencies.

***(i)*** ***Ascending Order Cummulative Frequency :*** The ascending order cummulative frequency at a particular value of the variable is the total number of observations which are less than or equal to that particular value. As value of the variable increases, cummulative frequency also increases; as additional observations are included in the cummulative frequency.

***(2)*** ***Descending Order Cummulative Frequency :*** The descending order cummulative frequency at a particular value of the variable is the total number of observations which are greater than or equal to the particular value. As value of the variable increases, in this case, cummulative frequency decreases as more number of observations are excluded.

The cummulative frequency distribution is a successive total of frequencies over previous classes. When the class interval are arranged in ascending order and frequency total of each previous class in added to the next class such that the cummulative frequency of the last class is total frequency over all classes. Then this type of cummulative frequency distribution is known as ascending cummulative frequency distribution or less than cummulative distribution.

The more than cummulative frequency is the total of successive frequencies from the highest class interval to the lowest class interval. Thus the successive total of frequencies are obtained from the last class to the first class. e.g. Following is the frequency distribution of marks in mid term examination and the number of students who secured the marks in particular range. Then obtain cummulative frequency less than and cummulative frequency more than.

| Class interval (marks) | 0-5 | 15-20 | 5-10 | 10-15 |
|---|---|---|---|---|
| Number of students ( f ) | 5 | 12 | 15 | 18 |

**Solution :** Arrange the classes in systematic order and write down the frequency of each class.

| Class interval | Frequency ( f ) | Cummulative frequency (less than) | Cummulative frequency (greater than) |
|---|---|---|---|
| 1 | 2 | 3 | 4 |
| 0-5 | 5 | 5 | 45+5=50 |
| 5-10 | 15 | 5+15=20 | 30+15=45 |
| 10-15 | 18 | 20+18=38 | 12+18=30 |
| 15-20 | 12 | 38+12=50=N | 12 |

From the above table we can answer the questions as given below.

1. How many students are having marks less than 15 ?
2. How many students are there who secured more than 10 marks ?

These two are similar type of questions can be answered by referring third column for the first question and column No. 4 for the second question .

### 11.2.4 Tabulation of Statistical Data

To draw meaningful conclusions from the raw data it is essential to tabulate the data. The tabulation is the systematic presentation of the information in rows and columns in accordance with some characteristics. The horizontal arrangement of the data is in rows and the vertical arrangement of data is in column.

Tabulation is the process of accumulating the data to find out reasoning and results from the gathered statistical information. Tabulation is the most important and convenient process of presenting the data in a condensed and readily understandable manner to furnish maximum information in the minimum possible space.

It is a intermediate process between collection of data and statistical analysis and interpretation of the same. By tabulating the data, we can compare the two or more than two groups. The diagrammatic and graphic representation of the data is possible. Further statistical treatments such as finding out relation between two variables is also possible. Accordingly a good table can only be obtained through the skill, expertise, experience and common sense of the tabulator in respect of nature, scope and objective of the statistical data.

**Parts of Table**

The various parts of a table vary from problem to problem depending upon the nature of the data and purpose of study. However following seven parts are desirable in a good table.

**(1)** **Table Number :** The table number may be placed on the top of the table, the table number is an essential part when there are more than one tables in a book, article or report. The table number helps to identify the desired information in easy and quick manner. By assigning table numbers, we can give a logical sequence to the matter to be presented in the report.

**(2)** **Title of the Table :** The title of the table can be given next to the table number, the title should be brief and it should describe the contents of the table in a concise manner. The title must be able to describe the nature of the data presented in the table. It should be unambiguous, properly worded and punctuated. The title should be prominently lettered. Thus the title is very essential without which the contents, nature and scope is not clear.

**(3) Head Notes :** The head note is given just below the title in a prominent type enclosed in brackets, to describe the contents of the table. The units of measurements are usually expressed such as "in hectares" "in millions" "in kg" "in rupees" or "in million tonnes" etc.

**(4) Captions and Stubs :** Captions are the headings or designations for the vertical columns. The stubs are headings or designations for the horizontal rows. The captions and stubs should be very brief, concise and self explanatory. If the units are different for different columns then the units should be given in the bracket below the caption. Similarly if each row has different measurements then the units of measurement such as ratios percentages etc. in case the same units are repeated for all columns then there is no necessity to mention the unit for separate columns.

**(5) Body of the Table :** Generally the body of the table contains the numerical data which is important part of the table for description, undesirable and irrelevant information should be avoided. The total at the end of rows and columns increase the usefulness of the table.

**(6) Foot Notes :** When the characteristic of the items in the body of the table is not fully explained to the reader or it is felt that the tabulated information requires additional explanation then the foot note is used. The foot note is given below the bottom of the table. Suitable caption is given to the foot note. Foot notes are identified with some captions and symbols such as * , **, @.

**(7) Source Note :** The source of the tabular information is important to know the reliability of the data or information. The source note is given below the foot note. The source note is written, when the table presented depends upon already published or processed data. Generally the volume number, table number, page number of a book, journal or research article is given in the source note. So that the reliability can be verified by referring the original document in case of further verification.

## 11.3 Glossary

**Array** : It is an arrangement of data in ascending or descending order

**Classification** : The process of arranging data into groups or classes

**Table:** A statistical table is an orderly and logical arrangement of data

**Variable** : A quantity or quality which varies from one individual to another

**Discrete Variable :** the variable which assumes the values in whole numbers.

**Frequency** : Repetition of item in the series

**Caption** : The heading of vertical coloumn

**Stub** : The heading of horizontal rows

## 11.4 Summary

The data collected by the investigator is in raw form, huge, unclassified and heterogeneous. It is difficult to draw the conclusions and inferences. It is therefore necessary to make classification of data as per the categories or groups. The total population or data are classified as per the characteristics which may be qualitative or quantitative. The number of data appearing of same value in a group is a frequency distribution. The basic principles of group frequency are types of classes, number of classes, intervals, etc. In order to draw meaningful conclusions from raw data, it is essential to tabulate in desirable form.

## 11.5 Exercises for Practice

**Answer the following questions in about 200 words each.**

1. What is classification of data ? State and explain basis for clasificatin of data.
2. What is tabulation of data ? Why it is essential in statistical analysis ?
3. Define frequency and frequency distribution, explain different frequency distributions used for discrete and continuous variables.
4. What is variable ? Explain types of variables in statistical analysis.

# Unit 12: Testing of Hypothesis

***Index***



## 12.1 Introduction

In behavioral science or biological sciences scientists conduct research and give different methods of extension work; different crop varieties, different methods of sowings, applications of fertilizers etc. All these methods and new things evolved require comparison with each other or with the established one or earlier entities. From the theories of behavioral science or from, the knowledge of certain things, researchers make certain assumptions. These assumptions lead to different types of research. The research hypothesis is the prediction derived from the theory under test. Hence prior to undertaking of a research, whatever rationale we make is known as hypotheses.

**After the study of this unit, you will be able to know and understand :**

- Hypothesis and types of hypothesis
- Testing of hypothesis
- Limitations of testing hypothesis

## 12.2 Content

### 12.2.1 Hypothesis and its Definition

A population is completely known, if the probability distribution of the population is completely specified. The probability distribution is completely specified, if values of the parameters of probability distribution are specified. The population parameters are generally unknown. The hypothesis is a statement regarding the population parameter, and trueness of this hypothesis is tooled in light of the sample available.

The hypothesis is a statement regarding the population parameter. Population parameter is a population characteristics or population constant generally unknown. The hypothesis is a statement regarding the population distribution of population constant. e.g.

(1) The mean yields of two sorghum varieties CSH-1 and CSH-2 are equal

(2) The statement that the average annual rainfall of Mumbai and Nasik is not equal.

(3) The cropping pattern of location A is similar to cropping pattern of location B.

(4) The adoption level of small farmers is different from large farmers.

The above statements can also be written briefly in the mathematical form

1. $\overline{Y}_1 = \overline{Y}_2$

   Where $\overline{Y}_1$ = mean yield of sorghum variety CSH-1

   $\overline{Y}_2$ = mean yield of sorghum variety CSH-2

2. $\overline{R}_1 \neq \overline{R}\ 2$

   Where $\overline{R}_1$ = Average rainfall of Mumbai

   $\overline{R}_2$ = Average rainfall of Nashik

3. CPA = CPB

   Where CP = Cropping Pattern and A and B are locations

4. ASF ≠ ALF

   Where

   ASF = Adoption level of Small Farmer

   ALF = Adoption level of Large Farmer

## 12.2.2 Types of Hypothesis

There are two types of hypotheses (1) Null Hypothesis and (2) Alternative Hypothesis

**1. Null Hypothesis ($H_0$) :**

Null Hypothesis is the Hypothesis which is set for possible rejection under the assumption that it is true. Thus null Hypothesis is the Hypothesis, which is set for possible equality or possible rejection. It is denoted by ($H_0$). The Null Hypothesis is the hypothesis of no difference. It is usually formulated for the purpose of being rejected. e.g.

(i) $H_0$: The median scores of two teaching methods are equal
$H_0$: $Me_1 = Me_2$
$Me_1$ = median of method one
$Me_2$ = median of method two

(ii) $H_0$ : the variables age and income are independent
$H_0$ : r ( A,I )
r = Correlation coefficient, A= Age and I = Income

**2. Alternative Hypothesis ($H_1$)**

The alternative hypothesis is the hypothesis which is a set for possible acceptance or possible inequality. The alternative hypothesis is denoted by $H_1$. The alternative hypothesis is the operational statements of the experimenter's research hypothesis. The alternative hypothesis is complimentary statement to the null hypothesis. It is very important to state the alternative hypothesis explicitly, because rejection of null hypothesis ($H_0$) imply acceptance of alternative hypothesis ($H_1$) and vice-versa. e.g.

The null hypothesis stated as above is

$H_0$ : $\overline{Me_1} = \overline{Me_2}$ has alternative hypothesis $H_1$ : $\overline{Me_1} \neq \overline{Me_2}$ ........ (A)

Thus the alternative hypothesis is the median of two methods are not equal and correlation coefficient between two variables is not zero.

Similarly other alternative statements can be

$H_1 : \overline{Me_1} > \overline{Me_2}$ ....... (B)

$H_1 : \overline{Me_1} < \overline{Me_2}$ ........ (C)

Thus in case of (A) it is not stated whether $Me_1 > Me_2$ or $Me_1 < Me_2$. It keeps the possibility open for the (B) and (C). Hence it is known as two tailed alternative hypothesis. The alternative hypothesis stated in (B) tests the possibility of "greater than" symbol between the two, hence it is known as Right tailed alternative hypothesis and in (C) it is given "less than" symbol, hence it is known as left tailed alternative hypothesis. Accordingly the statistical tests used are also called as two tailed statistical test, right tailed statistical test and left tailed statistical test.

## 12.2.3 Procedure for Hypothesis Testing

In order to reach an objective decision as to whether a particular hypothesis is confirmed or accepted on the basis of a data; it is required to have an objective procedure. The objective procedure should be based on the information, we obtain in our research and on the risk, we are willing to take, that our decision with respect to the hypothesis may be incorrect. This procedure involves several steps. Following steps are important in order of their performance.

**(1) Null Hypothesis ($H_0$) :** State the null hypothesis ($H_0$) in the mathematical form as well as in the statement form. The details of the formation of the null hypothesis ($H_0$) are already described earlier.

**(2) Alternative Hypothesis ($H_1$) :** This will enable researcher to decide whether to select single (right or left) tailed test or two tailed test.

**(3) The Choice of Statistical Test :** The statistical tests are available for almost all types of research designs and similarly alternative statistical tests are also available to come to a decision about a hypothesis. We must have a rational basis for choosing a particular statistical test from the several alternatives available. The power of a statistical analysis is partly a function of the statistical test used in the analysis. A statistical test is good one, if it has a small probability of rejecting null hypothesis ($H_0$) when $H_0$ is true; but a large probability of rejecting it when $H_0$ is false. If there are two statistical tests A and B which have the same probability of rejecting null hypothesis when it is true. Then we select the one that has the larger probability of rejecting $H_0$ when it is false. There are parametric and non-parameter statistical tests. Both the tests are equally powerful in detecting the differences, but the applicability of them depend upon the manner in which the samples are drawn, the nature of the population from which the samples are taken, the operational definitions of the variable and the type of measurement scale of the observation. Thus in brief the statistical test to be selected depends upon the hypothesis set, nature of the population distribution, sample size and measure scale of the scores. Whenever, the population distribution is normal and measurement scale is not weak; one can use parametric statistical tests and when the measurement level is weak and the assumptions are fewer or weaker in nature regarding the population distribution, it is advised to use non parametric statistical tests.

**(4) Statistical Errors :** There are two types of errors which may be made in arriving at a decision about null hypothesis.

***Type I Error :*** The first kind of error or type I error is made when we reject the null hypothesis ($H_0$) when in fact it is true. The probability of type I error is denoted by $\propto$ (alpha). The level of significance is the probability of rejecting null hypothesis ($H_0$) when if is ought to be accepted. If larger is $\propto$ then it is more likely that $H_0$ is falsely rejected.

***Type II error :*** The second kind of error or type II error occurs when $H_0$ is accepted, when infact it is false. The probability of committing type II error is denoted by $\beta$ (beta). The probability of making type I error is $\propto$ and probability of making type II error is $\beta$

P (Type I error)= $\propto$ and P (Type II error)= $\beta$

The values of $\propto$ and $\beta$ should be specified by the experimenter before he begins his research work. The values of $\propto$ and $\beta$ determine the required sample size (n) to meet the statistical test.

A decrease in type I error will increase type II error for any given sample size (n). If we wish to reduce the possibility of both types of errors, we must increase size of n. Hence whenever we are making a statistical inference, it is evident that it is not free from the two types of errors. Therefore, it is necessary to optimize or control them. This gives rise to the concept of power of the statistical test and confidence coefficient. Therefore, the probability of type I error is kept at a fixed level and a test procedure with minimises probability of types II error at affixed level of type I error is desired.

***The power of statistical test :*** The power of statistical test is the probability of rejecting null hypothesis $H_0$ when infact it is false

Power of the test : = 1- probability of type II error = $1 - \beta$

The power of the test can be increased by increasing sample size (n), because by increasing the sample size (n) the type II error decreases. Similarly the confidence coefficient can also be defined as $1- \propto$

**Confidence Coefficient :** The confidence coefficient is defined as the probability of accepting $H_0$ when infact it is true.

Confidence coefficient = 1 – probability of type I error = $1 - \propto$

The decisions of accepting and rejecting null Hypothesis (Ho) when it is true or false can be very easily described with the help of tabular form.

| True or false/Accept or reject | Decision from sample | |
|---|---|---|
| | Reject $H_0$ | Accept $H_0$ |
| $H_0$ True | Wrong type I error | Correct |
| $H_0$ False | Correct | Wrong type II error |

**(5) The Level of Significance and Sample Size :** After stating null hypothesis ($H_0$), alternative hypothesis ($H_1$) accurately, and an appropriate choice of a statistical test, the next step is to specify a level of significance ($\propto$) and to select a sample size (n). This is our decision-making procedure which specifies the set of all possible samples that could be drawn when $H_0$ is true. Thus we define the level of significance as the probability of rejecting $H_0$ when infact it is true. This probability of rejecting null hypothesis $H_0$ when it is true is kept at fixed level say 5% or 1% and it is called as level of significance. Common values of $\propto$

selected in biological and agricultural research are very small such as 0.05 and 0.01. The interpretation of level of significance $\propto = 0.05$ will be such that if we select 100 number of samples only five or less than five will have chances of showing that the null hypothesis is true but it will be rejected. Thus level of significance $\propto$ gives the probability of falsely rejecting null hypothesis ($H_0$).

**(6) The Sampling Distribution :** The sampling distribution is the theoretical distribution. It is that distribution we get if we take all possible samples of size (n) from the same population of size (N) with random selection of each sample.

The same thing can be stated in the manner that the sampling distribution is the distribution under null hypothesis $H_0$ of all possible values that some statistic can take when that statistics is computed. From the random samples of equal size. The statistic is function of observations from the sample values. It is a sample characteristics. Sample mean, sample median, sample variance are statistic. The same thing can be illustrated with an example; let the population size be (N) and sample size be (n). Then if we make random selection of a sample of size (n); we shall have $(\frac{N}{n})$ possible number of samples. Let the statistic we desire to estimate from the same be arithmetic mean ($\bar{x}$). Then $(\frac{N}{n})$ samples will have $(\frac{N}{n})$ values of sample mean ($\bar{x}$). Then if we group all these values in a frequency distribution table, we shall get a frequency distribution of these means. This is the sampling of distribution of sample means. Therefore sampling distribution is the probability distribution of the sample parameters.

**(7) Critical Region :** The critical region is also called as region of rejection. The region of rejection belongs to the sampling distribution; so it is a rejection region of sampling distribution. The sampling distribution contains all the possible values of a test statistic lender Null hypothesis. The region of rejection is the subset of these values. This is defined as the probability of the occurrence of a test statistic having a value which is in that subset under null hypothesis is $\propto$.

The region of rejection consists of a set of possible values, which are so extreme that when $H_0$ is true, the probability is very small that the sample we actually observe will give a value, which is among them. The probability associated with any value in the region of rejection is equal to or less then $\propto$.

The location of the region of rejections is affected by alternative hypothesis $H_1$ is one tailed the region of rejection is at one end and if the alternative hypothesis is two tailed then the region of rejection is at two ends of the sampling distribution. If the level of significance $\propto = 0.05$, then the region of rejection is 5 per cent of the entire space of the sampling distribution. This five-percent will be cover the extreme portion of the entire space.

**(8) The Decision :** If the test statistic gives a value, which is in the region of rejection, we reject $H_0$. When the probability associated with an observed value of a statistical test is equal to or less then the value of $\propto$; we conclude that Ho is false. Such an observed value is called as "Significant". $H_0$ is rejected wherever the test statistic yields a "significant" result. A significant value is one whose associated probability of occurrence under $H_0$ is equal to or less then $\propto$.

**(9) Degrees of Freedom :** Degrees of freedom denotes the extent of independence or freedom obtained by the number of observations in the given set. If there are 'n' number of observations and there are 'k' numbers of restrictions or constraints imposed on them the degrees of freedom is given by n-k.

Thus we subtract one from 'n' number of observations, if we estimate one parameter from the sample. If we are estimating two from the sample. If we are estimating two parameters from the sample, we subtract two from the size of sample 'n', assuming that there are only (n-2) observations, which are independent and can assume any value in the sample.

## 12.2.4 Types of Tests and Limitations of Hypothesis

The sampling distribution of statistic makes us enable to find out the probability that a statistic would differ from a given hypothetical value of the parameter by more than a certain amount.

Thus the difference between the sample value and assumed value with a given probability level is value of significance and gives rise to tests of significance. Thus the procedure to know the significance of difference between sample value and population value is known as test of significance. We say that the differences to be significant between the sample value and population value, when the differences are real differences and can not be attributed to the fluctuations of sampling.

Most of the distributions tend towards normal distribution (Bell shaped curve) when the sample size (n) becomes very large. If the sample size is larger, then it can be assumed that the distribution of parent population to be normal. Based on the Normality, we can have a statistical test; which is known as Z test, to test the difference between the sample value and assumed value of the population.

**1. Z Test**

If x is a random variable normally distributed with mean $\mu$ and variance $\sigma^2$, then the variable Z = $\frac{\bar{x}-\mu}{\sigma}$ is normally distributed with zero and variance equal to unity. The criteria for the test is to calculate the statistic Z

$$Z = \frac{|\bar{x}-\mu|}{\sigma}$$

If the calculated value of Z is greater than 1.96, then the null hypothesis $H_0$ is $\bar{x} = \mu$ can be rejected, otherwise it can be accepted at 5 per cent level of significance.

Similarly if the calculated value of Z is larger than 2.58 the null hypothesis Ho can be rejected at 1 per cent level of significance. Similarly if Z value is greater than 1.645, then null hypothesis $H_0$ is $\bar{x} = \mu$ can be rejected at 10 per cent level of significance or at 0.10 probability level. Thus 1 per cent, 5 per cent and 10 per cent are the probability values of Z

the P $(Z > Z\alpha) = \alpha$

Where $\alpha = 0.05, 0.01$

The value of $Z\alpha = 1.96$ for $\alpha = 0.05$

The value of $Z\alpha = 2.58$ for $\alpha = 0.10$

The procedure to use Z test is

(1) Set up null hypothesis $H_0$: $\bar{x} = \mu$.

(2) Alternative Hypothesis $H_1$: $\bar{x} \neq \mu$.

(3) Choose level of significance $\alpha$ 0.01 or 0.05 or 0.10

(4) Use test criteria $Z = \frac{|\bar{x} - \mu|}{SE(\bar{x})}$

Where X = any observation $X \sim N(\mu, \sigma^2)$

$\mu$ = mean of population of Hypothesis

SE ($\bar{X}$) = standard Error of $\bar{X}$

(5) We compute the value of Z as given in step number (4). Compare the value obtained in step number (4) with the theoretical value of $Z\alpha$.

If $Z > Z\alpha$ then the observed differences can be regarded as real differences at the level of significance $\alpha$. Otherwise the differences can be attributed to the fluctuations of sampling.

The Z test can be used for large samples ($n > 30$). The applications of Z test can be used for comparing two means, two attributes.

**Application**

**(i)** Z test for comparing sample mean with assumed mean or population mean i.e. $H_0 : \bar{X} = \mu$

the criteria is $Z = \frac{|\bar{X} - \mu|}{SE(\bar{x})}$

$\bar{X}$ = sample mean = $\Sigma X / n$

$$SE(\bar{x}) = \sqrt{\frac{\Sigma (X - \bar{X})^2}{n(n-1)}}$$

If the calculated value of Z is greater than table value of Z for the given level of significance (say 5 percent or 1 percent) then. The null hypothesis $H_0 : \bar{X} = \mu$ is rejected.

**(ii)** To compare two independent means $H_0 : \bar{X}_1 = \bar{X}_2$

if $X_1$ and $X_2$ are two samples of same size $n_1$ and $n_2$ of sample size in$_1$ and $n_2$ having their means as $\bar{X}_1$ and $\bar{X}_2$ and their samples variances as

$$S^2_1 = \frac{\Sigma (X_1 - \overline{X}_1)^2}{n_1 - 1} \qquad \text{and } S^2_2 = \frac{\Sigma (X_2 - \overline{X}_2)^2}{n_2 - 1}$$

then the criteria is $Z = \dfrac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\left(\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}\right)}}$

If Z> 1.96 reject $H_0$ at 5 % level of significance

If calculated value Z > 2.58 then Ho can be rejected at 1 % level of significance

**(iii)** To test significance of proportion.

$H_0$ : p=P : p = Sample proportion

Where p= E (p) = expected value of p

Z= (p – E (p) ) / SE (p) $\sim$ N (0,1)

Z = p-E(p) / SE (p) $\sim$ N (0,1)

**(iv)** For two samples we have the null hypothesis $H_0 : P_1 = P_2$ and test statistic Z as given below

If we wanted to compare to proportions of to independent samples then

$H_0 : P_1 = P_2$

$Z = | P_1 - P_2| / SE (P_1 - P_2)$

$$Z = \frac{| P_1 - P_2|}{\sqrt{\dfrac{P_1 (1-P_1)}{n1} + \dfrac{P_2 (1-P_2)}{n2}}}$$ where

$P_1$ = is the proportion of first sample with n1 as total number of observations in sample number one.

$P_2$ = is the proportion of second sample with $n_2$ as total number of observations in sample number two. $P_1 = m_1 / n_1$ and $P_2 = m_2 / n_2$ , where $m_1$ and $m_2$ are successful cases in sample one and sample two respectively, then Z $\sim$ N(0,1)

Z is normally distributed with zero mean and unit variance. Thus the value of Z can be refereed from the table and it is compared with the observed (calculated) value of Z. If the observed Z is greater than table Z, null Hypothesis Ho is rejected otherwise it is accepted.

**(v)** If the researcher has two independent samples coming from two different populations which are normally distributed then two sample means can be compared as per the procedure given below. Let $X_1$ and $X_2$ are two samples coming from two independently and normally distributed populations then the null hypothesis $H_0 : \overline{X}_1 = \overline{X}_2$ can be tested

$$Z = \frac{|\overline{X}_1 - \overline{X}_2|}{SE(\overline{X}_1 - \overline{X}_2)} = \frac{|\overline{X}_1 - \overline{X}_2|}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

where $\overline{X}_1$ = mean of $X_1 = \Sigma X_1 / n_1$

$\overline{X}_2$ = mean of $X_2 = \Sigma X_2 / n_2$

$S_1^2$ = sample mean square for $X_1 = \Sigma(X_1 - \overline{X}_1)^2 / (n_1 - 1)$

$S_2^2$ = sample mean square for $X_1 = \Sigma(X_2 - \overline{X}_2)^2 / (n_2 - 1)$

$n_1$ and $n_2$ are the sample sizes for $X_1$ and $X_2$ coming from large samples.

The value of Z can be obtained as per the procedure given above. The calculated value of Z can be compared with the table value of Z for given level of significance Null Hypothesis $H_0$ is accepted or rejected accordingly as per the procedure we have discussed earlier.

**(2) 't' Test**

If the sample size n is large, it can be assumed from the large sample theory that the sample has property of normality, that is the sample distribution of any normal variate can be converted into standard normal variate $\sim N(0,1)$,

when $n \to \infty$ (asymptotically),

$$\text{i.e } Z = \frac{|\overline{X} - E(X)|}{SE(\overline{X})} = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \qquad \text{as } n \to \infty$$

But if the sample size n is small, then all the samples do not satisfy the conditions of normality. Hence the formula of Z test can not be applied. It is very difficult to distinguish between small samples and large samples. For practical purposes it can be assumed that the sample is small, if it is less than or equal to 30 and the samples can be termed as large samples when the sample size n > 30.

W. S. Gosset has developed a small sample test and has given the test statistic known as Student's "t".

The basic assumptions used in this test are as given below

(1) The parent population or populations from which sample are taken as normal.

(2) The sample selected is random sample.

(3) The measurements of the observation in the related samples are independent.

(4) The observations are identically independently and normally distributed. In other words there is homogeneity of variances.

These are the four crucial assumptions which are required to be satisfied to have the conclusions of small samples to be valid.

**Students 't' test**

If $X_1$, $X_2$, ... $X_n$ is a random sample of size n from a normally distributed population with mean $\sigma$ and variance $\sigma^2$, then the student's 't' statistic is defined as

$$t = \frac{|\bar{X} - \mu|}{S/\sqrt{n}}$$

Where $\bar{X} = \frac{\Sigma X}{n}$ $\qquad$ $S^2 = \frac{\Sigma(X - \bar{X})^2}{n-1}$ = sample mean square

Sample mean square ($S^2$) is an unbiased estimate of population variance $\sigma^2$,

thus $\qquad t = \frac{|\bar{X} - \mu|}{S/\sqrt{n}}$

't' has student's t distribution with (n-1) degrees of freedom.

**The Applications of "t" Statistic**

(1) To compare a sample mean with population mean of one sample "t" test

(2) To compare the means of two dependent samples or to compare the mean of the same sample which are tried for "before situation" and "After situation" or paired "t" test

(3) To compare the two sample means coming from two independent samples or unpaired "t" test.

**(a) Case 1 : One Sample 't' test**

If $X_1, X_2, X_3 \dots X_n$ are n number of observation and $\overline{X}$ is a sample mean and $\mu$ is population of hypothetical mean and if $\sigma$ the standard deviation of population. Then the null hypothesis is $H_0 : \overline{X} = \mu$

$\overline{X}$ = sample mean or observed mean

$\mu$ = Assumed mean or Hypothetical mean or population Mean

$$t = \frac{|\overline{X} - \mu|}{\sigma / \sqrt{n}} \text{ with (n-1) degrees of freedom}$$

$\overline{X} = \Sigma X / n$

$\sigma$ = population standard deviation which is known, n = sample size

If population standard deviation is not known then an estimate of population standard deviation is taken which is computed from the sample and

$$S^2 = \frac{1}{n-1} \Sigma (X - \overline{X})^2$$

this $S^2$ used instead of population variance $\sigma^2$

S is used instead of population standard deviation $\sigma$ thus

$$t = \frac{|\overline{X} - \mu|}{S / \sqrt{n}} \text{ with (n-1) degrees of freedom}$$

The calculated value of t is compared with table value of "t" for (n-1) degrees of freedom for the selected level of significance. If calculated 't' > tabulated 't' then null hypothesis $H_0$ is rejected for the given level of significance

**(b) Case : 2 Paired "t" test (case of dependent samples )**

When the same sample observation are used for the two situations then paired "t" test is used. The change in level of weight before giving a particular diet and after giving a diet. The change in the level of Hemoglobin before taking tonic and after taking tonic. The change in the yield of farmers' field before giving fertilizers and after applying fertilizers are certain examples where paired 't' test can be used.

If X and Y are two variables which are the measurements of the same individuals for the two situations then student's paired "t" test in applied as given below

$H_0 : \overline{X} = \overline{Y}$

$t = \dfrac{|\bar{d}|}{SE(\bar{d})}$ with (n-1) degrees of freedom

Where $d = X-Y$ or $Y- X$ $\qquad \bar{d} = \Sigma d / n$

n = number of observations in sample X or Y

$$SE(\bar{d}) = \sqrt{\frac{\Sigma(d-\bar{d})}{(n-1)\,n}}$$

the null hypothesis $H_0$ tested is means of the before situation and after situations to be same.

Compare this observed 't' with Table 't' for the (n-1) degrees of freedom at 5 per cent or 1 per cent level of significance or any other chosen level of significance. If calculated 't' is greater than table 't' the Null Hypothesis $H_0$ is rejected, otherwise it is accepted.

**(c) Case 3 : Unpaired 't' test (the case of two independent samples)**

When the two samples are independent and it is desired to test the equality between two sample means then unpaired 't' test is used.

If $X_1$ is one sample having $n_1$ number of observations and $X_2$ is another sample having $n_2$ number of observations then the null hypothesis $H_0 : \overline{X}_1 = \overline{X}_2$ can be tested is given as follows :

$$t = \frac{\left|\overline{X}_1 - \overline{X}_2\right|}{\sqrt{S^2(1/n_1 + 1/n_2)}} \qquad \text{with } (n_1 + n_2 - 2) \text{ degrees of freedom}$$

$$S^2 = \frac{\sum(X_1 - \overline{X}_1)^2 + \sum(X_2 - \overline{X}_2)^2}{n_1 + n_2 - 2} = \text{estimate of variance of difference}$$

$\overline{X}_1 = \Sigma X_1 / n_1$ ; with sample size = $n_1$

$\overline{X}_2 = \Sigma X_2 / n_2$; with sample size =$n_2$

The observed 't' is compare with table 't' for $(n_1 + n_2 - 2)$ degrees of freedom and the null hypothesis is $H_0$ is accepted or rejected.

We have studied three cases of Students "t" test besides this Students' 't' test can be used to test the null hypothesis $H_0 : r = 0$ where r is = correlation coefficient. i.e significance of Correlation Coefficient.

Ho: $\beta = 0$ where $\beta$ = sample regression coefficient also test of partial correlation and partial regression coefficients.

The 't' test can also be used to test the significance of partial correlation coefficient and partial regression coefficient. In general student's 't' can be defined as the difference between two means to the standard error of the difference, with corresponding degrees of freedom.

### 3. F-Test

The conclusions of student's 't' tests are valid only when all the four assumptions are satisfied. Out of the four assumption the assumption number (4) of homogeneity of variances is very important. If the assumption of homogeneity is not satisfied by the samples obviously the conclusion will not be valid. Therefore, it is essential to test the validity of this assumption. This assumption can be tested with F test.

The F test is given by W. G. Snedecore, and he named it as F-test in the honor of R. A. Fisher. The F statistic is defined as the ratio of two variances.

The null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is tested as given below

$\sigma_1^2$ variance of first population and

$\sigma_2^2$ variance of second population

$$F = \frac{\sigma_1^2}{\sigma_2^2}$$

with $\sigma_1^2 > \sigma_2^2$

$\sigma_1^2$ is with $n_1$-1 degrees of freedom

$\sigma_2^2$ is with $n_2$-1 degrees of freedom

in absence of population variances we take sample mean square ($S^2$)

$$F = \frac{S_1^2}{S_2^2} \qquad \text{Where } S_1^2 > S_2^2$$

$$S_1^2 = \frac{\sum(X-\overline{X})^2}{n_1-1} = \text{ sample mean square for X}$$

$$S_2^2 = \frac{\sum(Y-\overline{Y})^2}{n_2-1} = \text{ sample mean square for Y}$$

$n_1$= sample size for X $\qquad$ $n_2$= sample size for Y $\qquad$ thus

$$F = \frac{\sum(X-\overline{X})^2/(n_1-1)}{\sum(Y-\overline{Y})^2/(n_2-1)}$$ the degrees of freedom associated are ($n_1$-1), ($n_2$-1)

The null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ is accepted if the table value of F associated with ($n_1$-1) and ($n_2$-1) is greater than calculated value of F.

In other words the null hypothesis is rejected if calculated F is greater than table value of F and we say that the two populations are not identically distributed or having no similarity in variances or having inequality of variances.

If the variances are not homogeneous. The "t" test can not be applied as such and modified t test is suggested by Cochron and Cox. (1950) can be applied.

If $n_1$ and $n_2$ are the degrees of freedom for sample one and sample two. If $S_1^2$, and $S_2^2$ are the sample mean squares (sample variances) for the two samples. Then from the table of t values (Given in appendix of this book) record the significance level $t_1$ and $t_2$. For the ($n_1$-1) and ($n_2$-1) degrees of freedom respectively then calculate fresh t as $t'$ is given as

$$t' = \frac{(W_1 t_1 + W_2 t_2)}{(W_1 + W_2)}$$

Where $W_1 = S_1^2 = \frac{\Sigma(X-\overline{X})^2}{(n_1-1)}$ and $W_2 = S_2^2 = \frac{\Sigma(Y-\overline{Y})^2}{(n_2-1)}$

$t_1$ is table t value associated with $(n_1$-1)df

$t_2$ is table t value associated with $(n_2$-1)df

The value of $t'$ will be in between $t_1$ and $t_2$.

Thus $t'$ is the table value for the given level of significance and given degrees of freedom. Compare the t with $t'$ thus $t'$ can be used as table value for comparison with observed t.

**(4) Chi square test ($\chi^2$- test )**

When the data is not obeying the assumptions of normality and the sample size is small either with frequency; we use small sample test or non parametric statistical tests or $\chi^2$ chi-square test for frequencies

**Chi-square :** If $X_1$, $X_2$, ....$X_n$ are n independent random variables following normal distribution with means $\mu_1, \mu_2 \ldots \mu_n$ and standard deviation as $\sigma_1, \sigma_2, \ldots \sigma_n$ respectively then the variate

$$\chi^2 = \frac{(X_1 - \mu_1)^2}{\sigma_1^2} + \frac{(X_2 - \mu_2)^2}{\sigma_2^2} + \ldots\ldots + \frac{(X_n - \mu_n)^2}{\sigma_n^2}$$

$= \Sigma\left[\frac{(X_i - \mu_i)}{\sigma_i}\right]^2$ which is a sum of the squares of n independent standard normal varieties follows $\chi^2$ chi-square distribution with n degrees of freedom.

**Application :-** $\chi^2$ chi-square test has many applications

(1) Testing goodness of fit

(2) Testing the independence of attributes

(3) Testing equality of variances for several population

**To Test the Goodness of Fit**

The null hypothesis $H_0$: is that there is no significant difference between observed and the theoretical or expected values of frequencies.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

If $O_1$, $O_2$, …. $O_n$ are observed frequencies and if $E_1$, $E_2$, ….$E_n$ are expected frequencies then $\chi^2$ can be estimated as above.

**Limitations**

***(i) Student's 't' Test***

(1) It is used for small samples

(2) It requires the samples from independently, randomly, identically and normally distributed population, if the above conditions are not satisfied the conclusions of 't' tests are not valid

***(ii) F-Test***

The F test is used for comparing variances of two populations and testing equality of several means. F test is also valid when the errors of observation are independently, identically and distributed with normal distribution. Thus normality of a population is a crucial assumption for these two tests.

***(iii) The $\chi^2$ Test***

(1) The $\chi^2$ (chi-square) statistic is valid only when the total frequency is reasonably large

(2) The sample observations should be independent. This indicates that the cell frequencies should not overlap on each other.

(3) The constraints on cell frequencies should be linear. They should not involve. Squares and higher powers of frequencies.

(4) The theoretical frequency (Expected.) frequency should not be very small. The theoretical frequency should not be smaller than 5. If the expected frequency of the cell is less than 5 then apply some other non parametric test or club the cells by merging the adjacent cells.

(5) The given distribution should be of original frequencies and should not be replaced by relative frequencies of proportions.

## 12.3 Glossary

**Hypothesis** : A statistical hypothesis is the statement regarding the population parameter

**Null Hypothesis** : A full hypothesis is the hypothesis set for possible equality or possible rejection

**Level of Significance :** The level of significance is the probability of rejecting null hypothesis $H_0$ when infact it is true.

**Type I Error (α) :** When the null hypothesis $H_0$ is rejected and it is ought to be accepted

**Type II Error (β) :** When the null hypothesis $H_0$ is false and the null hypothesis is accepted.

**Confidence Coefficient :** The probability of accepting $H_0$ when infact it is true.

**Statistic** : Statistic is a function of observations obtained from the sample. It is a sample characteristic

**Parameter** : Parameter is an unknown constant of the population distribution. It is a population characteristic

**Critical Region :** Critical region is the region of rejection .

**Test** : A statistical test is the criteria to reject or to accept the null Hypothesis.

## 12.4 Summary

The Research Hypothesis is the prediction derived from the theory under test. The Hypothesis is a statement regarding the population parameter. The Hypothesis are of two types i.e. Null Hypothesis ($H_0$) which is set for possible rejection under the assumption, that it is true and Alternative Hypothesis ($H_1$) is the set for possible acceptance or possible inequality. The procedure for Hypothesis testing involves several steps, such as Null and Alternative Hypothesis, Choice of Testing, Significance, Statistical Errors, Sampling Distribution, Critical Region, Degrees of Freedom and so on. The different statistical test such as Z test, Students ‘t’ test, F-test, $\chi^2$ (chi-square) test etc. are used for significance of results.

## 12.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. Define hypothesis and describe different types of hypotheses in statistical analysis.
2. What are the different steps involved in testing the hypothesis
3. Describe ‘t’ test, Z-test and $\chi^2$ (chi-square) test and differentiate among them.
4. Define and distinguish between parametric and non parametric statistical tests.

# Unit 13 : Statistical Concepts

***Index***



## 13.1 Introduction

The important objective of the data analysis is to determine and measure features of the frequency distribution. The first of such measure is Averages or Measurement of Central Tendency. To draw conclusions regarding the characteristics of huge information and data collected in the numerical form, we require a single or selected figure to grasp entire set of data or information. The information either qualitative or quantitative is separated in the form of classification, tabulation and frequency distribution. This data or information is required to be condensed and one of the typical representive values is required to describe the entire set. This typical value which is representative single figure of huge, unwieldy numerical data is known as measures of Central Tendency or Averages.

**After the study of this unit, you will be able to know and understand :**

- λ Measures of typical size and variability
- λ Correlation and regression
- λ Non-parametric tests

## 13.2 Content

### 13.2.1 Measures of Typical Sizes

They are useful in describing the data in a concise and meaningful manner, comparison of data becomes possible and the measure of distribution of observation or spread of the data can be studied.

If values in the data are plotted in a graph, it will be observed that points are clustering around a particular point or value; that means there is tendency of observation or values to clusters or center around a particular value. This tendency is called as central tendency and the point or value around which values of the data are clustering is called as a measure of central tendency or an average. As values or observations are located around a central value, it is known as a measure of location.

Some of the definitions central tendency or average given by the eminent statisticians are given below :

(1) An average is a single value selected from a group of values to represent them in some way; a value which is supposed to stand for whole group of which, it is a part, a typical of all the values in the group. *(A. E. Wavgh)*

(2) Averages are statistical constants which enable us to comprehend in a single effort the significance of the whole. *(A. L. Bowley)*

(3) An average value is a single value within the range of the data, that is used to represent all the values in the series. Since an average is somewhere within the range of the data, it is called as measure of central value. *(Roxton and Cowden)*

**Requisites of a good average or measure of central tendency :** According to *Prof. Yule* stated the ideal average is that which satisfies following conditions.

(1) It should be well defined or it should be rigidly defined. The definition of average should be rigid because there should not be any thing for the speculation of the investigator or interpreter. It should be unambiguous. There should not be any scope for bias from the observer. It should lead one and only one interpretation from the different observers.

(2) It should be readily comprehensible and easy to calculate. The measure must be easy to understand and calculate. It should not have complicated and more number of arithmetic calculations.

(3) It should be based on all observations. The average is a representative figure of series. Therefore it must utilize all the observations of the series. Thus whatever data is at our disposal must be utilized for computing the average and there should not be loss of information resulting from non utilization of certain figures from the series.

(4) It should be amenable for further a mathematical treatment. The average must be suitable for the further calculations, if desired and required for further interpretation. If the average is not suitable for further mathematical treatment the scope of the same will be limited.

(5) It should not be affected much by the fluctuations of sampling. There should be sampling stability to the average. To compute, we take a random sample and then calculate average. If the same procedure is repeated with equal size of sample, we may get different averages from different samples. The various averages, which we have obtained from different samples, must give values much closer to each other. The different values, which we get, from different samples is a sampling fluctuation. Thus this sampling fluctuation should not change the value of an average to larger extent.

(6) It should not be affected by the extreme observations. Sometimes there are unduly small values and very large values of observations, which are not truly representative for the sample. These values affect the value of average. Therefore ideal average should be such that it should be least affected by extreme observations.

## 1. Various Measures of Central Tendency

The various measures of central tendency are described below :

### (a) Arithmetic Mean (AM)

It is defined as the sum of the given observations to the total number of observations in the series. e.g. if 1, 2, 3, 4, & 5 are the five observations.

Then Arithmetic mean (AM) is given by $\text{AM} = \dfrac{1+2+3+4+5}{5} = 3$

Thus it can be defined as, if $X_1$ , $X_2$, $X_3$ ——— and Xn are 'n' number of observations in series the Arithmetic mean (AM) is denoted by $\overline{X}$ and is obtained by the formula

$$\overline{X} = \frac{X_1 + X_2 + X_3 - - - - + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Where Σn = Summation of 'n' number of individuals, i = 1, i = 1 to n

Some times the individual observations are repeated for different number of times. This leads to the data of discrete frequency distribution such as the data of number of flowers per plant and number of plants

| No. of flowers / plant | 5 | 4 | 2 | 10 |
|---|---|---|---|---|
| No. of plants | 4 | 3 | 4 | 1 |

Here number of flowers per plant is variable (X) and number of plants is frequency(F)

| Particulars | | | | | Total |
|---|---|---|---|---|---|
| X | 5 | 4 | 2 | 10 | - |
| F | 4 | 3 | 4 | 1 | 12 |
| FX | 20 | 12 | 8 | 10 | 50 |

The Arithmetic mean is worked out as given below

| F | X | $F_x$ |
|---|---|---|
| 5 | 4 | 20 |
| 4 | 3 | 12 |
| 2 | 4 | 8 |
| 1 | 8 | 8 |
| 0 | 1 | 0 |
| | ΣF= 20 | ΣFX=48 |

$$\text{AM} = \frac{F_1X_1 + F_2X_2 + - - - - - + F_5X_5}{F_1 + F_2 + F_3 + - - - - - + F_5}$$

$$= \frac{4\times5+3\times4+4\times2+8\times1+1\times0}{20}$$

$$= \frac{48}{20} = 2.4$$

Thus if $X_1, X_2, X_3, \ldots\ldots\ldots\ldots\ldots X_k$ are K number of distinct individual observations having their frequencies $F_1, F_2, F_3, \ldots\ldots\ldots\ldots F_k$ respectively then the arithmetic mean (AM) is defined as

$$AM = \frac{\sum FX}{\sum F} = \frac{F_1X_1 + F_2X_2 + ----+F_kX_k}{F_1 + F_2 + F_3 + -----+F_k}$$

$$= \frac{\sum_{i=1}^{k} Fi\ Xi}{\sum_{i=1}^{k} Fi} \qquad \text{where } \sum_{i=1}^{k} = \text{Sum of k number of individuals}$$

Similarly third situation can be as given below. The data given below is marks in statistics and number of students getting the marks. Calculate the arithmetic mean or average number of marks per student.

| Marks | 10-20 | 20-30 | 0-10 | 30-40 |
|---|---|---|---|---|
| No. of Students | 4 | 4 | 1 | 1 |

The marks are given in class intervals, No. of students is frequency of the class. If we write the data in systematic order and in tabular form or it will be

| Class interval | Frequency |
|---|---|
| 0-10 | 1 |
| 10-20 | 4 |
| 20-30 | 4 |
| 30-40 | 1 |

Then if we calculate the midvalue of class interval and write down the frequency of the above problem. It can be same as that of the earlier one i.e

| Class interval | Midvalue m | F | Fm |
|---|---|---|---|
| 0-10 | (0 + 10 ) / 2 = 5 | 1 | 5 |
| 10-20 | (10 + 20) / 2 = 15 | 4 | 60 |
| 20-30 | (20 + 30) / 2 = 25 | 4 | 100 |
| 30-40 | (30 + 40) / 2 = 35 | 1 | 35 |
| | Total | ΣF =10 | ΣFm= 200 |

$$AM\ (\overline{X}) = \frac{\sum Fm}{\sum F} = \frac{F_1m_1 + F_2m_2 ----+ F_km_k}{F_1 + F_2 + F_3 ----+ F_k}$$

$$= \frac{1\times5+4\times15+4\times25+1\times35}{1+4+4+1} = \frac{200}{10} = 20 \text{ marks / student}$$

Thus if the data is in continuous series with frequencies, then calculate the midvalues of the class intervals and use midvalues of class for calculating Average Mean (AM).

Thus if the series is distributed in class interval as Xo to Xi, Xi to X 2i, X2i to X3i … $2X_{k-1}$ to Xk with frequencies Fi ........ Fk then

$$AM = \frac{\sum F_iM_i}{\sum Fi} = \frac{\sum Fm}{\sum k} \qquad \text{Where mi} = \frac{X_{i-1} + X_i}{2} = \text{midvalue}$$

And Fi = frequents of the i$^{th}$ class, i = 1 to k

$$AM = \frac{F_1m_1 + F_2m_2 ----+ F_{k-1}m_{k-1} + F_kX_k}{F_1 + F_2 ----+ F_k}$$

**(b) Geometric Mean (GM)**

Geometric mean of set of ‘n’ observations is the ‘n’th root of their product.

***For example : 1***

The geometric mean (GM) of 16 and 4 is given below

The product: 16 x 4 = 64

n = number of observations = 2

n$^{th}$ root of the product = (GM) = $(64)^{½}$

$= \sqrt{64} = 8$

***For example : 2***

The geometric mean of 8,2,4 is

$$GM = (8\times2\times4)^{1/3} = \sqrt[3]{8\times2\times4} = \sqrt[3]{64} = \sqrt[3]{4\times4\times4} = 4$$

Here n=3 product: 64 and cube root of 64 is 4. Hence Geometric mean is 4

Thus for discrete series geometric mean can be defined is given as follows :

***(i) Geometric mean for Discrete Series***

If $X_1, X_2, X_3 \ldots\ldots\ldots X_n$ are 'n' number of observations, then geometric mean (GM) can be defined as 'n' th root of the product of 'n' number of observations

$$GM = (X_1 \times X_2 \times X_3 \times X_4 \ldots \times X_n)^{1/n} = \sqrt[n]{X_1 \times X_2 \times X_3 \times X_4 ----\times X_n}$$

This can be easily calculated by taking logarithms

$$\text{Log (GM)} = \frac{1}{n}(\text{Log } X_1 + \text{Log } X_2 \ldots\ldots\ldots\ldots\ldots + \text{Log } X_n)$$

$$= \frac{1}{n}\Sigma \text{ Log } X_i$$

$$GM = \text{Antilog}\left(\frac{1}{n}\Sigma \text{ Log } X_i\right)$$

***Thus geometric mean can be defined as the antilogarithm of the arithmetic mean of logarithms for the n number of observations.***

***For Example :***

Calculate the geometric mean of the following data 4,2,3,1 and 5, write down the data in the tabular form

| Sr. No. | Observations (Items X) | Logarithms (X) |
|---|---|---|
| 1 | 4 | 0.6020 |
| 2 | 2 | 0.3010 |
| 3 | 3 | 0.4771 |
| 4 | 1 | 0.0000 |
| 5 | 5 | 0.6987 |
| | No. of observations n=5 | Total= Σ Log X = 2.0788 |

$$\text{Log (GM)} = \frac{1}{n}\Sigma \text{ Log X} = \frac{2.0788}{5} = 0.4158$$

$$GM = \text{antilog } (0.4158) = 2.6047$$

*(ii) Geometric Mean for Grouped Discrete Series*

When the items or observations are with their frequencies the geometric mean can be worked out as per the formula we have used for ungrouped data.

Let us take the example. Following is the set of observations from the experiment for number of pods per plant and number of plants having pods. Calculate geometric mean.

| | | | | |
|---|---|---|---|---|
| No. of pods per plants | 2 | 3 | 4 | 5 |
| Number of plants | 4 | 3 | 2 | 1 |

| | | | | |
|---|---|---|---|---|
| X = No. of pods | 2 | 3 | 4 | 5 |
| F = Number of plants | 4 | 3 | 2 | 1 |

The item X has frequency F means that X is repeated F number of times.

The item X= 2 has frequency F= 4, implies that it is repeated for four number of times and the item (X) =5 has frequency F= 1, implies that it is repeated for only one number of time. Thus geometric mean can be calculated by taking ΣF= n.4+3+2+1=10=$10^{th}$ root of the product of all 'n' individual items or in other words

GM = $(2 \times 2 \times 2 \times 2 \times 3 \times 3 \times 3 \times 4 \times 4 \times 5)^{1/10}$

= $(16 \times 27 \times 16 \times 5)^{1/10} = (34560)^{1/10}$

Log (GM) = 1/10 Log (34560)

= 1/10 (4.53867) = 0.453867

GM = Anti log (0.453867) = 2.84

Or

GM = $(2^4 \times 3^3 \times 4^2 \times 5^1)^{1/10}$

Log GM = 1/10 [4 log 2 + 3 log 3 +2 log 4 + 1 log 5]

Log GM = 1/10 [4 x 0.3010 + 3 x 0.4771 + 2 x 0.6020 +1x 0.6989]

= 1/10 [1.204 + 1.4313 + 1.2040 + 0.6989]

= 4.5382 / 10 = 0.45382

GM = Antilog (0.45382) = 2.84

The same observations can be written in the following form

| X | F | Log X | F log X |
|---|---|---|---|
| 2 | 4 | Log 2 = 0.3010 | 4x0.3010 = 1.2040 |
| 3 | 3 | Log 3 = 0.4771 | 3x0.4771 = 1.4313 |
| 4 | 2 | Log 4 = 0.6020 | 2x0.6020 = 1.2040 |
| 5 | 1 | Log 5 = 0.6989 | 1x0.6989 = 0.6989 |
| Total | ΣF = 10 | - | ΣF log X = 4.5382 |

$$\text{Log GM} = \frac{\sum F \log x}{\sum F} = \frac{4.5382}{10} = 0.45382$$

GM = antilog (0.45382) = 2.84 pods per plant

***(iii) Grouped data : (Continuous distribution)***

Whenever the data presented is given for class intervals and frequency, the geometric mean can be estimated as per the procedure explained earlier by replacing the discrete (X) with the midvalue (M) of the class interval and carrying out the same procedure.

The same is explained with the given example

The data of milk yield and number of cows is as given from the herd of cows estimate geometric mean.

| Milk yield in liters | 0-2 | 2-4 | 4-6 | 6-8 | 8-10 |
|---|---|---|---|---|---|
| Number of cows | 1 | 2 | 4 | 1 | 2 |

| Class interval | 0-2 | 2-4 | 4-6 | 6-8 | 8-10 |
|---|---|---|---|---|---|
| Frequency F | 1 | 2 | 4 | 1 | 2 |

| Midvalue of class interval | $\frac{(0+2)}{2}=1$ | $\frac{(0+4)}{2}=3$ | $\frac{(0+6)}{2}=5$ | $\frac{(0+8)}{2}=7$ | $\frac{(0+10)}{2}=9$ |
|---|---|---|---|---|---|
| Frequency | 1 | 2 | 4 | 1 | 2 |

| Class | | M | F Log m | F log m |
|---|---|---|---|---|
| 0-2 | 1 | 1 | Log 1 =0 | 1 x 0 =0 |
| 2-4 | 3 | 2 | Log 3 =0.4771 | 2 x 0.4771 =0.9542 |
| 4-6 | 5 | 4 | Log 5 =0.6989 | 4 x 0.6989 =1.3978 |
| 6-8 | 7 | 1 | Log 7 =0.8450 | 1x0.8450 =0.8450 |
| 8-10 | 9 | 2 | Log 9 =0.9542 | 2x0.9542 =1.9084 |
| Total | | ΣF =10 | - | ΣF, log m = 5.1054 |

m = mid value of the class interval = $\frac{(\text{Upper limit} + \text{Lower limit of class})}{2}$

$$\text{Log GM} = \frac{\sum F \log m}{\sum F} = \frac{5.1054}{10} = 0.51054$$

Log GM = 0.51054

GM = Antilog (0.51054) = 3.239 = 3.24

GM = 3.24 liters per cow

Thus geometric mean for continuous series is defined as

If $m_1, m_2, \ldots\ldots\ldots$ mk are the mid values of the class intervals and

If $F_1, F_{2,} \ldots\ldots\ldots$ Fk are the frequencies of each class then

Geometric mean (GM) = $\left(\frac{\sum F \log m}{\sum F}\right)$ Where 'm' is middle value of class

Where 'm' = $\frac{\text{Lower limit of class} + \text{Upper limit of class}}{2}$ = mid value

and F = Frequency of class

**(c) Harmonic mean (Ungrouped data)**

If $X_1, X_{2,} X_{3,} \ldots\ldots$Xn are 'n' number of observations, then Harmonic mean (HM) can be defined as

$$\text{HM} = \frac{1}{H}$$

Where H = $\frac{1}{n}\Sigma\left(\frac{1}{X}\right)$ Or HM = $\dfrac{1}{\frac{1}{n}\Sigma\left(\frac{1}{X}\right)}$

In other words harmonic mean is defined as the reciprocal of the arithmetic mean of reciprocals.

Thus if $X_1$, $X_{2,}$ $X_{3,}$ ......... $X_n$ are 'n' number of observations, then Harmonic mean is given by

$$HM = \frac{n}{(1/X_1 + 1/X_2 ---- + 1/X_n)}$$

Estimate the average speed when a traveler travels by his bicycle from place A to B with a speed of 10 km per hour, from place B to C with 15 km per hour and from place C to A with a speed of 12 km per hour

Let Speed = X

| X | 1/X | |
|---|---|---|
| 10 | 1/10 | 0.1000 |
| 15 | 1/15 | 0.06666 |
| 12 | 1/12 | 0.08333 |
| | Σ 1/X | 0.24999 |

= 0.24999

Harmonic Mean (HM) = $\frac{1}{H}$

Where H = $\frac{1}{n}\Sigma\left(\frac{1}{X}\right)$

= $\frac{1}{3}\times(0.24999)$

= 0.08333

HM = $\frac{1}{H}$ = $\frac{1}{0.8333}$

= 12.0008 = 12

***(i) Harmonic Mean (Discrete series: Grouped data)***

Calculate harmonic mean for the following data :

X= item F= frequency

| X | 4 | 8 | 5 | 2 |
|---|---|---|---|---|
| F | 3 | 2 | 1 | 4 |

| X | F | F/X |
|---|---|---|
| 4 | 3 | 3 / 4 = 0.75 |
| 8 | 2 | 2 / 8 = 0.25 |
| 5 | 1 | 1 / 5 = 0.20 |
| 2 | 4 | 4 / 2 = 2.00 |
| Total | ΣF 10 | Σ F/X= 3.20 |

Harmonic Mean (HM) = $\frac{1}{H}$

$$H = \frac{1}{\sum F}\sum\left(\frac{F}{X}\right) = \frac{1}{10} = (3.20) \qquad H = 0.320$$

$$HM = \left(\frac{F}{H}\right) = \left(\frac{10}{3.20}\right) = 3.125$$

***(ii) Harmonic Mean (Grouped data continuous series)***

Calculate Harmonic Mean for following data :

| Class | 1-3 | 3-5 | 5-7 | 7-9 | 9-11 |
|---|---|---|---|---|---|
| Frequency | 2 | 8 | 6 | 4 | 5 |

| Class interval | Mid value | Frequency F | F/ X |
|---|---|---|---|
| 1-3 | 2 | 2 | 2/2 = 1 |
| 3-5 | 4 | 8 | 8/4 = 2 |
| 5-7 | 6 | 6 | 6/6 =1 |
| 7-9 | 8 | 4 | 4/8 = 0.5 |
| 9-11 | 10 | 5 | 5/10 =0.5 |
| | | ΣF = 25 | Σ(F/X ) = 5 |

HM = 1 / H where H = $\left(\frac{1}{\sum F}\right)\sum\left(\frac{F}{X}\right)$

ΣF = 25, Σ (F / X) = 5

$$H = \frac{1}{25} \times 5 = \frac{5}{25} = \frac{1}{5} = 0.20$$

$$HM = \frac{1}{H} = \frac{1}{0.20} = 5$$

***(iii) Combined mean of two samples***

The arithmetic mean, geometric mean and harmonic mean are known as mathematical averages because further mathematical treatment is possible for these three averages. e.g. if A and B are two samples having $n_1$ and $n_2$ number of observations respectively and arithmetic means as $\overline{X}_1, \overline{X}_2$ respectively. Then the arithmetic mean of the two is possible such as $\overline{X}$

$$\overline{X} = \text{Combined Mean of A \& B}$$

$$\overline{X} = \frac{n_1 \overline{X}_1 + n_2 \overline{X}_2}{n_1 + n_2}$$

**(1) Merits and demerits of mathematical averages**

*Merits*

1. Arithmetic mean is rigidly defined
2. It is easy to understand and simple to calculate
3. It is based on all observations.
4. The algebraic treatment is possible for further calculations
5. The arithmetic mean is least affected by the fluctuations of sampling

*Demerits*

1. It is very much affected by the extreme observations.
2. It can not be obtained, when the data is with open ended classes.
3. It leads to wrong conclusions, if the conclusions are drawn from incomplete or insufficient information.

**(2) Geometric Mean Uses and Demerits**

*Uses*

1. It is used in the construction of index numbers.
2. To measure the growth pattern when the population is multiplicative in nature e.g. Growth, population of insects, bacteria etc.

*Demerits*

1. It is not easy to understand.
2. If one of the observations from the series is zero, then the geometric mean comes to zero. Similarly if one of the items is negative, geometric mean is imaginary every though other items are having meaningful magnitude.

**(3) Harmonic Mean**

*Merits*

1. It is rigidly defined
2. It is based on all observations
3. It is suitable for further mathematical calculations.
4. It is also not affected much by the fluctuations of sampling
5. It gives more weightage to smaller items

*Demerits*

1. It is not easy to understand and interprete.
2. It can not be obtained, when one of the items is zero.
3. It can be used only when it is desired to give higher weightage to smaller items.

*Uses*

1. Harmonic mean is useful when the average of rates and ratios is to be calculated, where the time factor is variable and distance is constant.

**(d) Median**

Median is also known as the positional average by virtue of its position it can be located. Whenever the observations have a weak measurement scale and sometimes starting point of the scale is also not well defined, the items can only be ordered, for such types of situations the arithmetic mean, geometric Mean and harmonic means are not suitable averages to describe the situation. For such types of situation median and mode are good types of averages.

Median is the middle item of the series when items are arranged in ascending or descending order of magnitude.

*Example :* The items of the series are 14, 11, 51, 21, 31 then the observations can be arranged in ascending order such as 11,14,21,31,51 then middle item is 21. Hence median is 21.

Similarly the observations can also be arranged in descending order such as 51, 31, 21, 14, 11 thus mid items of the series is 21. Hence median is 21

The items of the series are 21, 10, 42, 33, 63, 74. If we arrange the data in ascending or descending order of magnitude we get

$B_1$ 10, 21, 33, 42, 63, 74. (ascending order)

$B_2$ 74, 63, 42, 33, 21, 10. (descending order)

Here the middle item lies in between two observations i.e. 33 and 42. Hence median is the arithmetic mean of these two figures i.e. median.

$$\text{median} = \frac{(33+42)}{2} = \frac{75}{2} = 37.5$$

- λ Thus in situation, there were five number of observations and median was located as item. No. $\frac{(5+1)}{2}$ i.e. 3rd item, when the observations were arranged in ascending order.
- λ In example (B) there are six items Hence the middle item was located by taking the average of 3rd and 4th item.

Thus it can be understood that when the number of items (N) are odd; then median is given by the item number $\frac{(N+1)}{2}$. Secondly when there are even number of items i.e. n is even number, then the median can be worked out by taking the arithmetic mean of the values of item number $\frac{N}{2}$ and $\frac{N}{(2+1)}$ th item.

Thus there are six items. The value of $\frac{6}{2}$ = 3rd item is 33 and the value of item number $\frac{N}{2+1}$ = 4th item is 42.

$$\text{Therefore median} = \frac{(33+42)}{2} = 37.5$$

***(i) Median (for grouped data discrete series)***

When the items and their frequencies are there and median is required to be estimated then the items are arranged in ascending or descending order of magnitude. The frequency of the item is written against the item. The cumulative frequency is worked out and the item bearing position N / 2 or 0.5 N is designated as median. Let us explain the same with an illustration. Let the data be a number of fruits per plant and number of plants.

| | | | | | |
|---|---|---|---|---|---|
| Number of fruits | 10 | 7 | 2 | 5 | 4 |
| Number of plants | 4 | 8 | 7 | 8 | 3 |

The items X are number of fruits per plant and the number of plants is frequency F. Let us arrange the items, in ascending or descending order of magnitude

| X | F | CF |
|---|---|---|
| 2 | 7 | 7 |
| 4 | 3 | 10 =7+3 |
| 5 | 8 | 18 =10+8 |
| 7 | 8 | 26 =18+8 |
| 10 | 4 | 30 = 26+4 |

Where CF= Cummulative Frequency

ΣF = N = 30.

The median is the item number $\frac{N}{2} = \frac{30}{2}$ = 15. The item number 15 is greater than 10 and smaller than 18. Hence the item 5 is median.

***(ii) Median (For grouped data- continuous series)***

Wherever the class intervals and frequency is given median class is located as per the same procedure given above. Here median class is located first and there after, the value of median is estimated by taking into consideration the width of class interval, frequency of median class, cumulative frequency of just preceding class to median class and lower limit of median class. Following steps are involved in locating median for continuous series.

1. Arrange the class intervals in ascending or descending order. Write down the frequency against each class.
2. Find out cummulative frequency of each class and ΣF = N

3. The item number $\frac{\Sigma F}{2} = \frac{N}{2}$ = median number.

Select the class interval in which item number $\frac{N}{2}$ (median number) lies.

4. Let lower limit of median class be L, and frequency of median class be F. The width of median class = U – L = I , U-upper limit L-lower limit

5. Find out the cummulative frequency of first earlier class to median class = C then median is worked out as median (Me) = $L + \left[\left(\frac{N}{2} - C\right) \times \frac{i}{F}\right]$

*Example :* Calculate Median from the following data

| Class intervals | Frequency |
|---|---|
| 10-15 | 14 |
| 15-20 | 18 |
| 25-30 | 14 |
| 5-10 | 17 |
| 0-5 | 15 |
| 20-25 | 22 |

Arrange the data in systematic order of class intervals

| Class | Frequency F | Cummulative frequency CF |
|---|---|---|
| 0-5 | 15 | 15 |
| 5-10 | 17 | 15+17 = 32 |
| 10-15 | 14 | 32+14 = 46 |
| 15-20 | 18 | 46+18 = 64 |
| 20-25 | 22 | 64+22 = 86 |
| 25-30 | 14 | 86+14 =100 $\therefore$ F = N |

$\Sigma F$= N= 100

Median number (N / 2) = (100 / 2) = 50. The item numbers 50 lies in the class 15-20 because the item number 50 is greater than 46 and smaller than 64. Hence 15 - 20 is median class. Lower limit of median class (L) = 15, upper limit of median class (U= 20), width of median class ( i ) = U – L = 20 – 15 = 5, frequency of median class (F) = 18, cummulative frequency of just proceedings class to median class (C) = 46

Median (Me) $= \dfrac{\left[\left(\dfrac{N}{2}-C\right)\times i\right]}{F} = 15 + \dfrac{[(50-46)\times 5]}{18}$

$= 15 + \dfrac{20}{18} = 15 + 1.11 = 16.11$

$\therefore$ Median = 16.11

**(e) Mode**

Mode is defined, as the common most item of the series. The item having the highest frequency is known as mode. Some times there are two items having the same frequency which is the highest of all. Then such type of series is called as bimodal series and the two items having the highest frequency are modes of the series. Thus we can expect more than two modes for a series. Suppose a sample of 10 random numbers is as given below 1,0,2,0,3,4,0,5,9,2. In the above series of 10 numbers the item 0 has occurred 3 number of times. The time 2 has frequency 2 and all other items are having frequency 1. Hence mode of the series = 0.

Hence to calculate mode from the series write down the items and their frequencies; inspect the item having the highest frequency and the item of maximum frequency is mode. The above problem can be solved by writing down a discrete frequency distribution of item.

| Item X | 1 | 0 | 2 | 3 | 4 | 5 | 9 |
|---|---|---|---|---|---|---|---|
| Frequency F | 1 | 3 | 2 | 1 | 1 | 1 | 1 |

Total items = 10 $\therefore$ F = 10, the item 0 has maximum frequency.

Hence Mode = 0

***Mode for Continuous Distribution***

When the data is continuous frequency distribution mode can be worked out by using the following procedure. We shall explain the method with suitable example.

1. Write the class intervals and their frequencies in systematic or ascending order.
2. Choose the class, which has the highest frequency as modal class. The frequency of model class is denoted as Fm

3. Find out the preceding class to modal class. The frequency of preceding be denoted as Fp. ssimilarly find out succeeding class to modal class. Call the frequency of succeeding class Fs

4. Let L is lower Limited of modal class, U is upper limit of modal class then the width of modal class = U – L = i

   The mode is given by

$$\text{Mode (Mo)} = L + \left(\frac{\Delta 1 \times i}{\Delta 1 + \Delta 2}\right) \text{ when } F_p > F_s$$

$$\text{Mode (Mo)} = U - \left(\frac{\Delta 2 \times i}{\Delta 1 + \Delta 2}\right) \text{ when } F_s > Fp$$

   Where $\Delta 1 = Fm - Fp$ $\quad \Delta 2 = Fm - F_s$

   i = width $\quad$ U = upper limit $\quad$ L = lower limit

   The above formula can also be written as

$$\text{Mode (MO)} = L + \left(\frac{(Fm - F_p) \times i}{(2Fm - F_p - F_S}\right) \text{ when } F_p > F_s$$

$$\text{Mode (MO)} = U - \left(\frac{(Fm - F_S) \times i}{(2Fm - F_p - F_S)}\right) \text{ when } F_s > F_p$$

***Example :*** Calculate mode of the following data

| Class | 4-6 | 2-4 | 8-10 | 0-2 | 6-8 |
|---|---|---|---|---|---|
| Frequency | 6 | 3 | 2 | 4 | 5 |

**Step 1**

Write down the class intervals in systematic order

| Class | 0-2 | 2-4 | 4-6 | 6-8 | 8-10 |
|---|---|---|---|---|---|
| Frequency | 4 | 3 | 6 | 5 | 2 |

Modal class is 4 - 6 :

Fm = frequency of modal class = 6

$F_p$ = frequency of preceding class = 3

$F_s$ frequency of succeeding class = 5

i = width of modal class 6-4 = 2

L = Lower limit of modal class = 4

U = upper limit of modal class = 6

$F_s > F_p$, hence the formula is used accordingly

$\Delta 1 = Fm - F_p = 6\text{-}3 = 3$

$\Delta 2 = Fm - F_s = 6\text{-}5 = 1$

**Step 2**

Here $F_s > F_p$

Model (MO) $= U - \left(\frac{\Delta 2 \times i}{\Delta 1 + \Delta 2}\right)$

$= 6 - \left(\frac{1 \times 2}{3+1}\right) = 6 - 0.5 = 5.5$

***Example :*** Calculate mode from the following

| Class | 0-10 | 10-20 | 20-30 | 30-40 | 40-50 |
|---|---|---|---|---|---|
| Frequency | 4 | 16 | 20 | 10 | 8 |

1. Modal class: 20-30,

   $F_p < F_s$ hence the formula is used accordingly

2. Fm = 20 $F_p = 16$; $F_s = 10$;

   $F_p < F_s$ hence the formula is used accordingly

   $\Delta 1 = Fm - F_p = 20\text{-}16 = 4$

   $\Delta 2 = Fm - F_s = 20\text{-}10 = 10$

   i = U – L = 30-20 = 10

3. L= 20

   U= 30

   Mode $= L + \left(\frac{\Delta 1 \times i}{\Delta 1 + \Delta 2}\right)$ because $F_p < F_s$

   $= 20 + \left(\frac{4 \times 10}{4+10}\right) = 20 + 40 / 14$

   $= 20 + 2.85 = 22.85$

**(f) Weighted Mean**

When individual values in the sample are of different importance, weighted mean is used.

Let $X_1$ .........Xn be n values with weights $w_1$, $w_2$ ........ $w_n$ respectively, then weighted mean is defined as

$$\overline{X} = \frac{w_1X_1 + w_2X_2 ---- + w_nX_n}{w_1 + w_2 ---- + w_n}$$

$$= \frac{\sum w_iX_i}{\sum w_i}$$

## 13.2.2 Variability

We have studied the measures of locations or averages, different measures of central tendencies give adequate knowledge of a series but this is not just sufficient to know the spread of the data. We shall explain the situation with following example.

Let there be A, B and C series as given below

A. 5, 3, 1, 2, 4

B. 3, 3, 3, 3

C. 2, 4

The arithmetic mean of series A, B and C is same = 3 i.e. for

5+3+1+2+4

A : AM $= \frac{5+3+1+2+4}{5} = 3$

B : AM $= \frac{3+3+3+3}{4} = 3$

C : AM $= \frac{2+4}{2} = 3$

Thus in series (A) there are five observations, the range of observations is from 1 to 5 and the mean is 3. In series (B) all observations are same; range is zero but mean is 3; similarly for series. (C) The range is 2 to 4 and mean is 3.

Thus in all three series, (A), (B) and (C) the number of observations are different but the arithmetic mean is same, Hence, we can not judge what is the spread of observations from arithmetic mean. It means that the average only will not be able to tell complete picture of the data. Therefore, we need another measure, which will give us the knowledge of spread of observations. This measure is measure of Dispersion or spread of observation. We define Measures of dispersion or Measures of variability as the Measures, which give us the knowledge of spread of observations from one individual to another, and variations of observations from the average value of the distribution.

Thus the measure of variations study the variability amongst observations as well as variability from the average value.

**(a) Measures of Dispersion**

The various measures of dispersion are

1. Range
2. Quartile deviation or semi Interquartles range
3. Mean deviation
4. Standard deviation
5. Coefficient of variations

Characteristics of ideal measure of dispersion

1. It should be rigidly defined
2. It should be easy for calculation and easy to understand
3. It should be based on all observations or it should utilize all the given information.
4. It should be accessible for further mathematical treatment.
5. It should be least affected by the fluctuations of sampling

**(1) Range**

Range is defined as the difference between the largest and smallest item. (11, 1, 2, 20, 25). There are five observations. (11, 1, 2, 20, 25) the range observations is from 1 to 25.

The value of Range is 25-1 = 24

The highest observation = 25

The smallest observation = 01

**(2) Quartile deviation**

Quartile deviation of semiquartile range is defined as Q = (Q3-Q1) / 2 where Q3 is third quartile and Q1 is first quartile. This is better than range since it is utilizing 50 % of the data of the distribution.

Calculate quartile deviation for the example, 2, 4, 6, 8, 10, 12, 14, 16

First quartile is the figure below which 25% observations are there and above which are 75 % observation.

The third quartile is that figure below which 75% of data observations are and above which 25% observations are there.

If we arrange the observations in ascending order we get

2, 4, 6, 8, 10, 12, 14, 16

N= Number of observations = 8

The figure above which 50 % observations are there and below which 50 % observations, these are median and it lies in between 8 and 10 and it is $\frac{8+10}{2}$ = 9

The first quartile is below which 25 % and above which 75 % observation are is the figure which will lie in between 4 and 6 i.e. $\frac{4+6}{2}$ = 5

The third quartile is the figure below which 75 % observations are there and 25% observations are there i.e. it will lie in between 12 & 14 i.e. $\frac{12+14}{2}$ = 13.

Thus $Q_1 = 5$, $Q_2 = 9$ and $Q_3 = 13$

Quartile Deviation = $\frac{Q3-Q1}{2} = \frac{13-5}{2} = 4$

***Coefficient of Quartile Deviation***

It is defined as the ratio of quartile deviation to the mean of first and third quartile.

Coefficient of Q.D. = $\frac{(Q3 - Q1)/2}{(Q3 + Q1)/2} \times 100$

$$= \frac{(13-5)/2}{(13+5)/2} \times 100 = \frac{(8)/2}{(18)/2} \times 100 = \frac{4}{9} \times 100 = 44.44$$

**(3) Mean Deviation**

Mean deviation from the average is the amount of spread or scatter of the items in a series from the average value ignoring the signs of deviation.

Thus mean deviation from the average is the arithmetic mean of absolute deviations from the average value. The average value can be arithmetic mean, median, mode or weighted mean. The absolute deviations imply the deviation, which ignores the signs or directions and considers only magnitude (positive or zero)

***Example :*** Calculate the mean deviation of 11, 9, 10, 25, 25 from the Arithmetic mean. The arithmetic mean (AM) of the series

$$AM = \frac{11+9+10+25+25}{5} = \frac{80}{5} = 16$$

The mean deviation from AM

$$\frac{|11-16| + |9-16| + |10-16| + |25-16| + |25-16|}{5}$$

the mean deviation from AM

$$= \frac{5+7+6+9+9}{5} = \frac{36}{5} = 7.2$$

Thus mean deviation from arithmetic mean is $\frac{\sum|X - \overline{X}|}{n}$

where $\overline{X}$ = AM; n = number of items; and X = individual observation

***Mean Deviation from Median***

For the series 11, 9, 10, 25, 25, the median can be located i.e. 9, 10, 11, 25, 25

Thus median is = 11

Therefore mean deviation from median is = $\frac{\sum|X - Me|}{n}$

where X = item, Me = median, n= number of observations, | | = absolute value

$$= \frac{\sum|X - Me|}{n} = \frac{|9-11| + |10-11| + |11-11| + |25-11| + |25-11|}{5}$$

$$= \frac{|2| + |-1| + |0| + |14| + |14|}{5} = \frac{2+1+0+14+14}{5} = \frac{31}{5} = 6.2$$

Thus for grouped discrete series mean deviation is given by

$$MD = \frac{\sum F|X - \overline{A}|}{\sum F}$$

$$= \frac{F_1|X_1 - \overline{A}| + F_2|X_2 - \overline{A}| - - - - + Fk|X_k - \overline{A}|}{F_1 + F_2 + F_3 - - - - + F_k}$$

Where $F_1$ is the frequency of $X_1$
$F_2$ is the frequency of $X_2$
Fk is the frequency of Xk
$\overline{A}$ = Average; ΣF = total of all frequencies

***Mean Deviation for Continuous Series***

When there are class intervals and frequencies. Mean deviation is obtained from the mid values of the class interval as $M_1$, $M_2$, ......Mk and frequencies. Put the values of M instead of X and calculate mean deviation as per the procedure explained above.

$$\text{Mean deviation} = \frac{\sum F|M - \overline{A}|}{\sum F}$$

M = Mid value of class, $\overline{A}$ = Average F= class frequency

***Example :*** Calculate mean deviation for the following data

X: 5, 3, 1, 7. X = item
F: 4, 3, 9, 4. F = Frequency

Write down the observations in systematic order.

| X | F | FX | $\lvert X-\overline{X}\rvert$ | $F\lvert X-\overline{X}\rvert$ |
|---|---|---|---|---|
| 1 | 9 | 9 | \|1-3.31 \| = 1-2.31 \| = 2.31 | 9x 2.31 |
| 3 | 3 | 9 | \| 3-3.31\| = 1-0.31 = 0.31 | 3x 0.31 |
| 5 | 4 | 20 | \| 5-3.31\| = 1-69 = 1.69 | 4x 1.7 |
| 7 | 4 | 28 | \| 7-3.31\| = 3-69 = 3.69 | 4x 3.7 |
| Total | ΣF 20 | 66 | | 43.2 |

$$\overline{X} = AM = \frac{\sum FX}{\sum F} = \frac{66}{20} = 3.3$$

$$MD = \frac{\sum F\lvert X-\overline{X}\rvert}{\sum F} = \frac{43.20}{20} = 2.16$$

To calculate mean deviation from median we get

| X | F | CF |
|---|---|---|
| 1 | 9 | 9 |
| 3 | 3 | 12 |
| 5 | 4 | 16 |
| 7 | 4 | 20 |
| | ΣF = 20 | |

N= ΣF = 20; median number = N/ 2 = 20 / 2 = 10

Thus the item Number 10 is 3.

Hence the item Number 10 = 3 is median

| X | F | F\|X – Me\| |
|---|---|---|
| 1 | 9 | 9\|1-3\| = 18 |
| 3 | 3 | 3\|3-3\| = 0 |
| 5 | 4 | 4\|5-3\| = 8 |
| 7 | 4 | 4\|7-3\| = 16 |
| | ΣF = 20 | ΣF \|X – Me \| = 42 |

$$MD = \frac{\sum F\lvert X-Me\rvert}{\sum F}$$

$$= \frac{42}{20} = 2.10$$

***Example :*** Calculate mean deviation from the Arithmetic mean for following data.

| Class | Frequency |
|---|---|
| 1-3 | 4 |
| 3-5 | 6 |
| 5-7 | 8 |
| 7-9 | 2 |

***Solution :*** To calculate mean deviation from arithmetic mean, arrange class intervals in systematic order (ascending order).

| Class | Mid value of class | Frequency F | FX |
|---|---|---|---|
| 1-3 | (1+3) / 2 = 2 | 4 | 8 |
| 3-5 | 4 | 6 | 24 |
| 5-7 | 6 | 8 | 48 |
| 7-6 | 8 | 2 | 16 |
| | Total | $\Sigma F = 20$ | $\Sigma F X = 96$ |

$$AM = \frac{\sum FX}{\sum F} = \frac{96}{20} = 4.8$$

| Mid value X = M | F | F \|M – AM\| |
|---|---|---|
| 2 | 4 | 4 \| 2 - 4.8 \| = 11.2 |
| 4 | 6 | 6 \| 4 - 4.8 \| = 4.8 |
| 6 | 8 | 8 \| 6 - 4.8 \| = 9.6 |
| 8 | 2 | 2 \| 8 - 4.8 \| = 6.4 |
| | $\Sigma F = 20$ | $\Sigma F\|M - AM\| = 32$ |

$$MD = \frac{\sum F|M - AM|}{\sum F} = \frac{32}{20} = 1.6$$

AM= Arithmetic mean

M= mid value of class

Mean deviation from median

| Class | F | CF |
|---|---|---|
| 1-3 | 4 | 4 |
| 3-5 | 6 | 10 |
| 5-7 | 8 | 18 |
| 7-6 | 2 | 20 |

N= ΣF= 20

N / 2 = 20 / 2 = median number median number = 10

The item number 10 lies in the class 3-5

C = cummulative frequency of just preceeding class to median class = 4

Hence median class = 3-5, with of the class = 5-3 = 2 (i)

$$Me = L + \frac{(N/2 - C)\times i}{F} = 3 + \frac{(10-4)\times 2}{6} = 3+2 = 5$$

F= frequency of media class = 6

Lower limit (L) = 3 of median class

| M | F | \|M-Me\| XF |
|---|---|---|
| 2 | 4 | 4\|2-5\| = 12 |
| 4 | 6 | 6\|4-5\| = 6 |
| 6 | 8 | 8\|6-5\| = 8 |
| 8 | 2 | 2\|8-5\| = 6 |
| | ΣF = 20 | ΣF \| M – Me \| = 32 |

$$MD = \frac{\sum F|M - Me|}{\sum F}$$

$$= \frac{32}{20} = 1.6$$

***Merits, Demerits and Uses***

*Merits*

1. It is easy to calculate and easy to comprehend.
2. It utilizes all the given data.
3. Absolute deviations are used for calculating mean deviation, which ignores signs of deviations. Thus the irregularities are minimized by absolute deviations.
4. It is least affected by extreme observation as compared to the standard deviation
5. It is a better measure of dispersion for comparison about the formation of different distributions.

*Demerits*

1. It ignores the signs of deviations. It does not give importance to the direction of observation, which is mathematically illogical.
2. It is not a satisfactory measure when it is computed from mode and when we deal with skewed distribution.
3. It is rarely used in sociological studies
4. It is not possible to compute from open end classes.
5. The value of mean deviation increases un proportionately as the sample size increase.

*Uses*

λ Mean deviations is useful in business economics and statistics. It is used in computing forecasting businen cycle. According to National Bureau Of Economic Research ; the mean deviation is the most practical measure for the purpose of studies relating to businen cycles and forecasting.

**(4) Standard Deviation**

**It is defined as positive square root of the arithmetic mean of squares of deviations of individual values from their arithmetic mean.** If $X_1$, $X_2$, — $X_N$ are N number of items then the standard deviation SD is defined as given below and is denoted by $\sigma$

$$SD = \sigma = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

Where N = Total number of observations, X = Observations

$\overline{X}$ = Arithmetic mean = ($\Sigma X$ / N) and $X - \overline{X}$ = deviation from arithmetic mean

SD = Standard Deviation = $\sigma$

***Example :***

Calculate standard deviation for 1,2,3,5,4

X : 1,2,3,5,4

$\Sigma X$= 1+2+3+5+4 = 15, N = 5

$$\overline{X} = \frac{\sum X}{N} = \frac{15}{5} = 3$$

$$SD = \sigma = \sqrt{\frac{(X-\overline{X})^2}{N}} = \sqrt{\frac{(1-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2}{5}}$$

$$= \sqrt{\frac{4+1+0+1+4}{5}} = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.4142$$

***(i) Standard Deviation for Grouped Data***

If $X_1$, $X_2$, ..... Xk are items and $F_1$, $F_2$, ...... Fk are the frequency of each item respectively then.

$$SD = \sqrt{\frac{\sum F(X-\overline{X})^2}{\sum F}} = \sqrt{\frac{F_1(X_1-\overline{X})^2+F_2(X_2-\overline{X})^2 ----+F_k(X_k-\overline{X})^2}{F_1+F_2 ---+F_k}}$$

where $\overline{X} = \frac{\sum FX}{\sum F}$ = AM

***Example :*** Calculate SD for the given data

| | | | | |
|---|---|---|---|---|
| F: | 5 | 4 | 3 | 8 |
| X: | 4 | 6 | 4 | 3 |

F= frequency ; X = item or observation

There $\Sigma FX = \frac{5\times4+4\times6+3\times4+8\times3}{5+4+3+8} = \frac{20+24+12+24}{20} = \frac{80}{20} = 4$

$$SD = \sqrt{\frac{5(4-4)^2+4(6-4)^2+3(4-4)^2+8(3-4)^2}{20}}$$

$$= \sqrt{\frac{0+8+0+8}{20}} = \sqrt{\frac{16}{20}} = \sqrt{\frac{4}{5}} = \sqrt{0.8} = 0.8944$$

***(ii) Standard Deviation for Grouped Data Continues Series***

If the data is given with class internal and frequencies then the same formula as given above is used by replacing X with mid value ( M ). If $X_0$- $X_1$,$X_1$-$X_2$, $X_{k-1}$- Xk are k classes and if $F_1$, $F_2$, ......... Fk are the frequencies respectively for each class then.

$$SD = \sqrt{\frac{\sum F(X-m)^2}{\sum F}} \qquad \text{Where mi} = \frac{X_{i-1}+X_i}{2} \qquad i = 1,2,.......,k$$

***Example :*** Calculate the standard deviation for continuous series

| CI | F | Mid value = m | Fm |
|---|---|---|---|
| 1-3 | 2 | (1+3) / 2 = 2 | 4 |
| 3-5 | 4 | (3+5) / 2 = 4 | 16 |
| 5-7 | 8 | 6 | 48 |
| 7-9 | 3 | 8 | 24 |
| 9-11 | 1 | 10 | 10 |
| | | | 102 |

AM = 102 / 20 = 5.1

$$SD = \sigma = \sqrt{\frac{\sum F(m-\overline{X})^2}{\sum F}}$$

$$= \sqrt{\frac{2(2-5.1)^2+4(4-5.1)^2+8(6-5.1)^2+3(8-5.1)^2+1(10-5.1)^2}{2+4+8+3+1}}$$

$$= \sqrt{\frac{2\times(3-1^2)+4\times(1-1^2)+8\times(0-9)^2+3\times(2-9)^2+1\times(4-9)^2}{2+4+8+3+1}}$$

$$= \sqrt{\frac{19.22+7.84+6.48+25+23+24.01}{20}} = \frac{79.78}{20}$$

$$= \sqrt{3.989} \quad = 1.9972$$

## *(iii) Variance*

If $X_1$, $X_2$ ....... $X_n$ are n numbers of items then variance of the series in defined as

Var (X) = 1/n $\Sigma(X-\overline{X})^2$

Where $X$ = item

$\overline{X}$ = mean

$\Sigma$ = sum over n items

Alternatively the same can be written as

$$\text{Var}(X) = \sigma^2 \qquad = \frac{1}{n}\Sigma(X-\overline{X})^2 = \frac{1}{n}\left[\Sigma x^2 - n\overline{X}^2\right] = \frac{1}{n}\left[\Sigma x^2 - (\Sigma x)^2/n\right]$$

$\Sigma X$ = total of n observations

$\Sigma X^2$ = total of squares of n item

n = total number of items

***(i) Variance for Discrete Series Grouped Data***

When $X_1, X_2 .... X_k$ are k items having frequencies $F_1, F_2, ..... F_k$ respectively then

$$\text{Var}(X) = \frac{\sum_{i=1}^{K} F_i (X_i - \overline{X})^2}{\sum Fi}$$

If $\Sigma F = N$ then

$$\text{Var}(X) = \frac{\sum F_i (X_i - \overline{X})^2}{N}$$

Alternatively the same formula can be written as

$$\text{Var}(X) = \frac{1}{N}\left[\sum FX^2 - \frac{(\sum FX)^2}{N}\right] \text{ Or } \quad \text{Var}(X) = \frac{1}{N}\left[\sum FX2 - N\overline{X}^2\right]$$

$N = \Sigma F$; $\overline{X} = (\Sigma FX / \Sigma F)$

***(ii) Variance for Grouped Data (continuous series)***

When the data is in class intervals and frequencies. Instead of X, here m is used, where m is midpoint of each class interval.

Variance (X) = $\sigma^2 = \Sigma$ Fi (mi $- \overline{X})^2 / \Sigma F$

mi = mid point of item class interval i = 1,2,3 … k

$$\Sigma F = N \qquad \overline{X} = \frac{\sum Fm}{\sum F} \qquad \overline{X} = \text{mean} = \Sigma Fm / \Sigma F$$

Variance is always non-negative i.e. either it is positive or zero when all the items are alike variance is zero. Positive square root of variance is standard deviation of the series.

**(5) Coefficient of Variation (C.V.)**

**It is defined as the ratio of standard deviation to the arithmetic mean of the series in percentages. The coefficient of variation is the standard deviation expressed in terms of percentage of arithmetic mean.**

***Example :*** If the item of series are x = 1,2,3,4 and 5 then variance (X) = 10/5

Standard Deviation SD = $\sqrt{\frac{10}{5}} = \sqrt{2} = 1.4142$

AM = 3, then Coefficient of Variation (C.V.) = $\frac{1.4142 \times 100}{3} = 47.14$

***Merits and Demerits of Standard Deviation***

1. Of all the measures studies, standard deviation is the best measure it is easy to understand and easy to calculate
2. It is based on all observations
3. It is least affected by the fluctuations of sampling
4. It can be used for further mathematical treatment
5. However it is affected by extreme observation to less extent and less used by business statisticians as they are interested in the results of modal class.
6. This is a powerful measure and is used in calculating regression, crrelation and sampling theory.

## 13.2.3 Correlation and Regression

The measures of central tendencies, measures of dispersion, Skewness and kurtosis can be used for the purpose of analysis for a single variable only. However there can be two variables or more than two variables which are related with each other or their interrelation may affect each other. If we study the two variables height and weights of same individuals; we shall observe that there is some relationship between these two variables. Similarly the phenomena of marks in Statistics and marks in Mathematics of the same individuals can show some relationship. Thus one individual person has two observations of height and weight or marks in Statistics and Mathematics. Thus a person is a unit and this unit has two values; one relating to height and other relating to weight. Such distributions are known as bivariate distributions. Thus bivariate distribution is a distribution in which a unit (or individual) of series has measurements or values on two characteristics or variables.

If more than two variables are observed for individuals in the population, it is called as Multivariate distribution. If we take an example of a farmer, sowing a crop we can have land, labor, seed, fertilizer and production obtained from the crop as different values for a single farmer and if we take another farmer we shall have a set of different values for another farmer.

In bivariate analysis researcher might be interested to know the type of relationship between two variables. And when we wanted to establish a relationship between two variables; correlation is a statistical tool, which measures the degree of relationship between two variables.

The correlation is thus an extent of relationship between two variables and it measures the association between two variables, if the variables are linearly related. The correlation analysis is the study of methods and techniques for measuring extent or degree of relationship between two variables.

When the measurement is of quantitative nature, the correlation is an appropriate measure or statistical tool to express the degree of relationship. The correlation is an analysis of co-ordination between two or more variables. The two variables are said to be correlated if the change in one variable results in a corresponding change in the other variable. The correlation between two variables is defined as a measure of extent of linear association between two variables.

When the two variables change in the same direction, we can expect positive correlation between two variables. If the change in two variables occurs in opposite directions of each other, then there is negative correlation between two variables. For example if we take height and weight as two variables, we may expect positive correlation between these two variables. If we take two variables such as pressure and volume we may expect negative relationship between two variables.

The examples of positive correlation can be income and expenditure of family amount of rainfall and production of a crop. Price and supply of commodity. Similarly the examples of negative correlation can be use of pesticides and population of insects, temperature and woolen garments, price and demand of commodity, in crease in education and decrease in doubts and ignorance in adoption of new technology.

**(a) Scatter Diagram**

Scatter diagram is the diagrammatic representation of a bivariate distribution and provides a tool for the ascertaining the nature of correlation between two variables. Let $(X_1,Y_1)$, $(X_2, Y_2)$, —— $(X_n, Y_n)$ be 'n' pairs of two variables X and Y. Let X be the height and Y be the weight of the individuals.

Thus there are 'n' pairs of X and Y which will represent 'n' points on a graph paper. If the X axis of the graph paper represents X variable and Y axis of the graph represents Y variable, we shall get 'n' points of $(X_i, Y_i)$. The diagram of points is known as scatter diagram. From a scatter diagram, we may get fairly good idea about the relationship between two variables.

The following conclusions can be drawn from the scatter diagram.

1. If we get fairly dense points or the points are very close to each other, a fairly good amount of correlation may be expected between the two variables. If the points are widely scattered a poor amount of correlation may be expected between them.
2. If the points on the scatter diagram reveal either upward or downward trend, then the variables are said to be positively or negatively correlated. If there is no trend in the direction upward or downward then the two variables are said to be uncorrected with each other.
3. Thus if there is rising trend from lower left hand corner and going upward to the upper right hand corner the correlation is positive. On the other hand if the trend of the points depicts a downward line from the upper left hand corner to the lower right hand corner, the correlation is negative.
4. If all the points lie on a straight line starting from the left bottom and going up towards the right top; the correlation is perfect and positive and if all the points lie on a straight line starting from left top and coming down to right bottom; the correlation is perfect and negative.
5. If the shape of the curve drawn is circular in nature or 'U' type or does not depict any particular trend then the two variables are said to be uncorrelated.

*(ii) Karl Pearson's Correlation Coefficient*

**The Karl Pearson's correlation coefficient is defined as the ratio of covariance between two variables (X,Y) to the product of the standard deviations of X and Y**

$$r\,(X,Y) = \text{Correlation Coefficient} = \frac{\text{Cov}\,(X, Y)}{SD(X)\ SD(Y)}$$

$\rho\,(X,Y)$ = Population Correlation Coefficient between X & Y

Generally we deal with samples and the correlation between two variables X and Y is denoted by r (X,Y) or $r_{x\,y}$ or simply r and is the measure of relationship between two variables X and Y for samples. Thus if $(X_1,Y_1), (X_2,Y_2) \ldots (X_n, Y_n)$ are n number of pairs of X an Y then correlation (X, Y) can be given as :

$$r\,(X,\,Y) = \frac{\text{Cov}\,(X,Y)}{Sd(X)\,.\,Sd\,(Y)}$$

$$\text{Cov}\,(X,Y) = 1 / n\, \Sigma(X - \overline{X})(Y - \overline{Y})$$

$$Sd(X) \quad = \sqrt{1/n \sum (X-\overline{X})^2}$$

$$Sd(Y) \quad = \sqrt{1/n \sum (Y-\overline{Y})^2}$$

$$r(X,Y) = r = \frac{1/n \sum (X-\overline{X})(Y-\overline{Y})}{1/n \sqrt{\sum (X-\overline{X})^2 \sum (Y-\overline{Y})^2}}$$

$$= \frac{\sum Dx\,.\,Dy}{\sqrt{\sum Dx^2 . \sum Dy^2}}$$

Where $\overline{X}$ = AM of X = ΣX / n

$\overline{Y}$ = AM of Y = ΣY / n AM = Arithmetic Mean

Dx = $X - \overline{X}$ = deviation of X from its mean ($\overline{X}$)

Dy = $Y - \overline{Y}$ = deviation of y from its mean ($\overline{Y}$)

n = number of pairs of ( X and Y )

***Example :*** Calculate the correlation coefficient for the data of no of branches per plant (X) and total number of flowers per plan (Y)

| | | | | |
|---|---|---|---|---|
| X | 2 | 5 | 6 | 3 |
| Y | 4 | 8 | 12 | 8 |

***Method I***

*The steps in obtaining Correlation Coefficient.*

1. Write down the data in the tabular form as given below. Count the number of pairs. Call these pairs as N. N = 4 in present case.

| Sr. No | x | y | $x^2$ | $y^2$ | xy | |
|---|---|---|---|---|---|---|
| 1 | 2 | 4 | 4 | 16 | 8 | N=4 |
| 2 | 5 | 8 | 25 | 64 | 40 | $\Sigma x = 16$ |
| 3 | 6 | 12 | 36 | 144 | 72 | $\Sigma y = 32$ |
| 4 | 3 | 8 | 9 | 64 | 24 | $\Sigma x^2 = 74$ |
| 5 | 16 | 32 | 74 | 288 | 144 | $\Sigma y^2 = 288$ |
| N | $\Sigma x$ | $\Sigma y$ | $\Sigma x^2$ | $\Sigma y^2$ | $\Sigma xy$ | $\Sigma xy = 144$ |

2. Obtain the squares of each individual variable X and Y
3. Obtain the product of individual X & Y
4. Find out their totals
5. Use the following formula to calculate correlation coefficient.

$$\text{Cor}(x, y) = \frac{\sum xy - \dfrac{(\sum x)(\sum y)}{N}}{\sqrt{\left[\sum x^2 - \dfrac{(\sum x)^2}{N}\right]\left[\sum y^2 - \dfrac{(\sum y)^2}{N}\right]}} \qquad \text{......(A)}$$

$$\Sigma xy - \frac{(\sum x)\,(\sum y)}{N} = 144 - \frac{16 \times 32}{4} = 144 - 128 = 16$$

$$\Sigma x^2 - \frac{(\sum x)^2}{N} = 74 - \frac{16^2}{4} = 74 - 64 = 10$$

$$\Sigma y^2 - \frac{(\sum y)^2}{N} = 288 - \frac{32^2}{4} = 288 - 256 = 32$$

By putting these values in the formula ….. (A)

$$\text{Cor}\,(x, y) = \frac{16}{\sqrt{10 \times 32}} = \frac{16}{\sqrt{320}} = \frac{16}{17.8883} = 0.8944$$

***Method II***

1. Calculate mean of x and mean y
2. Obtain deviation of x from $\bar{x}$, i.e. ( x - $\bar{x}$), and deviation of y from y i.e. (y – $\bar{y}$).
3. Calculate the product of (x - $\bar{x}$) and (y - $\bar{y}$)
4. Calculate the total of $(x - \bar{x})^2$ and $(y - \bar{y})^2$
5. Use the following formula and calculate Correlation Coefficient.

$$r(x,y) = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 \sum (y-\bar{y})^2}}$$

$$\Sigma x = 16, \qquad \bar{x} = \frac{\Sigma x}{N} = \frac{16}{4} = 4$$

$$\Sigma y = 32, \qquad \bar{y} = \frac{\Sigma y}{N} = \frac{32}{4} = 8$$

| X | Y | $(x-\bar{x})$ | $(y-\bar{y})$ | $\Sigma(x-\bar{x})^2$ | $(x+y)^2$ | $(x-\bar{x})(y-\bar{y})$ |
|---|---|---|---|---|---|---|
| 2 | 4 | 2-4 = -2 | 4-8=-4 | 4 | 16 | 8 |
| 5 | 8 | 5-4 = +1 | 8-8 = 0 | 1 | 0 | 0 |
| 6 | 12 | 6-4=+2 | 12-8 | 4 | 16 | 8 |
| 3 | 8 | 3-4= -1 | 8-8=0 | 1 | 0 | 0 |
| $\Sigma x=16$ | $\Sigma y=32$ | $\Sigma(x-\bar{x})=0$ | $\Sigma(y-\bar{y})=0$ | $\Sigma(x-\bar{x})^2=10$ | $\Sigma(y-\bar{y})^2=32$ | $\Sigma(x-\bar{x})^2(y-\bar{y})^2=16$ |

$$= r(x,y) = \frac{16}{\sqrt{10 \times 32}} = 0.8944$$

## 13.2.4 Non Parametric Tests

There are two types of statistical tests one in which certain stringent conditions regarding population distributions and another, in which the conditions are less stringent regarding the population distribution.

The statistical tests are known as parametric statistical tests in which there are stronger assumptions regarding the population distribution such as variable should have normal distribution; there should be homogemetry of variances. The samples under consideration should be independently identically and randomly selected from the population.

The statistical tests are known as non parametric statistical test or distribution free statistical test in which no conditions are laid down regarding the population distribution. The model does not specify the condition on the population distribution and population parameters. The measurement required for the observation in weaker as compared to parametric tests.

The only requirement of non parametric statistical test is that the observation should be independent and the variable should have the observations which can be either in nominal or ordinal scale. There are some non parametric statistical tests which are alternative to parametric tests which require stringent assumptions.

The Binomial test, the one sample $\chi^2$ test are alternative tests for one sample 't' test. The sign test, Wilkoxon matched pairs rank test are alternative non parametric statistical tests for testing equality of two paired sample averages instead of paired 't' tests.

The median test, Fisher's exact probability test, the Mann-Whitney 'U' test are alternative non parametric tests for 't' test with independent (unpaired) samples.

## 13.3 Glossary

**Average** : A typical value from the set of data around which most of the observation are clustersed.

**Arithmetic Mean :** The sum of observations to the total number

**Geometric Mean :** The 'n' th root of product of 'n' number of items

**Harmonic Mean :** Reciprocal of the arithmetic mean of reciprocals

**Median :** The mid item of the series when the items are arranged in ascending or descending order of magnitude.

**Mode :** Mode is the common most item of the series.

**Dispersion :** The spread of observation from the individual value and from the average value

**Mean Deviation :** The arithmetic mean of absolute deviations from the average

**Variance :** The arithmetic mean of squared deviations from the arithmetic mean.

**Standard Deviation :** Positive square root of variance

**Correlation coefficient :** The measure of linear association between two variables.

**Coefficient of variation :** It is the standard deviation expressed as percentage of arithmetic mean.

## 13.4 Summary

The data eitehr qualitative or quantitative is to be separated in the form of calsssification, tabulation and frequency distribution. The data are clustered is called as Measures of Central Tendancy. The various measures of central tedndancy are arithmatic mean, geometric mean, hormonic mean, median, mode and weighted mean. The measures of central tendancies, measures of dispersion, skewness and kurtosis can be used for the purpose of analysis for a single variable only. However for more than two variables, such distribution are known as bivariate distribution which is also known as Correlation. There are two types of statistical test i.e. Parametric tests and Non-parametric tests which are mainly used in Agriculture and Social Sciences.

## 13.5 Exercises for Practice

1. What is average ? Write on different types of averages ?
2. What are the characteristics of an ideal average ?
3. Describe the term variability and describe different measures of variability.
4. What is mean by Scatter Diagram ? How it is useful in statistical interpretations.
5. Explain in detail Correlation Co-efficient and Regression Co-efficient.
6. State and explain non-parametric statistical tests used in Agricultural research.

# Unit 14 : Scaling Techniques

***Index***



## 14.1 Introduction

The measurement mean the assigning of numbers to observation for helping Statistical analysis. The measurement are mostly used by Social Scientists. The Sociometry means the measurement of social variables and techniques to give measurement dimensions to these variables. There are four levels of measurements i.e. Nominal scale, Ordinal scale, Interval scale, Ratio scale etc. which will be studied in detailed in this unit. There ate different scaling techniques which are scored and applied in the analysis of Social Science data.

**After the study of this unit, you will be able to know and understand :**

- λ Scales for measuring social status.
- λ Scaling techniques.
- λ Scaling scores and testing.

## 14.2 Content

### 14.2.1 Scales for Measuring Social Status

Measurement means the description of the data in terms of numbers which is useful in mathematical operations. Measurement means the assigning of numbers to observations in such a way that the numbers are amenable to statistical analysis by manipulating the operations with certain rule. The analysis will lead to new information about the objects which are used for measurement. The measurements obtained by physical scientists are capable of different mathematical and statistical operations because of the structure of numbers which is isomorphic to the numerical arithmetic. However, the measurement mostly used by a social scientist is not always as strong as the physical scientist and in many a times the measurement is weak by assigning qualitative trait or simply comparative scale.

The operations such as adding, dividing are possible in measurements of physical sciences but the same is not always possible in social sciences for the attributes such as honesty, temperament, behavior. There are certain sociological or psychological variable frequently used in social science research, sometimes the measurement of these variables is just possible by making use of the existing scales; but when new items are added or some old concepts became obsolete or required to be deleted ; the new scale of measurement is required. The sociometry means the measurement of social variables and the study of techniques to give measurement dimensions to these variables. Most of the times the mathematical operations are rarely possible; under these circumstances it is necessary to measure the trait of the variable on the scale. Attitude, honesty beauty etc. a given set of scores are dependent on the level of measurement. There are four levels of measurements these are (1) Nominal scale (2) Ordinal scale (3) Interval scale (4) Ratio scale.

**1. Nominal Scale**

When number or other symbols are used to classify an object or person according to a characteristic, it is the weakest scale, the of numbers or symbols which are used to identify groups, characteristics or category is called as nominal scale. The members of a group or class are regarded as equivalent with respect to the characteristic under consideration. e.g (i) The students in a class can be grouped into batch 1, batch 2, batch 3, etc. (ii) The set of cars can be grouped as white, red, green black colour etc.

The Classification of human beings according to race, religion, colour etc. Thus a equality of objects in some respect is the basis for putting them in the categories. Sometimes objects are placed in one category because they are indistinguishable by the prevailing methods of observations. At other times there are observable differences yet these differences might be tolerable to avoid too many categories. And therefore the objects are placed in same category. For example the vehicle licenses can be an example of this type such as MH-15, MH-20, MH-21 etc.

The nominal scale partitions the given class into mutually exclusive subclasses. The members of the subclass obey the equivalence relation. That is the members of any one subclass must be equivalent in the property being used for scaling. The equivalence relation means the relation is reflexive, symmetrical and transitive, the symbols or labeling can be interchanged without altering the information of a group. We can estimate mode, frequency counts from the nominal scale. Non parametric statistical test $\chi^2$ test , binomial test, contingency coefficient C, can also be used when the measurement is in nominal scale

**2. The Ordinal or Ranking Scale**

When the measurement utilizes the property of rank and order then the scale used in ranking or ordering the objects is known as ordinal or ranking scale. The typical relationship used in this type of measurement are, higher, less preferred, more difficult, less mature, more disturbed, warmer, louder, etc.

These relations may be designated by the symbols “>” (greater than) or the “<” (less than) to designate the higher or lower than the object. e.g in four point scale grading system of students A>B>C>D>F grade, the students are graded in this manner.

In the army the ranks of officers can be Second Lieutenant < Lieutenant < Captain < Major < Colonel. The ordinal scale has the relationship of greater than or less than in addition to equivalence. Thus the relation is asymmetrical and transitive.

The transformation does not change the order of the classes as is admissible because it does not involve loss of information. Hence the numbers applied to groups or designations can be changed in any way without altering the order of ranks. Quartile, percentile can be applied as statistical tools in such situations.

Median is an appropriate statistic to measure the average of a group. The Spearman's rank correlation, Kendall's rank correlation coefficient can be applied to know the relationship. When the measurement is in ordinal scale the parametric statistical tests should not be used, because the properties of ordinal scale are not similar to the numerical system. The distances in the two ranks might not be same which will result in calculating standard deviation and mean.

**(3) The Interval Scale**

When the measurement obeys all the characteristics of ordinal scale and in addition to that the distances between any two numbers are of known size, the measurement is known as interval scales. If there are three point A, B, and C, we can say that the distance between A to B plus the distance between B to C is equal to the distance between A to C. The interval scale has a characteristic of common and constant measurement which assigns a real number to all pairs of objects in the ordered set. For example A and B are two points which can be assigned with two numbers 10 and 20 respectively; as a part of the scale can also assume the number 50 and 60 respectively keeping the distance same during the two measurement.

In this measurement the ratio of any two intervals in independent of the unit of measurement and the of the zero point. In an interval scale the zero point and the unit of measurement are arbitrary. For examples. Take any two numbers on interval scale such as 10 and 20. The first number 10 which is 10 units from the arbitrary zero and the second is 20 units from the arbitrary zero. If we add to the two, we get the sum equal to 30 units from the zero point. Suppose the arbitrary zero is at 5 above the true zero point. Then the two values should have been 15 and 25; if we wanted their actual distances from the absolute and meaningful zero. If we add 15 and 25, we get 40 and not 30. Thus the sum varies as the position varies from the zero on the scale. Calendar time is an interval scale because the arbitrary zero time is set by convention.

Similarly the measurement of an altitude of a mountain which will vary from the sea level or location level. The other example can be the measurement of temperature such as Fahrenheit, Centigrade or Rumour. The true zero point and the unit of measurement are arbitrary. They are different for the three scales. However, if we take Fahrenheit and Centigrade both these scales furnish the same kind of information; because they are linearly related. The linear transformation of the three scales is given below.

(F-32) / 9 = (C / 5) = (R / 4) or if we want to convert centigrade into Fahrenheit we can use F = (9 C / 5) + 32

F= number of degrees on Fahrenheit scale

C= number of degrees on centigrade scale

It can be shown that the ratios of temperature differences (intervals) are independent of the unit of measurement and of the zero point for example the freezing occurs at zero degrees in centigrade and boiling occurs at 100 degrees centigrade scale. Whereas the freezing occurs at 32 degrees Fahrenheit and boiling occurs at 212 degrees in Fahrenheit scale. The other reading of two temperature scale are as given below.

| | | | | | |
|---|---|---|---|---|---|
| Centigrade | 0 | 15 | 30 | 40 | 100 |
| Fahrenheit | 32 | 50 | 86 | 104 | 212 |

Thus it can be verified from the above table that the ratio of differences between temperature readings on one scale is equal to the ratio between the equivalent differences on other scale. For example take 0, and 15 centigrade and find out the ratio of difference

We get $\frac{(30-15)}{(15-0)} = \frac{15}{15}$ = 1 Similarly for Fahrenheit we get $\frac{(86-59)}{(59-32)} = \frac{27}{27}$ = 1

Thus the ratio is the same in both cases. Thus in an interval scale the ratio of any two intervals is independent of the unit used and of the zero point or starting point, both of which are arbitrary. Most of the psychological scales are interval scales.

The information yielded by the scale is not affected by, if each number is multiplied by a positive constant and then if a constant is added to this product i.e. $f(x)$ = ax+6.

### 4. Ratio Scale

The scale which obeys all the properties of interval scale and in addition to that it has a true zero point. The zero stands for neither more nor less than none of the property represented by the scale. In a ratio scale, the ratio of any two scale points is independent of the unit of measurement For example, we measure mass or weight in a ratio scale. The scale of grams and kilograms or the scale of pounds and ounces are similar. If we determine the weights of two different object in pounds and grams we would find that the ratios of two pound weights is identical to the two gram weight. The operations and relations give rise to the numerical values and the ratio scale is isomorphic to the true structure of arithmetic. Therefore, the operations of arithmetic are permissible on the numerical values assigned to the objects.

Ratio scales are commonly used in physical sciences also. It achieves all the operations and relations such as (i) equivalence (ii) greater than (iii) known ratio of any two intervals and (iv) known ratio of any two scales values. Any statistical test is usable, when ratio measurement has been achieved. The numbers achieved by ratio scale are true numbers and true zero points and only the scale of measurements is arbitrary.

## 14. 2.2 Scaling Techniques

For the measurement of educational and psychological characteristics which are abstract in nature and where true starting point or zero is not known, the distances from the arbitrary point are taken to distinguish to tests. Many of these scales are based on the normal probability curve and it is assumed that two scales or different scales are comparable.

In brief, sociological, psychological scales are interval scales and not ratio scales because there is no absolute zero point. The commonly used scaling procedures are described below. σ scaling (Scaling Individual Test Items in Terms of a trait ).

Let us take an example to illustrate the technique. Let us suppose there are five problems A, B, C, D and E. These five problems were asked to solve to the large number of students. The response of solving the problem was 15%, 34%, 50%, 62% and 80% respectively. However the comparison of two percentages in measuring the trait ability to solve the possible is a crude method.

Because the problem solved by 80% of the individuals is comparatively easier as compared to the problem answered by 15%. These two percentages do not reflect the differences in the trait difficulty.

Suppose we wanted to construct a difficulty scale then we assume the ability or the trait being normally distributed with mean $\mu$ and standard deviation $\sigma$. Under the assumption of the normality of the trait or variability of the group provides a better scale which is known as $\sigma$ scale.

The difficulty value of an item is usually defined as the minimum ability to answer the item correctly under the assumption that the ability is distributed normally with mean 0 and variance $\sigma^2$. If $p_i$ is the proportion of the individuals answering $i^{th}$ item successfully then its difficulty value is given by $\sigma$ Zi where Zi is determined from the following relation

$$p(Z>Zi) = \frac{1}{\sqrt{2\pi}} \int_{zi}^{\infty} e^{-t^2/2} . dt = p_i$$

where Z is Zi distributed with N(0,1) (i= 1,2,…n)

For a given $p_i$'s the value of $z_i$'s can be read from the table of areas under normal probability curve. $Z_i$'s are also known as $\sigma$ distances from the mean.

Let us calculate the $\sigma$ value of each percentage explained for solving problems A, B, C, D & E.

| A | B | C | D | E |
|---|---|---|---|---|
| 15% | 34% | 50% | 62% | 80% |

$p_1 = 0.15$, $p_2 = 0.34$, $p_3 = 0.50$ $p_4 = 0.62$, $p_5 = 0.80$

The $\sigma$ values for corresponding p can be obtained by using.

$p( Z > Z_1) = 0.15 \Rightarrow p (0<Z<Z_1) = 0.35$

by referring the area table for normal distribution we get. $Z_1 = 1.04$

Similarly $p(Z>Z_2) = 0.34 \Rightarrow p (0<Z<Z_2) = 0.16 \Rightarrow Z_2 = 0.42$

$p (Z > Z_3) = 0.50 \Rightarrow p (0<Z<Z_3) = 0 \Rightarrow Z_3 = 0$

$p (Z > Z_4) = 0.62 \Rightarrow p (0<Z<Z_4)= 0.12$ (by symmetry)

$Z_4^1 = 0.31$ and $Z_4 = - Z_4^1 = -0.31$

Also p (Z > $Z_5^1$ ) = 0.80 $\Rightarrow$ p (0 < Z< $Z_5^1$ ) = - 0.30

$Z_5$ = - $Z_5^1$ = - 0.84

The required σ values will be

| Problem | A | B | C | D | E |
|---|---|---|---|---|---|
| σ distance from Mean | 1.04 | 0.42 | 0 | (-) 0.31 | (-) 0.84 |

Some problems are having negative values if we choose arbitrary zero as – 3 σ distances for A, B, C, D, and E will be from – 3 as 4.04, 3.42, 3,2.69, and 2.16 respectively. Thus three is added to each value

**1. Scaling Scores on a Test**

If a number of students have appeared for different test in English, Statistics, Physics, Psychology and General aptitude. Then usual procedure to judge the performance of a student is to sum the raw score is five, tests and rank the students on the basis of grand total of the score. Then the question arises whether we have justified making the comparison on the basis of total of raw scores? Obviously the answer is no, because, English and Statistics may require different degrees of a qualities. Hence the comparison will not be meaningful. Hence to have valid comparison between raw scores, we need a common scale which is obtained under some assumptions regarding the distribution of the trait being measured by the test, the standard scores and t scores furnish such common scale.

**2. Z or σ Scores and Z or σ Scaling**

If the distribution of raw scores (x) under test has mean μ and a standard deviation = σ there σ score or Z – score is defined by Z =( X- μ) / σ

We have E (Z) = E (X- μ) / σ = 0

= 1/σ [E(X)-μ]

and var (Z) = var (X- μ / σ ) = I / $\sigma^2$ var (X) = $\sigma^2$ / $\sigma^2$ = 1

Hence the mean of a set of σ scores is always zero and its standard deviation is unity. Accordingly, σ scale is that scale which has mean zero and standard deviation as 1 in the construction of σ the assumption as the distribution of the triad differ only in mean and standard deviation. Theoretically σ scores give valid comparison but practically they have following drawbacks.

(a) About 50% of the raw scores will lie below mean μ; hence approximately 50% of σ scores will have negative signs.

(b) Another disadvantage is the very large unit, viz. one standard deviation will make σ scores in fraction which is awkward for comparison and computation.

## 3. Standard Scores

The above two objections of σ scores can be over come by adding to σ scores a constant μ so that all of them become positive and multiplying them by another constant σ this amounts to transforming a new scale with mean μ and standard deviation σ.

This σ scores transformed to the new scores having mean μ and standard deviation σ. The σ scores transformed to the new scores having mean μ and SD = σ are called as standard scores. Thus standard score $X^1$ with mean μ and SD = σ corresponding to the raw score X with mean μ and SD = σ can be given as

$X^1 - \mu^1 / \sigma^1 = X - \mu / \sigma$ Or $X^1 = \mu^1 + \sigma^1 ( X- \mu / \sigma)$

$X^1 = \mu^1 + \sigma^1 Z$

Where $Z = X- \mu / \sigma$ is the σ score corresponding to X

Or $X^1 = (\sigma^1 / \sigma) x - [ (\sigma^1 / \sigma) \mu - \mu^1]$

## 4. Normalized Score

If p is the proportion of the individuals getting scores below X then the number ξ given by *e*

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\xi} e^{-(t^2/2)} .dt = p \qquad \phi(\xi) = p$$

Where Ø (ξ) is the distribution function of standard normal variate is called as normalized score corresponding to x. Like σ scores normalized scores are having their mean zero and variance unity and they lie almost in the range of – 3 to + 3.

For practical convenience normalized scores are transformed to new scale with mean μ and standard deviation σ by relation

$\eta' - \mu \;/ \sigma = \xi$, $\eta' = \sigma\xi + \mu$ ……………………………………..(A)

where $\mu$ and $\sigma$ are pre assigned $\eta'$ are called as $\sigma$ normalized standard scores.

## 14.2.3 T Scores

This score was devised by Mc. Call William and are named as T score in the memory of two psychologists **Terman and Thorndyke Scores.**

If we put $\sigma = 10$ and $\mu = 50$ in (A) of normalized scores, we get T score

The T scores are normalized standard scores converted into a distribution with mean = 50 and standard deviation 10. $\therefore$ T = 50 + 10 $\xi$

*Use following steps to calculate T score*

1. Arrange the test scores in descending order of magnitude
2. Obtain cummulative frequency from the bottom of the distribution
3. Obtain the cummulative frequency (CF) below the mid value of each class interval under the assumption that the frequencies are uniformly distributed over the class intervals (By subtracting 1 / 2 fi from each class respectively)
4. Obtain the cummulative frequency (CF) below the mid-value of each class interval under the assumption that the frequencies are uniformly distributed over the class intervals
5. Obtain the normalized scores $\xi$ given by $\int_{-\infty}^{\xi} \left(1 / \sqrt{2\pi}\right) \text{-}\, e^{\frac{1}{2}u^2} du = p$
6. Finally T score are obtained from normalized scores $\xi$ 's on using normal curve

Hence E. Garret has given in following table for obtaining T scores directly from the percentages or proportions $\rho$ in column number 5. However the steps outlined above can be elegantly explained by making use of following table.

Calculation of T scores for a given frequency distribution :

*The procedure for obtaining T scores is as given below*

1. Arrange the test scores in descending order of magnitude
2. Obtain cummulative frequency from the bottom of the distribution
3. Obtain the C.F. below the mid value of each class interval, assuming that the frequencies are uniformly distributed over the class intervals.

4. Express these cummulative frequencies or proportions ‘ρ’ of the total frequency N

5. Obtain the normalized scores “ξ” given by $\Sigma \int_{-\infty}^{\xi} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du = \rho$

6. Finally T scores are obtained from normalized ξ’s by using T= 50 + 10ξ

The steps explained by E Garret is given below :

**E Garret Table**

| Class interval | F | CF | C.F. below mid value of each class | Column 4 as proportion ρ of N | ξ | |
|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| $X_{1-}X_2$ | $F_1$ | N | N= ½ $F_1$ = $A_1$ | $A_1$ / N | | |
| $X_2$-$X_3$ | $F_2$ | $\sum_{i=2}^{n} Fi$ | $\sum_{i=2}^{n} fi - \frac{1}{2}\ Fn = A2$ | $A_2$ / N | $\int_{-\infty}^{\Xi} \frac{e^{-\frac{1}{2}\mu 2}\ \delta^{\mu}}{\sqrt{2\pi}} = \rho$ | T= 50+10ξ |
| $X_{n-1}$-$X_n$ | $F_n$ | $F_n$ - $F_{n-1}$ | | | | |
| $X_n$-$X_{n+1}$ | $F_n$ | | Fn – ½ $F_n$ = An | $A_n$ / N | | |

***Example***

| Class | F | Midvalue | F |
|---|---|---|---|
| 0.5-1.5 | 5 | 1 | 5 |
| 1.5-2.5 | 10 | 2 | 10 |
| 2.5-3.5 | 20 | 3 | 20 |
| 3.5-4.5 | 5 | 4 | 5 |
| 4.5-5.5 | 4 | 5 | 4 |
| 5.5-6.5 | 4 | 6 | 4 |
| 6.5-7.5 | 2 | 7 | 2 |
| 7.5-8.5 | | | |

***Solution :*** Arrange the scores in descending order of magnitude

| X | F | CF | C.F. below mid value of each score | ρ = col 4/N proportion ρ of N | ξ | T=50+10ξ |
|---|---|---|---|---|---|---|
| 7 | 2 | 50 | 50 - (2 / 2) = 49 | 49 / 50= 0.98 | 2.5 | 75 |
| 6 | 4 | 48 | 48 – (4 / 2) = 46 | 46 / 50 = 0.92 | 1.45 | 64 |
| 5 | 4 | 44 | 44 – (4 / 2) = 42 | 42 / 50 = 0.84 | 0.995 | 60 |
| 4 | 5 | 40 | 40 – (5/2) = 37.5 | 37.5 / 50 = 0.75 | 0.675 | 57 |
| 3 | 20 | 35 | 35–(20 / 2) = 25 | 25 / 50 = 0.50 | 0 | 50 |
| 2 | 10 | 15 | 15–(10 / 2) = 10 | 10 / 50 = 0.20 | -0.84 | 42 |
| 1 | 5 | 5 | 5-(5 / 2) = 2.5 | 2.5 / 50 = 0.5 | -1.64 | 34 |

Normalized ξ are obtained from normal area curve.

## 14.2.4 Likert Scale

Let there are N number of persons and they are rated by different judges with respect to some stimulus or trait. Let the trait be honesty or beauty or acting. The judges have given their ranks and the frequency distribution of the ratings is given to you or it is known to us. Then the problem is "can we assign weightage or numerical scores (values) to them so as to make them comparable from judge to judge ? the answer to this problem is yes, provided the following assumptions are true".

(i) The distribution of the trait say honesty has normal distribution

(ii) The judges are equally competent.

Let us assume the distribution of the trait (say honesty x) N (0,1) as normally distributed let the individuals with trait values in the interval ($X_1$-$X_2$) are given rating A by judge. The values corresponding to this rating "A" is defined to be the average trait value of all these individuals and is accordingly given by the formula by **Likert.**

$$\text{Scale value} = \frac{\int_{X_1}^{X_2} U\Phi(u)du}{\int_{X_1}^{X_2} U\phi(u)du} = \frac{\int_{X_1}^{X_2} U\Phi(u)du}{\phi(X_2) - \phi(X_1)}$$

Where Ø(u) = probability density function of standard normal variate

$\phi(u)$ = distribution functions of standard normal variate

$$\text{Scale value} = \frac{\left|-1/e^{-(1/2)\mu 2}/\sqrt{2\mu}\right|_{X_1}^{X_2}}{\Phi(X_2)-\Phi(X_1)} = \frac{\phi(X_2)-\phi(X_1)}{\Phi(X_2)-\Phi(X_1)}$$

By using normal probability table for area and ordinates the values of the number can be estimated. The denominator forgive the proportion of individuals for the ratings “A”.

The numerical score for the rating “A” can be obtained by using Likert Scale formula and the origin of the scale values can be shifted to (-3) to avoid fractions the score values can be multiplied by 10 and rounded to the nearest integer.

### 14.2.5 Reliability of Test

A test is like a measuring instrument. The important characteristic of any instrument or evaluation device is how reliably it measures. In simple terms reliability means consistency. If the instrument is reliable it should give consistent results. Thus a reliable instrument will always give trustworthy and stable results if it is applied to the same individuals or objects from time to time.

Similarly a reliable test is one which when applied to same objects persons on different occasions yields stable and trustworthy results, relatively free from the errors of measurement. For example, if an individual in an intelligence test obtain a raw score 90, we would expect to find the same or nearly the same score, if we test him after 2-3 weeks. On the other hand scores made on unreliable test are subject to larger errors of measurement. These scales are not trustworthy and stable.

An unreliable test will give you inconsistent and erroneous results when the same is repeated again. Thus reliable test should give consistent results when it is repeated over a span of time provided the trait of measurement it self is not changed in the mean time.

Statistically the raw score obtained can be written in a linear model is given as follows :

$X_R = X_T + X_E$

Where $X_R$ = raw score obtained or measured, $X_T$ = true score

$X_E$ = the error score, $X_E = X_R - X_T$

$X_T$ is the genuine value of the trait being measured. This is the value we expect on using a perfect instrument under ideal conditions. A true score can not be determined experimentally. However, it can be defined as the mean of a very large number of determinations given under identical conditions.

Xe the error score is the part which is attributed to such factors as temporary characteristics of an individual such as health fatigue, emotional upset, differences in motivations, etc. which are the factors beyond human control.

It is assumed that error components are independent and randomly distributed with zero mean. This implies that for certain observations it is positive and for some other observation it is negative. It is uncorrelated with the true value therefore,

$$\text{Var}(X_R) = \text{Var}(X_T) + \text{Var}(X_E) \quad \text{Or} \quad \frac{\text{Var}(X_R)}{\text{Var}(X_R)} = \frac{\text{Var}(X_T)}{\text{Var}(X_R)} + \frac{\text{Var}(X_E)}{\text{Var}(X_R)}$$

$$1 = \frac{S_T^2}{S_R^2} + \frac{S_E^2}{S_R^2}$$

Where $S_T^2$ = variance for true score

$S_E^2$ = variance for error score

$S_R^2$ = variance for raw score

Thus reliability can be defined as the part of the variance which is true variance coefficient of reliability $(r_{iR}) = \frac{S_T^2}{S_R^2}$ i.e $(r_{iR}) = 1 - \frac{S_E^2}{S_R^2}$

Thus the coefficient of reliability varies in the range of 0 to 1 i.e $0 \le r_{iR} \le 1$ Thus the reliability of the test depends upon the variance of error scores. If the variance of error scores is larger the reliability is smaller.

The reliability can be increased by decreasing error variance Thus the important factors affecting reliability are (i) Length of the test (ii) Range of talent (iii) Ability levels of the subjects (iv) Testing conditions.

## Method of Determining Test Reliability

There are four methods for estimating the reliability tests.

**(1) The Test – Retest Method**

This method consists in submitting a group of individuals to a particular test and compiling their respective scores. After some time the same test is repeated on same candidates, and their scores are noted again. The two series of scores are arranged pair wise. A pair is the score of two outcomes which is received from the same candidate at first stage and second stage. **The Karl Pearson's correlation coefficient is a measure of coefficient of reliability.**

This method is simple and easy to apply but has following drawbacks.

1. If the test is repeated immediately after the first, the scores are likely to be improved on account of memory effects and practice
2. On the other hand, if time lag for the retest is more then due to lack of memory and introduction of some other factors; the response for the retest may be poor and there by the coefficient of reliability will be decreased.
3. If the test contains novel features then the practice of the test will improve its application, hence, the performance of the test at first stage will be poor and it will improve in subsequent stages. Thus the correlation will be affected giving smaller coefficient of reliability.

**(2) Alternate or Parallel Forms Method**

Let there be a test "A" with items 1,2,3 —n . Manipulate a parellel test "B" with item 1,2,3, —n, administer the two tests A and B parallely to all the candidates.

Calculate the correlation coefficient for the two tests A & B. The self correlation is measure of reliability. This is a satisfactory method when the sufficient time has past between the administration of the two forms of A & B

**(3) The Rulon Method of Estimating Reliability**

Rulon gave the following formula for estimating the reliability from the scores on two halves of the same test

$r_{tt} = 1 - \frac{\sigma_d^2}{\sigma_t^2}$ where $\sigma_t^2$ = is the variance of raw score in the test

$\sigma_d^2$ = is the variance of difference of raw scores on the two halves of the test given by $\sigma_d^2 = (\Sigma di^2 / n)$

di = is the difference between two scores of half tests for i th individual

**(4) Kuder Richardson Formula**

This test is based on the assumption that all the items of the test are of equal or nearly equal difficulty and inter correlation. The most accurate and useful of all the formula is given by Kuder Richardson. This formula is based on interrelations.

$$r_{tt} = \frac{n}{n-1}\left[\frac{\sigma_t^2 - \sum \rho i\, qi}{\sigma_t^2}\right]$$

$r_{tt}$ = the reliability coefficient of test

n= number of items in the test

$s_t$ = standard deviation of the test scores

$q_i = i - \rho_i$ (i = 1 … n)

$\rho_i$ = the proportion of group answering i th test item correctly

This method of coefficient of reliability is free from the objections raised against the earlier methods. It stresses the intercorrelations of items in the test and the correlation of the item the test as a whole.

## 14.2.6 Validity of Test Scores

A measuring instrument is said to be valid if the measurements made by it are accurate and comparable with those made by a standard instrument. Because a test is like a measuring instrument one additional parameter required by a test to be satisfied with reliability is validity of test. For example a home made yardstick is valid if its readings are accurate in terms of a standard measuring rod.

Similarly a test is said to be valid, if the performance measured by it are accurate and comparable with the measurements as otherwise independently obtained. In other words, the validity of test depends upon the accuracy with which it measures, what it intends to measure.

**1. Calculation of Validity**

The validity of the test is determined experimentally by obtaining the coefficient of correlation between the scores of ‘n’ individuals on the given test (X) and some independent standard test (Y).

The test (Y) is called as criterion test. Thus it is the correlation between the criterion test and obtained test. The difficult aspects of validity are choice of a proper, and adequate criterion- variables and obtaining measures on the variable which are to be compared with the scores on the given test. The criterion can be an objective measure of performance or a qualitative measure such as judgement or excellence of work done.

A high correlation between X & Y is an evidence of validity provided that (i) The criterion Y was set up independently (ii) Both X and Y are reliable.

For example the validity of a typing test may be judged by correlation between the errors (score X) in the matter typed and the speed (criterion Y).

It may be pointed out that the validity is a highly relative concept. A test may be valid for a particular purpose, trait, group or situation and not always. If a test is used for measuring different traits then we must obtain its validity for different traits separately.

**2. Comparison between Reliability and Validity**

Let us take an example of clock to understand reliability and validity. Let us consider a clock which is quite good and accurate but is set ahead say by 10 minutes. Its time readings are reliable (consistent) but are not valid as judged by a standard time.

The reliability and revalidation are essential properties of a test. By reliability of a test we mean the stability or consistency of test scores; when the rest is administered to number of individuals on different occasions. Validity of a test is concerned with the accuracy of the test scores, when compared with the performance on the independent standard test (criterion). For valid test, it must be reliable. A test which is not quite reliable is hardly valid. The correlation of a test with a criterion is limited by its own index of correlation which provide the maximum correlation. The test is capable. A highly valid test can not be unreliable. The index of reliability is the square of coefficient of reliability. This can be taken as a measure of validity.

### 3. Attitude Scale Analysis

Always we ask the persons regarding the attitude towards job; attitudes towards work, attitude towards the working of a party. All these questions are intentended to know the liking, disliking or individuals feeling regarding the job, work or the working of a party. This is also known as the concept of attitude.

The concept of attitude can be defined as the degree of positive or negative feeling (affect) associated with some psychological trait or object. The attitude can be positive, negative, favourable or unfavorable towards the objects.

The attitude about a particular object can be judged through a planned questionnaire or schedule. This is possible when the atmosphere is free, but when the atmosphere is not free and individuals are reluctant to give their opinions the questionnaire should be balanced and should include all the possible outcomes of an object or trait. There are different types of psychological tests. These tests consist of items. We are required to respond to the items the psychological tests are useful in education, industry and research. The measurement requires attitude scales. The attitude scale provides the assessment degree of the respondent towards the object under study. The attitude scale consists of a number of items. These items are critically selected and edited in accordance with the criteria of the test. The experts of the subject or the individuals dealing with the trait prove to be the proper persons in editing and dealing with the items. The items which make an attitude scale are known as statements. Anything said about the psychological test is known as psychological statement. The class of such statements regarding the psychological object is known as Universe of content

There are three types of statements :

1. The statements relating to the objects for which a person accepts or agrees
2. The statement relating to the objects for which a person disagrees or rejects.
3. The statements regarding the objects that a person does not believe or does not disbelieve. The person is undecided or doubtful.

The response of several individuals must be tried to evaluate the statements. The statements for which high degree of positive or negative attitudes is received are retained as statements. The ambiguous statements can be eliminated from the universe.

**4. Suggestions regarding Scales**

Allen Edwards (1969) has summarized certain suggestions regarding the construction of scale as given below

1. The statements should pertain to the present and should not allow the respondent to strain his memory
2. Factual statements should be deleted.
3. The equivocal statements carrying more than one meaning statements should be eliminated. In the same manner the sentences carrying double negatives should be avoided.
4. The statement which are not relevant regarding the object should be avoided.
5. The plausible statement, which are likely to be responded by none or all should not be included
6. The entire range of the scale should be taken in to account while constructing the statement.
7. The statement should be simple, clear and direct. The words used should be understood by those who are given to complete the scale.
8. The statement should be small and direct. The compound and coplex sentence should be avoided.
9. One thought should be covered by one statement.
10. The statements containing the words such as all, everyone, always, none, nobody which universalize the thing should not be used
11. The words like rarely, not frequent, only just should be cautiously used.

## 14.3 Glossary

**Attitude** : Attitude is the positive or negative degree of feeling associated with psychological object

**Scale** : A psychological scale is a continum or continuity along which items tasks, problems, and likes are located in terms of some attribute.

**Reliability** : Reliability means consistency and stability of measurement scale

**Validity** : validity means accuracy and comparability in which what it intends to measure.

**Measurements** : Measurement is the process of assigning numbers or symbols to the object or observations. It enables further statistical analysis of the data.

**Nominal Scale :** Assigning the objects with symbols or classifying the objects with similarly and naming them with letters, numerals or symbols.

**Ordinal Scale :** Assigning the objects with ranks of ascending or descending order.

**Interval Scale :** A scale of measurement in which the zero point is arbitrary and the ratio of distance between any two points is constant.

**Ratio Scale** : The scale of measurement which has true zero point and where the ratios of two intervals are constant

**Statements :** The items making up an attitude scale are known as statements.

**Universe of Content :** The set of all possible statements about an object is known as universe of content.

**Reliable test Score** : A test score is said be reliable, when we can believe that the scores of the tests are stable and trustworthy.

**Valid Test Score :** A test score is said to be valid, when the measurements made by it are accurate and comparable with the standard measurement.

**Sociometry :** It is the science of techniques of measurement in Sociology

## 14. 4 Summary

There are certain Sociological or Psychological variables used in social science research and measurement of social variables and dimensions is a Sociometry study. There are four levels of measurements i.e. Nominal scale, Ordinal scale, Interval scale and Ratio scale. The sociological and psychological scales are interval scales. The Likerts Scale is given by different judges with respect to some stimulus or trait e.g. scales given for beauty or acting competitions. The reliability test means consistency test and applied to same object or persons on different occasions, relatively free from the errors of measurement. There are four methods of testing reliability i.e.

The Test-Retest methods, Alternate or Parallel Forms methods, Rulon method of Estimating Reliability and Kuoler Richardson methods. Thus in Sociometry the different scales and measures has valuable importance.

## 14.5 Exercises for Practice

**Answer the following questions in about 200 words each.**

1. Define attitude and describe the preliminaries required for attitude scaling.
2. What is T Score ? Explain the steps involved in T Score.
3. Describe Likert's scale for ratings and write down the procedure to obtain the scale
4. What is the concept of reliability ? Write different methods of studying reliability.
5. Define and compare reliability and validity of a test.

# Unit 15 : Graphic Presentation

***Index***



## 15.1 Introduction

The statistical data can be presented in tabular form for ready interpretation. However, the same information can also be presented with the help of diagrams and graphs. The statistical data when it is presented in the form of lines, bars circles, cubes, points, it is a graphic and diagrammatic presentation of data.

*Diagram and graphs are useful because :* (i) They provide bird's eyeview as a visual aid. (ii) As compared to numerical figures they are more appealing to eye and a layman can understand the results easily. They provide interesting, attracting and impressive set of information as compared to numerical data. (iii) They are more catching and extensively used in the exhibition, industrial and trade fairs. Human mind is naturally attracted towards the beautiful pictures. (iv) They save lot of time and effort required to grasp the results. (v) They highlight comparative performance and establish relationships and (vi) They can be used to know the future trend from the past information.

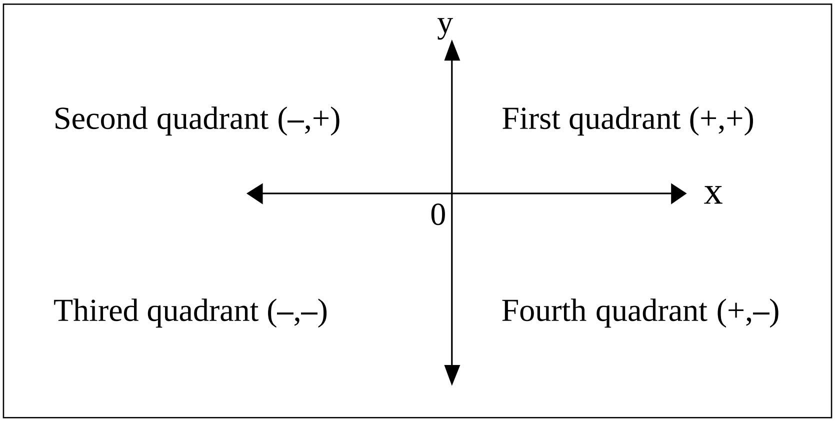
**After the study of this unit, you will be able to know and understand :**

- λ Rectangular Co-ordinate Graphs
- λ Frequency Distribution Graphs
- λ Diagrams and flowcharts

## 15.2 Content

### 15.2.1 Rectangular Co-ordinate Graphs

Graphing or plotting is done with reference to two lines or coordinate axes, the vertical axis or y axis and the horizontal axis or x axis. These basic lines are perpendicular to each other. The figure given below represents a system of coordinate axes.




The origin is the zero point or point of reference for both the axes. Distances are measured along the x axis to the right of ‘0’ are called positive distances. Distances measured along the x axis to the left of ‘0’ are called negative distances. In the same way the distances measured on the y axis above ‘0’ are positive and below ‘0’ negative distance.

The intersection of x and y axis at ‘0’ cause to have four quadrants. The upper right quadrant where both x and y are positive (+,+), the upper left or second quadrant x is negative and y is positive that is (-, +), the lower left or third quadrant both x and y are negative that is (-, -), where in the fourth quadrant or lower right quadrant x is positive and y is negative that is (+, -).

The distance of a point from ‘0’ on the x axis is commonly called as *abscissa* and the distance of the point from ‘0’ on the y axis is called as *ordinate*.

Locate the points A(3, -4), B (4,-5), C(-5,-6) D(-8,4) on the graph paper.

### The Guidelines for Graphing

**(1) Neatness :** The graphs are a sort of visual aid for presenting the numerical data; hence it should be fascinating to eyes and give lasting impression to mind. The appropriate type of colors, thickness of lines, dots and dashes should be used to show the significance and relevance of appropriate things on the graph.

**(2) Title and Footnotes :** The title of the graph explains nature, period and the facts relating to certain object. Therefore, the title should be self explanatory, brief and catchy. All the details of a graph should be clear from the title. The title should not be ambiguous. It should be displayed on the top of the graph. The other details should be explained by making use of footnotes. The footnotes can be given on, the right of the top or at any appropriate space where the space is available explaining certain fact which are essential to cover the relevant information given by the graph.

**(3) Structural Frame Work :** For every independent variable (x) there is a corresponding value of dependent variable (y). It is customary to depict the values of dependent variable on y axis and take values of independent variable on x axis. While choosing the x and y axes a point should be borne in mind that the movement in the dependent variable is to be exhibited from the changes of independent variable. Hence, the position of axes and proportion of graph gives importance to the frame work and structure of the graph. If area of certain crop varies from years to years or there is a type of peculiar trend, then it must be clearly explained by the frame work of the graph.

**(4) Scale :** This is very important aspect of a graph. Given a graph paper of certain size, you have to accommodate all the facts and figures by plotting a graph. There can not be a specific rule for choosing a scale. However, by making use of the graph paper, given data and other resources, the appropriate scale for x and y should be selected. By making a proper selection of a scale, the minute changes and variations in the data can be explained very nicely. The utilization of graph paper should be done judiciously. To display relative changes in the magnitudes, the ratios, the logarithmic scale is used instead of natural or arithmetic scale. For example if we plot the trend values from the graph $y = ab^x$ or $y = a^{bx}{}_e$ we shall obtain a straight line graph; if we use logarithmic scale for y values.

**(5) False Base Line :** The vertical scale must start from zero. If the values of y are very small the graph will be very nearer to x axis and some times becomes difficult to measure the change, similarly when the y values are very large; the graph papers do not accommodate the points. Hence to have a balance in between the two situations, we can use a false base line. If the minimum of the y values are very distant or far greater than zero; then the point of origin for an effective portrayal of the fluctuations of y values is stretched by using false baseline, the vertical scale is broken ends space between the origin and minimum value of the dependent variable is omitted by drawing two zigzag horizontal lines above the base line. The use of false base line is extensively used for magnifying the minor fluctuations in a time series data. It gives economy to the space. However, it is necessary to remember the use of false base line while interpreting the results from the graph.

**6. Simplicity :** The graph should be as simple as possible, so that it should be easy to understand and prediction. If more than one aspect is covered in the same graph, it becomes confusing. However when a comparative picture of two entities is to be shown, then two graphs can be plotted on the same graph paper by making use of colors dots, etc.

**7. Source Note and Number :** The source note, where it is possible should be placed at the bottom of the graph. This gives reliability to the information which is different from different sources. When there are more number of graphs in addition to the title, a separate number should be given to the graph, which is easier to distinguish one graph from another. Secondly the relevance of a particular table can be explained or can be depicted, if appropriate table number is given to the data.

8. **Index :** The various types of colors shades can be used to know the two distinct features of graph, this will be attractive and self explanatory. The graphs are more appealing to eyes than the tabulated data. The frequency graphs are designed to reveal the characteristic features of frequency data. They facilitate comparative study of two or more frequency distributions regarding their shape and pattern.

## 15.2.2 Frequency Distribution Graphs

The commonly used frequency distribution graphs are (i) Histogram (ii) Frequency polygon (iii) Frequency curve (iv) Ogive or cummulative frequency curve.

**1. Histograms**

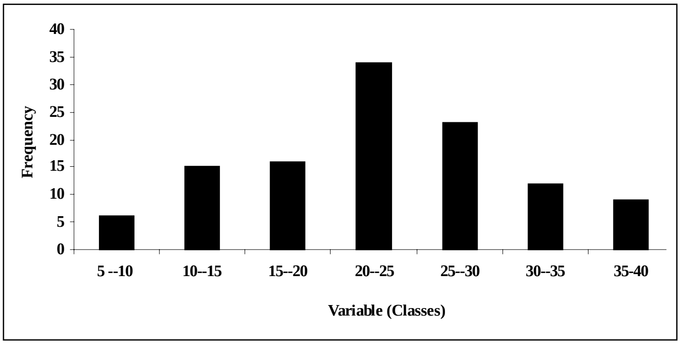
It is most commonly used graph charting continuous frequency distribution. The variable under study can be discrete or continuous. It consists in erecting a series of adjacent vertical rectangles on the sections of the horizontal axis (x-axis), with bases (sections), equal to the width of the corresponding class interval and heights proportional to the corresponding frequency of each class. The area of the rectangle is proportional to the frequency of the class.

When all classes are not of same size then the frequency density of each class can be worked out by dividing the width of class to the class frequency. Then height of the rectangle can be taken proportional to the frequency density of the class and histograms are worked out. There is a difference between bar diagram and histogram. A histogram is two dimensional, whereas bar diagram is one dimensional. In bar diagram the width of bar is at arbitrary and only height is taken into consideration. In histogram the area under rectangle is under consideration. Hence width and length, the two dimensions are important. In bar diagrams there is a proper spacing between two bars whereas in histograms the rectangles are adjacent to each other.

Histograms can not be constructed for open end classes. It can not be used to read the frequency over a fraction of a class interval because the height is not proportional to the fraction of a class. When the distribution is grouped but not continuous; histograms can not be drawn from two distinct classes. The drawing of histogram is possible only when the frequency distribution is continuous distribution. In case it is not continuous, the class limits are changed to form a continuous frequency distribution and then the drawing of histogram is possible.

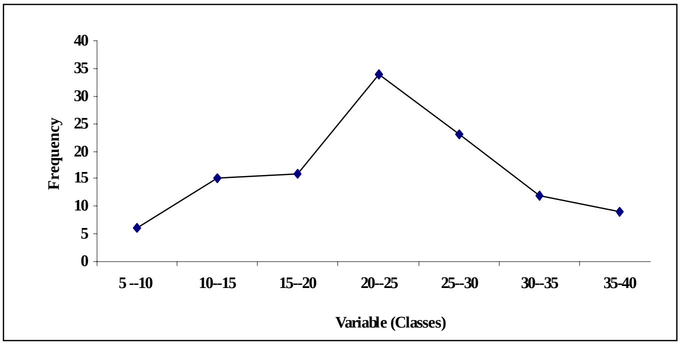
Draw the histogram from the following data

| Class | 5-10 | 10-15 | 15-20 | 20-25 | 25-30 | 30-35 | 35-40 |
|---|---|---|---|---|---|---|---|
| Frequency | 6 | 15 | 16 | 34 | 23 | 12 | 9 |




## 2. Frequency Polygon

Frequency polygon is another device of graphic presentation of frequency distribution. If the data is from discrete frequency distribution, frequency polygon is obtained on plotting the frequencies on the vertical axis or y axis against the corresponding values of the variable on the horizontal axis or x axis and joining the points by straight lines.



When histogram is drawn from the continuous frequency distribution. The frequency polygon is obtained by joining the top mid points of the rectangles in the histogram with straight lines. The joining of mid points of the histogram by straight lines is called as frequency polygon.

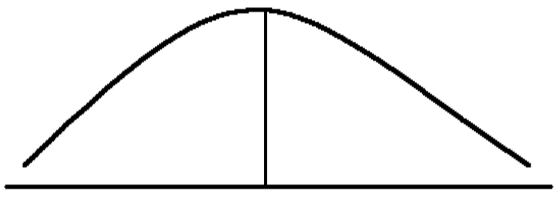
### 3. Frequency Curve

A frequency curve is a smooth free hand curve drawn through the vertices of a frequency polygon. The object of smoothening of the frequency polygon is to eliminate the random or erratic fluctuations that might be present in the data. The area enclosed by the frequency curve is same as that of histogram or frequency polygon. Frequency curve may be regarded as a limiting form of the frequency polygon.

Frequency curve gives the idea about the skewness and kurtosis of the distribution. A frequency polygon can be drawn with or without frequency polygon but for a frequency curve a logical sequence of frequency distribution is essential. Frequency curve should be tried when the distribution is continuous. The frequency curve should be obtained for sample studies which give the fair idea regarding the population distribution.

***Different types of data give rise to the variety of frequency curves.***

*(i) Curve of Symmetrical Frequency Distribution :* this curve rises steadily and reaches to a peak value and there after start going down slowly with the same pattern as it was increasing. The curve when it is divided at the peak point gives two identical halves. The maximum frequency line makes two equal halves of the symmetric frequency distribution.

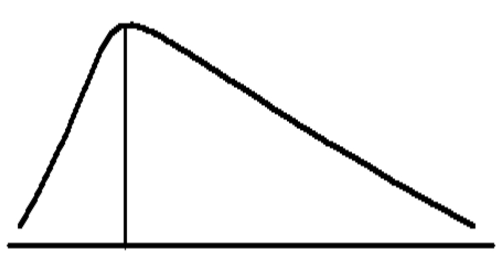


The normal probability curve.

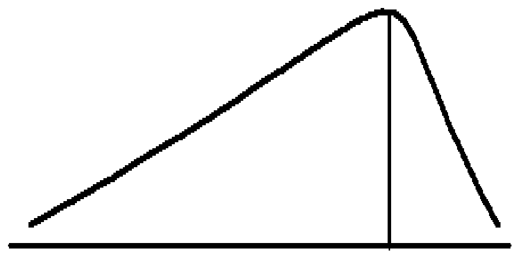
The most commonly used symmetric curve is normal probability curve. When the sample size is very large; the frequency curve resembles the symmetric probability distribution. Most of the curves which are based on random or natural experiment frequency distribution curves resemble with symmetric frequency curve. However, many of the samples coming from sociological, economic, and psychology studies give rise to asymmetrical curves.

*(ii) Skewed or Asymmetrical Curves :* A frequency curve is said to be asymmetrical or skewed when the curvature obtained is not symmetrical.

When the curve has a longer tail to the right side or when the curve is stretched more to the right side then the curve is called as positively skewed curve. The skew curve is obtained, when the frequency distribution is skew. In a positively skewed distribution most of the frequencies are associated with smaller values of the variable.



Mode Positively skewed curve



Mode Negatively skewed curve

Similarly a negatively skewed frequency curve is that curve, which has longer tail towards the left side and thus, it is stretched more to the left. In a negatively skewed distribution, most of the frequencies are associated with larger values of the variable.

## 4. Ogive Curve (Cummulative Frequency Curve)

Ogive curve is a graphic presentation of the cummulative frequencies (CF) distribution of continuous variable. To draw ogive curve, class intervals are plotted on x-axis and cummulative frequencies are plotted on y axis. There are two types of cummulative frequency distribution viz. Less than cummulative frequency and more than cummulative frequency and accordingly there are two ogive cures such as (i) Less than ogive curve (ii) More than ogive curve.

Less ogive is obtained by plotting upper class limits on x-axis and corresponding less than cummulative frequencies on y axis, then the points on the graph are joined by smooth hand curve. Less than ogive curve is slopping upwards from left to right. This is an increasing curve.

Similarly more than ogive curve or more than cummulative frequency curve is obtained by plotting lower limits of class intervals on x-axis and more than cummulative frequencies on y-axis. The smooth hand curve joining the points gives more than ogive curve, this is a decreasing curve and slopes downwords from left to right. We can draw less than cummulative frequency curve and more than cummulative frequency curve on the same graph paper. The less than ogive curve and more than ogive curve cintersect each other at point, where from the perpendicular length on y axis and x axis gives the values of median number on y axis and median on x axis.

They are useful to determine graphically, the proportion of observations below or above ogive value of the variable.  For symmetrical distribution mean, mode and median coincide. Ogive curves are useful for the comparative study of two or more distributions  If the frequencies are very large, they can be expressed in terms of percentages of the total frequency. The graph so obtained of the cummulative percentage frequencies in called as percentile curve.

## 15.2.3 Diagrams

The statistical data can be depicted with the help of diagrams. The diagrammatic presentation and types of diagram are described below :

### 1. Diagrammatic Presentation

The statistical data can also be presented with the help of diagrams and graphs in addition to classification and tabulation. The diagrams are geometric figures like point’s bars, lines, circles, pictures, maps, charts, etc.

*The advantages of diagrammatic presentation are given below* :

1. The data can be presented readily, easily and in a understandable form. It serves the purpose of visual aid, by giving birds eye view.

2. They create lasting impression in mind of an observer. Diagrams are more attractive, fascinating and impressive as compared to mere classification or tabulation. A layman can also understand the diagrams.
3. They are extensively used because they have universal applicability. Diagrams create meaningful impression upon our mind, before giving thinking to the statistical aspects of the data.
4. To highlight the salient features of the data, diagrams are most essential otherwise; the important information is lost in the numerical tabulation. The nature of the trend of the data can be easily understood by taking a look on the diagrams.

## 2. Types of Diagrams

The diagrammatic presentation of the data varies with number of diagrams presented and with different devices used in practice. The important types of diagrams used are :

*(1) One-dimensional diagrams* such as line diagrams and bar diagrams

*(2) Two-dimensional diagrams* such as squares and rectangles, circle and pie diagrams

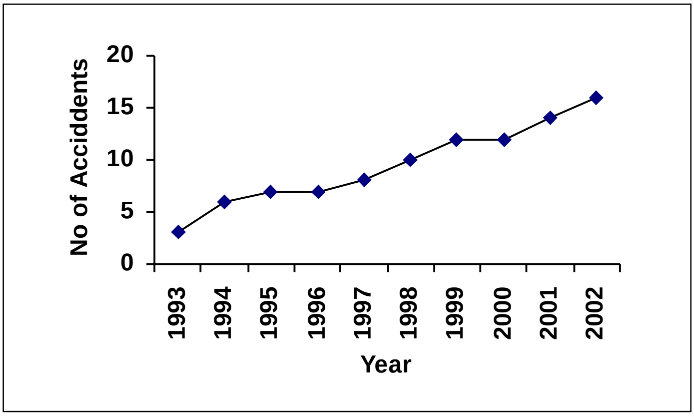
*(3) Three dimensional diagrams* such as cubes, prisms, pyramid type structure, cylinders and blocks are used

*(4) Pictograms and Cartograms.*

**(1) Line Diagram :** This is the simplest type of all the diagrams. The line diagram consists of drawing vertical lines or bars or horizontal lines. The height or length of the line is proportional to the frequency of the variable (x). the frequency of the variable is on y axis and the value of the variable is presented on x-axis with suitable scale. With the help of line diagrams, we can compare the magnitude of two or more things at a time.

**Illustration :** The year wise number of road accidents in Nashik are given below. Show the same with the help of line diagram.

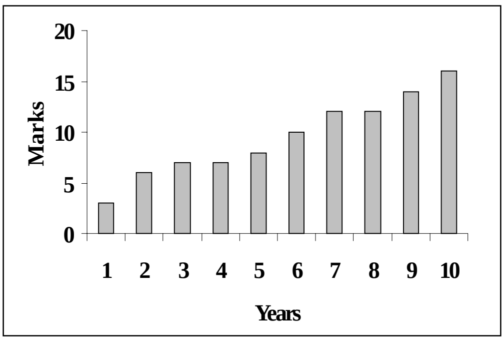
| Year | 1993 | 1994 | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|---|---|---|---|---|
| No. of accidents | 3 | 6 | 7 | 7 | 8 | 10 | 12 | 12 | 14 | 16 |



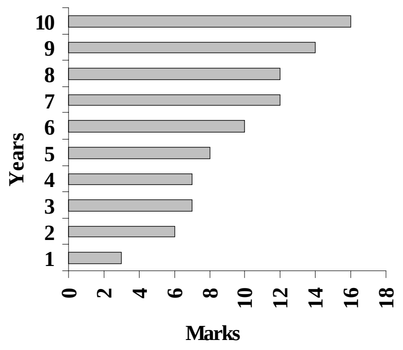

**(2) Bar Diagram :** Bar diagrams are equi-distance rectangles drawn vertically or horizontally for each category or group. The width of the rectangle is arbitrary whereas the length or height of the rectangle is proportional to the magnitude of the values presented. The bar or rectangles must be of uniform width as they are used to separate the category. The height scale which is used to present the group should be uniform for all the rectangles and it should be proportional to the value of the group. The bars must be placed with uniform spacing or distance between two rectangles. The base line for the construction of all the bars must be same. Bars can be drown vertically or horizontally. Bars should be arranged farm left to right in case of vertical bars. And they should be arranged from top to bottom in case of horizontal bars. The figures should be given on the top of the diagrams.

***Types of Bar diagrams***

*(a) Simple bar diagrams :* In simple bar diagrams only rectangles are erected with their length or height proportional to the values of observation. The graphs are dipicted as above.

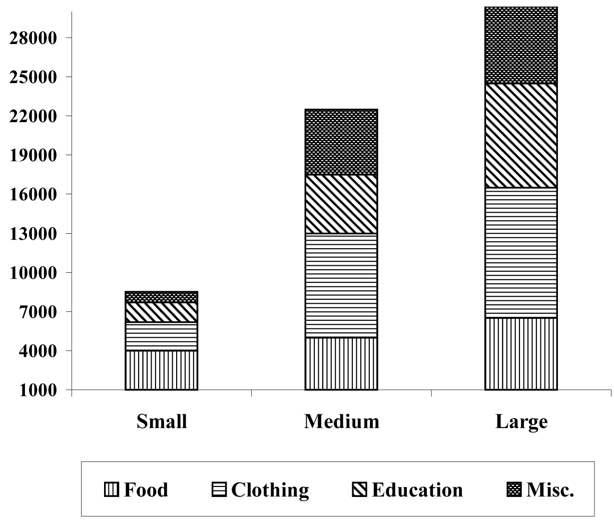



Vertical Bars



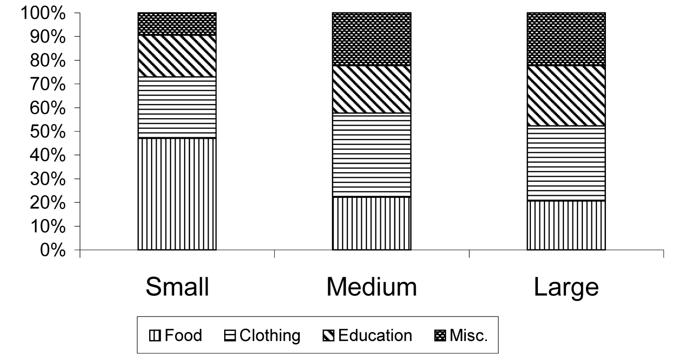

Horizontal Bars

*(b) Subdivided bar diagrams :* A component bar diagram facilitates comparison of various components. Different components can be shown by using index or key showing different colours, shades, dotting lining etc. The expenditure on different items of consumption of small medium and large farmers is presented here



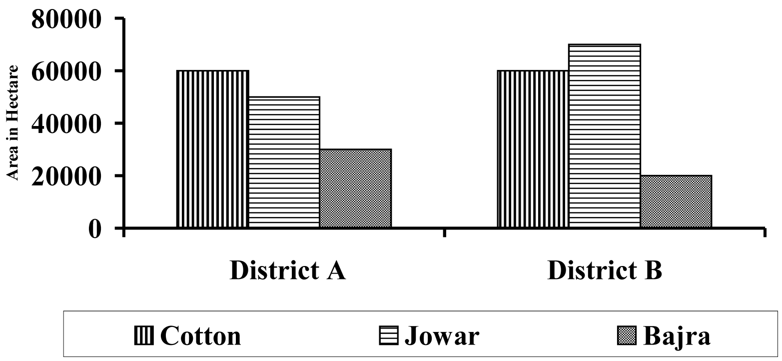

Subdivided bar diagrams

*(c) Percentage bar diagrams :* In percentage bar diagram instead of components their percentage total are worked out and are presented in terms of percentages. e.g. in above diagram the total expenditure of small farmer on items of expenditure are given in percentages and presented in singal bar diagram



Percentage Bar Diagram

*(d) Multiple bar diagrams :* Multiple bar diagrams are drawn in the same manner and procedure as that of simple bar diagrams. If there are two or more inter related aspects of a variable to be presented with the help of diagrams, then multiple bar diagrams are used. In simple bar diagram only one aspect, category or characteristic of a variable can be shown, whereas in multiple bar diagram, we can combine inter related characteristics of variable and they are presented for the sake of comparison.



Multiple Bar Diagrams

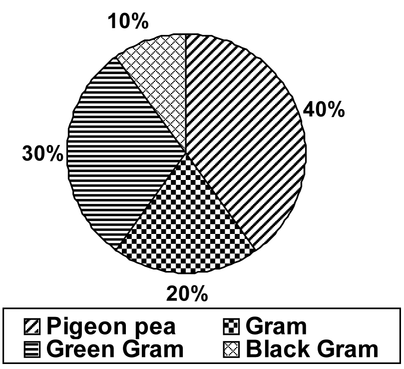
1. If you want to show the profits of two related firms owned by a same company or owner.
2. The cropping pattern of two districts with respect to major crops can be shown with the help of multiple bar diagrams. To distinguish one aspect from others dotting, shading, colouring or, key or index can be used.

### 3. Pie-chart or Angular Diagram

Pie-chart is a circle diagram in which various components of total magnitude are exhibited with the help of segments. The segments of the circle are decided with the help of percentage contribution of the components to the total and accordingly the angle at the center is calculated for the components. The total angle at the center is 360 which is equivalent to 100 percent.

*Illustration :* Draw a pie diagram for areas of a crop in certain district.

| Crop | Pigeon pea | Gram | Green Gram | Black Gram | Total |
|---|---|---|---|---|---|
| Area in Hectares | 4000 | 2000 | 3000 | 1000 | 10000 |
| Area Percentage | 40 | 20 | 30 | 10 | 100 |
| Angle in Degree | 144 | 72 | 108 | 36 | 360 |




Pie-chart or Angular Diagram

To draw a pie chart; the original data is expressed in terms of percentage to the total. Secondly the percentages are converted into angles. One percent is equal to (360/100) = 3.6 degrees; hence the percentages are converted into degrees by multiplying factor of 3.6. Draw a circle with appropriate radius depending upon the space of the paper. If two same characteristics are compared, two circles of equal radius should be drawn. Draw any line from the center to the circumference. Taking this line as base, draw the angle at the center. Use different colors, shades to depict different segments or components.

**4. Pictograms**

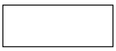
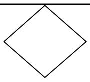
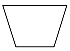
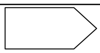
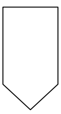
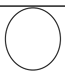
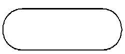
Pictogram is the technique presetting statistical data through appropriate pictures this type of diagram is useful for laymen. In this diagram the magnitudes of a particular phenomenon under study are presented through appropriate pictures, the number of pictures, drawn or the size of the pictures being proportional to the values of the different magnitudes to be presented.

**5. Cartograms**

In cartograms, statistical facts are presented through maps accompanied by various types of diagrammatic representation. Different types of rainfall zones, geographical zones can be shown with the help of cartograms.

## 15.2.4 Flowcharts

A flowchart is a graphical representation of a specific sequence of steps to be performed by the scientists or research worker to reveal a solution of a given problem. The flowchart is drawn with many symbols such as squares, circles and diagrams to represent many of the basic operations of the program. These symbols are connected by directed line segments to indicate the flow of information and processing.

| Flowchart Symbol | Represents |
|---|---|
|  | A processing operation such as addition or subtraction of two objects and movement within a unit |
|  | A decision point at which a branch to one of two or more alternate paths is possible |
|  | An input output operation |
|  | Modification or change in the steps instructions of a program |
|  | Entry from or exit to another aspect or phase, it indicates the direction of flow |
|  | Entry from or exit point on the same phase |
|  | The entry or exit point of the flowchart eg. Start, stop half and interrupt |

The flowcharting is an essential diagrammatic representation in computer programming now a days. If the same symbols are used in social sciences to evolve a programming model, it will be easier for a social scientist as well as the computer programmer.  To draw a flowchart or working model of the scheme for the schematic performance of the situation, following symbols are commonly used in computer programming.

## 15.3 Glossary

**Cummulative Frequency graph or Ogive Curve :** The ogive curve is a way of representing a frequency distribution by means of a diagram, ogive curve is obtained by plotting cummulative frequency on y axis and class intervals on x axis.

**Histogram :** If consists of erecting a series of adjacent vertical rectangles on the horizontal x axis with the base equal to the width of the class intervals and heights proportional to the frequency of each class such that area of the rectangle is equal to frequency of the class

**Frequency Polygon :** The frequency polygon is curve obtained by joining the midpoints of the upper side (top side) of rectangles with straight lines. The height of the mid point is proportional to the frequency of the class.

**Frequency Curve :** If the mid points of the histograms are joined with smooth hand curve or free hand curve, the curve obtained is called as frequency curve

**Symmetrical Distribution Curve :** In a symmetrical distribution curve the class frequencies first rise steadily and reach to a maximum, then diminish in the same identical manner, it is also called as bell shaped type of curve

**Line Diagrams :** Line diagrams are lines or bars whose height is proportional to frequency of the class

**Pie Chart :** Pie- chart is a circular diagram and used to show the contribution of different components to the total distribution

**Pictograms :** Pictograms are presentation of the data with the help of pictures.

## 15.4 Summary

The statistical data in tabular form requires lot of attention of the reader. However, if the same information is presented in the form of lines, bars, circles, etc. in the diagrammatic form it gives bird's eye view, save lot of time to grasp, highlight the relationship and thus human mind is attracted towards the beautiful pictures. The commonly used frequency distribution graphs are histogram, frequency polygon, frequency curve or ogive curves. The data is also presented in the form of diagrams which serves the purpose of visual aids and create lasting impression. The different types of diagrams used are one, two or three dimensional diagram as well as pictograms and cartograms.

## 15.5 Exercises for Practice

**Answer the following questions in about 200 words each.**

1. What is graphic representation of the data ? describe different methods of graphic presentation of data.
2. What are the various types of graphs used for presenting frequency distribution ?
3. What is cummulative frequency curve ? How the ogive less than and ogive more than curves drawn on the graph.
4. What is line diagram, bar diagram, multiple bar diagram and pie chart and explain them with graphic presentation ?
5. What is flowchart ? Explain the flowchart concept by taking a village communication model.

# Unit 16 : Analysis of Variance and Covariance

***Index***



## 16.1 Introduction

Experimentation and making inferences are the twin essential features of general scientific methodology. Statistics as a scientific discipline is mainly designed to achieve these objectives. Research is a continuous process and lot of valuable data are generated through this process. It is therefore essential to analyse the raw data to draw valid conclusions.

Analysis of variance (ANOVA) and Analysis of Covariance (ANCOVA) are the most commonly used techniques for statistical analysis. The analysis of variance (ANOVA) is a powerful Statistical tool for testing of significance. The ANOVA, sorts out the variance components attributable to different sources of variation like blocks, treatments, error, etc, similarly the analysis of Covariance (ANCOVA) sorts out the covariance effect and effects attributable to the different sources.

**After the study of this unit, you will be able to know and understand :**

- λ What is Analysis of Variance.
- λ What is Analysis of Covariance.
- λ Analytical procedure of these Techniques.
- λ Applications of ANOVA and ANCOVA in various fields.

## 16.2 Content

### 16.2.1 Analysis of Variance and Assumptions

The Analysis of variance is a powerful statistical tool for testing of significance of different factors. Also the test of significance based on 't' distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation, when we have three or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis, that all the samples are drawn from the same population with same mean.

There are two main objectives of the analysis of variance. Firstly, it helps in sorting out the variance due to different sources. If we consider the performance of varieties in an experiment, the variation observed can be partitioned into two parts, one representing the variation between the varieties and another within varieties. The second objective of analysis of variance is to provide basis for test of significance.

According to Prof. R.A. Fisher, **Analysis of variance (ANOVA) is the "Separation of variance describe to one group of causes from the variance ascribable to other group"**

In analysis of variance, the total variation in the data is distributed over different component causing variation. Thus ANOVA consists in the estimation of the amount of variation due to each of the independent factors separately and then comparing these estimates of variation due to assignable factors or causes with the estimate of variation due to chance factor, the later being known as experimental error or simply error.

**Assumptions in the ANOVA**

The assumptions of analysis of variance procedure :

(i) The observations are independently and normally distributed.

(ii) The effects of factors are additive.

(iii) The error variance remains more or less constant (homogeneous) for all observations under different treatment.

## 16.2.2 Types of Classification of Data

**1. One Way Classification**

When observations are classified over the different levels of a single factor, it is one way classification or one way ANOVA, we consider only one factor which is responsible for the variation except error.

*Model for one way classification :* A linear expression of effects of factors.

Yij : $j^{th}$ observation on $i^{th}$ treatment or $i^{th}$ level of the factor

$Yij = \mu + \alpha i + \varepsilon ij$ where

$\mu$ = General mean of the data

$\alpha i$ = Effect of $i^{th}$ treatment

$\Sigma ij$= Error of chance effect.

*Assumptions*

(i) Effects of factors are additive

(ii) $\Sigma ij$ are independently and normally distributed with mean zero and variance $\sigma^2$,

$$\sum_i eij = \sum_j eij = 0$$

(iii) $\alpha i$ is effect of $j^{th}$ treatment which is defined as deviation from the general mean therefore $\Sigma \alpha i = 0$

(iv) Error variance remains same under all treatments.

**2. Two Way Classification**

In two way ANOVA the data are classified on the basis of two factors e.g. the agricultural output may be classified on the basis of two factors i.e. different varieties and various fertilizers used

*Model for two-way classification :* An observation is considered as a linear combination of different factors including error.

$Yij = \mu + \alpha i + \beta j + \Sigma ij$

$\mu$ = General mean, $\alpha i$ = effect of $j^{th}$ level of factor A, $\beta j$ = effect of $j^{th}$ level of fastor b, $\Sigma ij$ = error of chance effects

*Assumption :* As given in the one-way classification. Additional assumption is $\Sigma\beta j = 0$

## 16.2.3 Analysis and Interpretation of Data

### 1. Statistical Analysis of One Way Classified Data

We shall illustrate the procedure of statistical Analysis and various formulae of one way classified data with following example.

| Treatments | Scores out of 10 obtained by five students selected at random in a test | Treatment total |
|---|---|---|
| School I ($T_1$) | 9 7 6 5 8 | 35 |
| School II ($T_2$) | 7 4 5 4 5 | 25 |
| School III ($T_3$) | 6 5 6 7 6 | 30 |
| | | 90 (G) |

Let 't' be the number of treatments, 'r' the number of replications in this case t=3, and r=5, the total number of observations r x t = 15

**Calculation of sum of squares**

(1) Correction factor (Cf ) $Cf = \frac{(G)^2}{rt} = \frac{(90)^2}{15} = 540$

(2) Total Sum of Squares (TSS) = $\Sigma\Sigma Yij^2 - CF = (9)^2 + (7)^2 \ldots\ldots + .(6)^2 - CF$

$= 568 - 540 = 28$

(3) Sum of Squares between treatments

$$= \frac{\sum Ti^2}{r} - CF$$

$$= \frac{(35)^2 + (25)^2 + (30)^2}{5} - CF = 550 - 540 = 10$$

(4) Sum of Squares within treatments or Error Sum of Square.

= Total SS - SS between treatments

= 28 – 10 = 18

**Analysis of variance Table (ANOVA)**

| Sources of variation | Degres of freedom(DF) | Sum of Squares (SS) | Mean Sum of Squares(MSS) MSS=SSdf | Variance Ratio (F) |
|---|---|---|---|---|
| Between treatment | (t-1) = 2 | 10 | 5 | 3.33 |
| Within treatment (Error) | (rt-1)-(t-1) = 12 | 18 | 1.5 | |
| Total | rt-1 = 14 | | | |

If F calculated > F table value then, Treatment differences are **significant**

If F calculated < F table then, treatment difference are **non-significant**

In the present case F calculated value 3.33 is less than the table value 3.88 at 5% level of significance. Therefore, treatment differences are non significant indicating that there may not be differences in the scores obtained by the students of different schools in a city.

**2. Procedure for Statistical Analysis of Two Way Classified Data**

We shall illustrates the analytical procedure and various formulae of two way classified data with the help of following example. The following table gives quality rating of service stations by five professional raters

| Rater | Service stations | | | | | | | | | | Treatment |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total (Ti) |
| A | 99 | 70 | 90 | 99 | 65 | 85 | 75 | 70 | 85 | 92 | 830 |
| B | 96 | 65 | 80 | 95 | 70 | 88 | 70 | 51 | 84 | 91 | 790 |
| C | 95 | 60 | 48 | 87 | 48 | 75 | 71 | 93 | 80 | 93 | 750 |
| D | 98 | 65 | 70 | 95 | 67 | 82 | 73 | 94 | 86 | 90 | 820 |
| E | 97 | 65 | 62 | 99 | 60 | 80 | 76 | 92 | 90 | 89 | 810 |
| Total Tj | 485 | 325 | 350 | 475 | 310 | 410 | 365 | 400 | 425 | 455 | 4000 (G) |

$H_0$ : No significant differences between ratings as well as between service stations.

**Calculation of sum of squares**

(1) Correction factor (CF) : $CF = \dfrac{(G)^2}{N} = \dfrac{(4000)^2}{5\times10}$ = 3,20,000

(2) Total sum of squares

$= \Sigma\Sigma\ Xij^2$ - CF

$= (99)^2 + (96)^2 + \ldots\ldots\ldots. (89)^2$ - CF

= 9948

(3) Rows sum of squares

$= \dfrac{\sum Ti^2}{C} - Cf$ C: 10 of column

$= \dfrac{(830)^2 + \ldots.. + (810)^2}{10} -$ 3,20,000 = 400

(4) Column SS $= \dfrac{\sum Tj^2}{r} - cf$ R = Number of Rows

$= \dfrac{(485)^2 + (325)^2 + \ldots. + (455)^2}{5} -$ 3,20.000 = 6800

Error SS = TSS - Row SS - Column SS

= 9948 - 403 –6810 = 2738

**ANOVA Table**

| Sources of variation | DF | SS (SS/ DF) | MSS value | F-cal | F table Value |
|---|---|---|---|---|---|
| Between columns | (r-1) 9 | 6810 | 756.67 | 9.95 | 2.15 |
| Between Rows | (c-1) 4 | 400 | 100 | 1.31 | 2.63 |
| Error | 49-13=36 | 2738 | 76.06 | | |
| Total | rc-1 50-1=49 | | | | |

In case of service stations F- calculated value i.e 9.95 is greater than table value of F at 9 and 36 degrees of freedom i.e. 2.15

Hence $H_0$ is rejected at 5% level of significant and we conclude that there is significant difference between service stations .

In case of ratings the F- calculated value i.e 1.31 is less than table value $F_{4,}$ 36, at 5% = 2.62. Hence, $H_0$ is accepted and we conclude that there are no significant differences in ratings.

## 16.2.4 Analysis of Covariance (ANCOVA)

When blocking is not effective for reducing variability due to the particular source of variability present or when experimental error remains large in spite of blocking ,other statistical tools for error control are available. One of the most effective technique in agricultural research is the Covariance Technique

The object of the experimental design in general happens to be to ensure that the results observed may be attributed to the treatment variable and no other casual circumstances .For example , the researcher studying one independent variable X, may wish to control the influence of some uncontrolled variable known as covariate or the concomitant variable, Z, which is known to be correlated with the dependent variable, Y, then one should use the technique of analysis of co-variance for a valid evaluation of the outcome of the experiment. In psychology and education, primary interest in the analysis of co- variance rests in its use for the statistical control of an uncontrolled variable.

### ANCOVA Technique

While applying ANCOVA technique, the influence of uncontrolled variable is usually removed by simple linear regression method and the residual sum of squares is used to provide variance estimates which in turn is used to make tests of significance.

In other words co-variance analysis consists in subtracting from each individual score (Yi) that portion of it $Yi^1$ that is predictable from uncontrolled variable (Zi) and then computing the usual analysis of variance on the resulting $(Yi\text{-}Yi^1)$, making the due adjustment to the degrees of freedom because of the fact that estimation of using regression method causes loss of degrees of freedom.

### Assumptions of ANCOVA

The ANCOVA technique requires one to assume that there is some (Relationship should be linear) relationship between the dependent variable and the co-variate. We also assume that this form of relationship is the same in various treatment groups. Other assumptions are

(1) Treatments, blocks and regression effects must be additive.

(2) Various treatments groups are selected at random from the population.

(3) The co-variate is not affected by treatments of the present experiment.

(4) The groups are homogeneous is variability.

(5) The residuals must be normally and independently distributed with zero means and the same variance.

(6) The regression is linear and same from group to group.

## 16.2.5 Need and Uses of Covariance Technique

Co-variance analysis should be used with care and the results should be interpreted in light of particular covariate used and the particular values of covariate obtained. The effects eliminated should be irrelevant to the objectives of the experiment, and the test of significance of treatments comparisons should not be invalidated.

The ANCOVA technique can be used to great advantage in agricultural research in the following applications.

(1) Adjustment of soil heterogeneity.
(2) Adjustment for crop stand irregularities
(3) Adjustment for non-uniformity of pests and disease incidence
(4) Estimation of missing data.
(5) Adjustment of competition effects in green house trials.

## 16.2.6 Analysis and Interpretation of ANCOVA

We shall illustrate various formulae and analytical procedure of ANCOVA by following example

Let Y be the covariate or concomitant variable, using ANCOVA, calculate the adjusted total, within and between groups, sum of squares on X and test of significance of difference between the adjusted means on X by using appropriate F- ratio. Also calculate the adjusted means on X.

***Example***

The following are the paired observations for three experiments groups.

| Group I | | Group II | | Group III | |
|---|---|---|---|---|---|
| X | Y | X | Y | X | Y |
| 7 | 2 | 15 | 8 | 30 | 15 |
| 6 | 5 | 24 | 12 | 35 | 16 |
| 9 | 7 | 25 | 15 | 32 | 20 |
| 15 | 9 | 19 | 18 | 38 | 24 |
| 12 | 10 | 31 | 19 | 40 | 30 |
| 49 | 33 | 114 | 72 | 175 | 105 |
| **(Total)** | | | | | |
| 9.8 | 6.6 | 22.8 | 14.4 | 35.0 | 21.0 |
| **(Mean)** | | | | | |

***Solution***

$\Sigma X = 49 + 114 + 175 = 338$

Correction factor (CF) for X = $\dfrac{(\Sigma X)^2}{N}$

$= \dfrac{(338)^2}{15} = 7616.27$

$\Sigma Y = 33 + 72 + 105 = 210$

CF for Y = $\dfrac{(\Sigma Y)^2}{N} = \dfrac{(210)^2}{15} = 2940$

Also calculate $\Sigma X^2 = 9476$

$\Sigma Y^2 = 3734$ $\quad \Sigma XY = 5838$

$$\text{CF for XY} = \frac{\sum X.\sum Y}{N} = \frac{(338)\ (210)}{15} = 4372$$

Hence : Total S S for X = $\Sigma X^2$ - CF for X = 9476- 7616.27 = 1859.73

$$\text{SS between groups for X} = \frac{(49)^2}{5} + \frac{(114)^2}{5} + \frac{(175)^2}{5} - (\text{Correction factor for X})$$

= (480.2+2599.2+6125) –(7616.27)

= 1588.13

SS within groups for X = (total SS for X)-(SS between groups for X)

= (1859.73)-(1588.13) =271.60

Similarly we work out the following values in respect of Y

Total SS for Y = $\Sigma Y^2$ -Correction factor Y

= 3734-2940=794

$$\text{SS between groups for Y} = \left\{\frac{(33)^2}{5} + \frac{(72)^2}{5} + \frac{(105)^2}{5}\right\} - \text{Correction factor Y}$$

= (217.8 + 1036.8 + 2205 ) - ( 2940) = 519.6

SS within groups for Y = (Total SS for Y) – SS between group for Y)

= (794)-(519.6) = 274.4

Then we work out the following values in respect of both X and Y Total sum of product for XY = $\Sigma XY$ - Correction factor for XY = 5838-4732=1106

$$\text{SP between groups for XY} = \left\{\frac{(49)(33)}{5} + \frac{(114)(72)}{5} + \frac{(175)(105)}{5}\right\} - \text{CF for XY}$$

= (323.4+ 1641.6+3675) -(4732) = 908

SP within groups for XY (Total sum of product)- (SP between groups for XY)

= (1106)-(908)=198

**ANOVA table for X,Y and XY**

| Source | d.f. | SS for X | SS for Y | Sum of product XY |
|---|---|---|---|---|
| Between groups | 2 | 1588.13 | 519.60 | 908 |
| Within groups | 12 | $E_{XX}$ 271.60 | $E_{YY}$ 274.40 | $E_{XY}$ 188 |
| Total | 14 | $T_{XX}$1859.73 | $T_{YY}$ 794.00 | $T_{XY}$1106 |

$$\text{Adjusted total SS} = T_{XX} - \frac{(T_{XY})^2}{T_{YY}} = 1859.73 - \frac{(1106)^2}{794} = (1859.73)\text{-}(1540.60) = 319.13$$

$$\text{Adjusted SS within group} \quad = E_{XX} - \frac{(E_{XY})^2}{E_{YY}} = 271.60 - \frac{(198)^2}{274.40}$$

= (271.60) – 142.87) = 128.73

Adjusted SS between groups = (Adjusted total SS) – (Adjusted SS within groups)

= (319.13-128.73) = 190.40

**ANOVA table for Adjusted X**

| Source | d. f. | SS | MS | F-ratio |
|---|---|---|---|---|
| Between Groups | 2 | 190.40 | 95.2 | 8.14 |
| Within Groups | 11 | 128.73 | 11.7 | |
| Total | 13 | 319.13 | | |

At 5% level the table value of F for $v_1 = 2$ and $v_2 = 11$ is 3.98 and at 1% level the table value of F is 7.21 Both these values are less than the calculated value (i.e., calculated value of 8.14 is greater than table values) and accordingly we infer that F-ratio is significant at both levels, which means the difference in group means is significant.

Adjusted means on X will be worked out as follows:

Regression coefficient for X on Y

$$\text{i.e., } b = \frac{\text{Sum of product within group}}{\text{Sum of Squares within group for Y}} = \frac{198}{274.40} = 0.7216$$

Regression SS = bExy, Test of Significance Fije = (Regression SS) / Exx

e = error d.f.

| Deviation of initial group means from general mean (=14 ) in case of Y | | Final means of means group in X (unadjusted) |
|---|---|---|
| Group I | -7.40 | 9.80 |
| Group II | 0.40 | 22.80 |
| Group III | 7.00 | 35.00 |

Adjusted means of groups in X

= (Groups mean) – b (deviation of the group mean from general mean of Y)

Hence, (i) Adjusted mean for Group I = (9.80) – 0.7216 (-7.4)= 15.14

(ii) Adjusted mean Group II = (22.80)-0.7216 (0.40) = 22.51

(iii) Adjusted mean for Group III = (35.00)- 0.7216 (7.00) = 29.95

## 16.3 Glossary

**Variance :** It is described as the mean of the squares of deviations taken from the mean of the given series of data.

**Standard Deviation** : The square root of the variance is known as standard deviation.

**ANOVA (Analysis of variance**) : Separation of variance ascribable to one group of causes from the variance ascribable to other group.

**One way ANOVA** : we consider only one factor e.g., effect of one factor fertilizer on yield of crop.

**Two way ANOVA** : Data are classified on the basis of two factors e.g., output may be classified on the basis of different varieties of seeds and also on the basis of different fertilizers used.

**ANCOVA (Analysis of Co-variance**) : In ANCOVA Besides measuring main character, X, an additional variable, Y, called covariate is measured from each experimental unit.

**Variance ratio test (F-test) :** F-test is based on F distribution and used to test the equality of variances of two normal populations.

## 16.4 Summary

Analysis of variance and co-variance are extremely useful techniques concerning research in the field of Economics, Biology, Education, Psychology, Sociology, Industry and of several other disciplines. Through ANOVA technique one can in general investigate any number of factors which are said to influence the dependent variable. In a two or more way ANOVA the interaction (i.e., inter relation between two independent variables), if any, between two independent variables affecting a dependent variable can also be studied.

When blocking is not effective for the particular source of variability present or when experimental error remains large inspite of blocking, other statistical tools for error control is available. One of the most effective technique used in agriculture research is the covariance technique. Covariance is not only used to reduce experimental error, it is now used to assist in interpretation and characterization of the treatment effects upon the character of interest in much the same way that regression and correlation analysis is used.

## 16.5 Exercises for Practice

**Answer the following questions in 200 words each**.

1. What is analysis of variance (ANOVA) and analysis of covarience (ANCOVA) and explain their significance in statistical analysis.
2. State and explain types of classification of data and their interpretation.
3. What is co-variance analysis ? Illustrate its analytical procedure with assumption.
4. Differentiate between ANOVA and ANCOVA and explain the application of ANCOVA in the field of agricultural research.

# Unit 17 : Multivariate Analysis

***Index***



## 17.1 Introduction

The technique which simultaneously analyses more than two variables on a sample of observation is categorised as multivariate technique. Most of the research studies involve more than two variables, in such a situation analysis is desired of the association between the dependent variable and several independent variables or we may be required to study the association among various independent variables.

All such analyses are termed as multivariate analyses or multivariate techniques. Thus the technique in which we simultaneously study the various relationships among variables are termed as multivariate analysis. Multiple correlation and regression, **Mahalanobis's $D^2$ analysis** and discriminate function are the important techniques of multivariate analysis by which we study the relationship among variables.

**After the study of this unit, you will be able to know and understand :**

- λ Various techniques for multivariable relationship
- λ Multiple correlation analysis
- λ Discriminate function and $D^2$ Analysis

## 17.2 Content

### 17.2.1 Concept of Multivariate Analysis

The univariate analysis carried out separately for each variable may, at times, lead to incorrect interpretation of results. This is so because univariate analysis does not consider the correlation or interdependence among the variables. Therefore , it was necessary to have a technique which simultaneously study the association or relationship amongst variables. Such a powerful technique is **Multivariate Analysis Technique**. Today because of its usefulness these techniques are being applied in many fields such as economics, sociology, psychology, agriculture, biology and medicine. The basic objective underlying multivariate techniques is to represent a collection of massive data in simplified way.

For understanding multivariate technique one must be familiar with fundamental concepts of linear algebra, vector spaces and univaried analysis. Also before applying multivariate technique one should consider the nature and structure of data and aim of analysis. The use of multivariate technique is increasing day by day because of advanced computing facilities available on your personal computer.

### 17.2.2 Multiple Correlation Analysis

**1. Multiple Regression**

The simple linear regression analysis can accommodate only one independent variable X, which must be the major, if not the only variable, affecting the dependent variable Y.. Hence where more than one variables are involved, simple linear regression analysis is clearly inappropriate. Therefore, there is increasing need of regression. procedure, that can handle more than one independent variable, particularly for researchers involved in survey data, where it is extremely difficult and at the same times impossible to control the number of variable affecting the dependent variable.

The multiple linear regression analysis, where a requirement for a single independent variable is relaxed, can very often satisfy these needs. In this procedure any number of independent variables suspected to affect the dependent Y can be included in the multiple regression analysis. Thus, for K independent variables, $X_1, X_2, \ldots Xk$, the functional form of the multiple linear regression is

$$Y = a + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots + \beta_k Xk + e$$

Where $\beta_k$ are the partial regression coefficients associated with each Xi and 'a' is the intercept or constant (the value of Y when all Xi's have zero values). $\beta_k$ the regression coefficient for $i^{th}$ variable. measures change in value of dependent variable (Y) for unit change in the values of independent variables (Xi) and 'e' is error term.

## 2. Multiple Correlation

A simple correlation measures the degree of relation between a dependent variable and one independent variable. A multiple correlation coefficient measures a degree of combined relation between a dependent variable and series of independent variables. It can also be explained as the correlation between the observed values of the dependent variable and its estimated values for the independent variate values , estimated with the help of multiple regression equation.

Let the dependent variable be Y and independent variable be $X_1, X_2, X_3 \ldots\ldots Xk$. Then the estimated multiple regression equation will be of the form.

$$Y = a + b_1 X_1 + b_2 X_2 + \ldots\ldots + bk\ Xk \ldots\ldots(1)$$

Where 'a' is intercept or constant. $b_1, b_2, b_3, \ldots\ldots b_k$ are the partial regression coefficients for $X_1, X_2, \ldots\ldots X_K$ respectively.

Expressing it in terms of $y, x_1, x_2, x_3 \ldots\ldots x_k$ the deviations of $Y, X_1, X_2, X_3 \ldots Xk$ from their respective means $\overline{Y}, \overline{X}_1, \overline{X}_2, \overline{X}_3 \ldots\ldots \overline{X}k$ we get.

$$y = b_1 x_1 + b_2 x_2 + \ldots\ldots + bk\ xk \ldots\ldots \quad (2)$$

If $Y^1$ represents the estimated value of Y, estimated from the regression equation (1), then

$$Y^1 = a + b_1 X_1 + b_2 X_2 + \ldots\ldots + bk\ Xk + e \quad \ldots\ldots\ldots\ldots(3)$$

then the coefficient of multiple correlation (R) is the simple correlation between Y and $Y^1$. In terms of deviations from means, it will be simple correlation between y and $y^1$, where $y^1$ is the estimated value of y estimated from equation (2)

$$y^1 = b_1 x_1 + b_2 x_2 + \ldots\ldots bk\ xk \ldots\ldots \quad (4)$$

$y = Y-\bar{Y} \quad e = Y-Y^1 \quad \ldots\ldots \; (5)$

This coefficient of multiple correlation, expressed by Ry x1,x2,…xk or simple R is given by following relation.

$$1-R^2 = \frac{\sum(\text{Deviations of observed Y from Estimated values})^2}{\sum(\text{Deviations of observed Y from the mean } \bar{Y})^2} = \frac{\sum(Y - Y^1)^2}{\sum(Y - \bar{Y})^2}$$

It is also given by the relation

$$R^2 = \frac{\sum(Y - \bar{Y})^2}{\sum(Y - \bar{\bar{Y}})^2} = \frac{\text{SS (Multiple Regression)}}{\text{Total sum of square}}$$

$R^2$ is also called as coefficient of multiple determination and it explains total variability in Y accounted jointly by the various independent variables under study . The values of $R^2$ lies between O to 1

Also the multiple correlation coefficient should be necessarily not less than zero and lies between O to 1. If R= O, then it indicates that the variate Y to which it refers is completely uncorrelated with other variates. If R=1 it proves that all the residuals are zero.

### 17.2.3 Discriminate Function

The use of discriminate function for plant selection was first proposed by Smith in 1936. He suggested that a better way of exploiting genetic correlation with several traits having high heritability is to construct an index, called **selection index**, which combines information on all the characters associated with the dependent variable, say yield. Thus selection index refers to a linear combination of characters associated with yield. The best known selection indices involve discriminate functions based on the relative economic importance of the various characters. This technique provides information on yield components and thus aids in indirect selection for the genetic improvement of yield. Since the desirable genotypes are discriminated from the undesirable ones, based on the combination of various characters, this technique is known as discriminate function analysis. Later on in 1943, Hazel developed a Simultaneous selection model based on the approach of path analysis; subsequently several modifications were made in the technique by different workers for specific breeding requirements. We shall describe below different types of selection index.

### 1. Type of Selection Index

Selection indices are of following three types :

**(1) Classical Selection Index :** This was developed by Smith in 1936 . This involves several characters simultaneously in selection index and discriminates between desirable and undesirable genotypes on the basis of selection efficiency. This was first applied for plant selection by Smith (1936) and then by Hazel (1943), for animal selection.

**(2) General Selection Index :** This was first proposed by Hanson and Johnson in 1957. This is a modification of the scheme of Smith. In this model the weightage for various traits are based on the average statistics for several populations. This selection index has wide application in plant breeding. This model has been modified by several workers to meet their specific breeding requirements.

**(3) Restricted selection index :** This was proposed by Kempthorne and and Nordskog in 1959. This helps in improving a set of characters keeping the value of other characters intact. Sometimes the restriction is put on single character and sometimes on double characters.

An experimental comparison of the model of Smith with that of Hanson and Johnson for soybean yields showed that only the specific index as proposed by Smith was more effective than direct selection for yield. Worked out examples of above three models are given by Singh and Choudhary (1995).

### 2. Computation of Selection Index

The estimation of all types of selection indices is based on the estimates of phenotypic and genotypic variances and covariance's of the characters involved in the selection index. Selection indices can be worked out from replicated data only. The Computation of selection indices involves following steps.

(1) Crossing parents according to mating plan to produce the progenies for evaluation.

(2) Evaluation of these progenies in replicated field trials and recording observations.

(3) Calculation of phenotypic and genotypic variances and covariances.

(4) Construction of Selection indices using one of the three models described above. The construction of selection index is based on the following statistical and geneticaly assumption.

**(i) Statistical Assumptions :** These include random selection of parents for mating, absence of genotype x environment interactions and linear or additive nature of component characters to be included in the index.

**(ii) Genetical Assumptions :** These assumptions consist of diploid segregation, lack of maternal effects, absence of linkage, absence of epistasis, absence of multiple alleles and equal survival of all genotypes in a population. However, these assumptions are seldom fulfilled.

### 3. Merits and Demerits

**(1) Merits** : (1) This technique provides information on yield components and thus aids in indirect selection for genetic improvement of yield, (2) This technique can be applied to both parental population as well as segregating populations.

**(2) Demerits :** (1) The construction selection index is difficult task which requires lot of statistical calculations and some times selection indices are not as effective as yield alone in spite of making extra efforts for their estimation. The calculation becomes still complicated, when the index is extended to cover many component characters. (2) Selection index is applicable to individual plant selection only. However, family selection is not greatly improved by the use of an index. (3) Selection indices have limited applications in practical plant breeding because of inaccuracies associated with the estimation of variances and covariances. (4) Different estimation procedures generally provide different values, when applied to the same set of population.

In short it can be summarized that discriminate function technique involves development of selection criterion on a combination of various characters, and aids the researcher in indirect selection for genetic improvement in yield. In plant and animal breeding, selection index refers to a linear combination of characters associated with yield. Selection indices are of three types, viz., classical, general and restricted. The construction of selection index consists of three main steps, viz., calculation of (1) weightage coefficients, (2) expected genetic advance and (3) relative efficiency of selection. Selection indices do not appear to be in widespread use in practical plant breeding due to several reasons.

## 17.2.4 $D^2$ Analysis and Interpretation of Results

The concept of $D^2$ statistics was originally developed by P.C. Mahalanobis in 1928. He used this technique in the study of Anthrapometry and psychometry. Rao (1952) suggested the application of this technique for the assessment of genetic diversity in plant breeding. Now this technique is extensively used in genetics and plant breeding for the study of genetic divergence in the various breeding materials and in many other fields. This is one of the potent techniques of measuring genetic divergence. In plant breeding, genetic diversity plays an important role because hybrids between lines of diverse origin, generally, display a greater heterosis than those between closely related parents. This has been observed in maize, cotton and several other crops. Genetic diversity arises due to geographical separation or due to genetic barriers to crossability. The analysis of $D^2$ statistics consists of the following main steps.

### 1. Important Steps in $D^2$ Analysis

**(1) Selection of Genotypes :** The material which has to be evaluated for genetic diversity may include Germplasm lines, strains and varieties. The genotypes are generally selected on the basis of phenotypic variability or geographical origin.

**(2) Evaluation of Material :** The selected genotypes are evaluated in replicated field trial and observations are recorded on various quantitative characters.

**(3) Biometrical Analyses :** First variances for various characters and covariances for their combinations are estimated. Then $D^2$ analysis is done as per Mahalanobis (1928).

### 2. Mahalanobis $D^2$ Analysis

It consists of the following important steps.

**(1)** Computation of $D^2$ values and testing their significance. The significance of $D^2$ values is tested against the table value of chi-square for ρ degrees of freedom, where ρ is the total number characters included in the study. If the calculated value of $D^2$ is higher than table value of chi-square, it is considered significant and vice versa.

**(2)** Finding out the contribution of individual character towards total divergence.

**(3)** Grouping of different genotypes into various clusters.

**(4)** Estimation of average distance at (a) intra-cluster and (b) inter-cluster levels

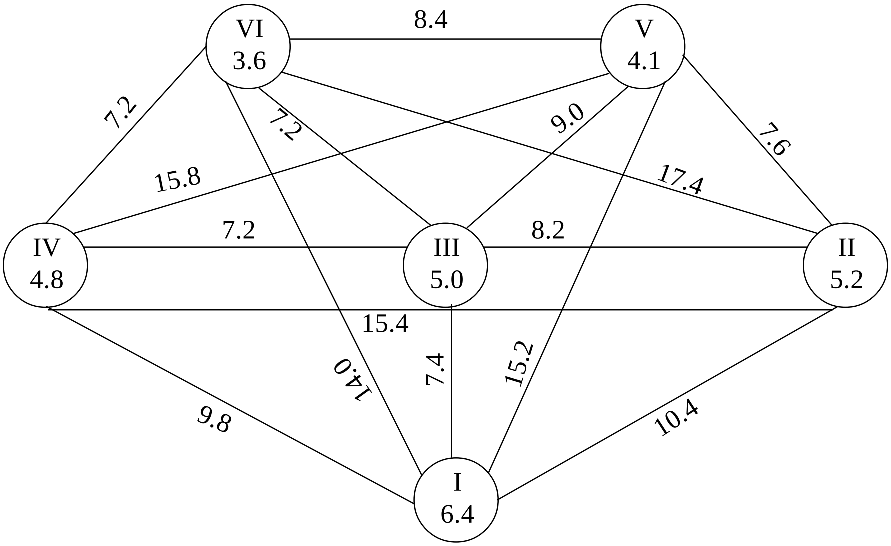
**(5)** Construction of cluster diagram.

In $D^2$ analysis a diagram is constructed with the help of $D^2$ values which is known as cluster diagram. The square roots of average intra and inter cluster $D^2$ values in the construction of cluster diagram. This diagram provides information on the following aspects:

**Average intra and inter-cluster $D^2$ and D values for 30 genotypes and 11 characters in blackgram**

| Clusters | I | II | III | IV | V | VI |
|---|---|---|---|---|---|---|
| I | 40.96<br>(6.40) | 108.60<br>(10.40) | 64.06<br>(7.40) | 96.04<br>(9.8) | 231.04<br>(15.20) | 196.00<br>(14.00) |
| II | | 27.04<br>(5.20) | 67.24<br>(8.20) | 237.16<br>(15.40) | 57.76<br>(7.60) | 207.36<br>(14.40) |
| III | | | 25.00<br>(5.00) | 51.84<br>(7.20) | 81.00<br>(9.00) | 51.84<br>(7.20) |
| IV | | | | 16.84<br>(4.80) | 249.64<br>(15.80) | 51.84<br>(7.20) |
| V | | | | | 16.81<br>(4.10) | 70.56<br>(8.40) |
| VI | | | | | | 12.96<br>(3.60) |

***(The values in parenthesis are square roots of $D^2$ values.)***




**Cluster Diagram**

(i) This depicts the genetic diversity in a easily understandable manner.

(ii) The number of clusters represent the number of groups in which a population can be classified on the basis of $D^2$ analysis.

(iii) The distance between two clusters is the measure of the degree of diversification, greater the distance between two clusters, greater the divergence and vice versa.

(iv) The genotypes falling in the same cluster are more closely related than those belonging to another cluster. In other words, the genotypes grouped together in one cluster are less divergent than those which are placed in a different cluster.

(v) It provides information about relationship between various clusters.

In selection of parents on the basis of $D^2$ values three important points should be taken into consideration. These points are : (1) the relative contribution of each character to the total divergence, (2) the choice of cluster with maximum genetic distance, and (3) the selection of one or two genotypes from such clusters. Other characters, like disease resistance, earliness, quality, etc., should also be considered. Crossing of the lines thus selected., in a dialed fashion may yield some useful sergeants. In plant breeding, $D^2$ techniques has been widely used in assessing the variability in crops like maize, sorghum, pearl millet, wheat, linseed, cotton, tobacco, alfalfa and brassica (Murty and Arunachalam 1965).

**3. Merits and demerits**

**(1) Merits** : (i) This technique helps in the selection of genetically divergent parents for their exploitation in hybridization programmes. (ii) $D^2$ statistics measures the degree of diversification and determines the relative proportion of each component character to the total divergence. (iii) It measures the forces of differentiation at two levels, i.e., intra cluster and inter-cluster levels. (iv) It Provides reliable estimates of genetic divergence and a large number of Germplasm lines can be evaluated at a time for genetic diversity by this technique.

**(2) Demerits** : (i) The calculation of $D^2$ statistics is difficult as it involves estimation of variances and covariance in its estimation. (ii) Since this technique is based on second order statistics, the estimates are not statistically very robust. (iii) The $D^2$ technique cannot be applied to unreplicated data.

Thus, it can be summarized that $D^2$ statistics measures, the forces of differentiation at intra and inter cluster levels and determines the relative contribution of each component trait to the total divergence. The clusters which are separated by the largest statistical distance show the maximum divergence.

## 17.3 Glossary

**Multivariate Analysis Technique** : The technique which simultaneously analyse more than two variable on a sample of observations is categorized as multivariate technique.

**Multiple Correlation** : It is joint effect of two or more independent variables on a dependent variable which is denoted by R

**Coefficient of Multiple Determination** :The square of multiple correlation coefficient (R) is the coefficienet of multiple correlation ($R^2$). It measures joint contribution of independent variable in explaining variability in dependent variable (Y).

**Discriminate Function** : A statistical approach used in the construction of selection index.

**$D^2$ Statistics** : A biometrical technique used for the study of genetic divergence in a population.

**Qualitative Trait** : A character showing discontinuous variation and governed by Oligogenes.

**Quantitative Trait** : A character showing continuous variation and governed by polygenes.

## 17.4 Summary

In this chapter the multivariate analysis techniques are described in short. Also the guidelines for statistical analysis are given for these techniques.

The multiple regression and multiple correlation techniques are useful for determining the contribution of each variable in the dependent variable and also to know the joint contribution of the independent variable in the dependent variable. Also we can know the percent variation explained by the independent variables under study. By this techniques, we can also select the important independent variables in the study. The discriminate function technique is also important multivariate techniques which involves development of selection criterion on a combination of various characters and helps the researcher in indirect selection of genetic improvement on yield. The $D^2$ statistics techniques is extensively used technique in plant breeding and genetics for the study of genetic divergence in the various breeding material. All these multivariate technique have also been used extensively in the other fields

## 17.5 Exercises for Practice

**Answer the following question in 200 words each.**

1. What is multiple regression ? Explain its application in agricultural and social sciences research.
2. Explain the analytical technique of multiple regression and multiple correlation.
3. Describe the various steps involved in the assessment of variability through $D^2$ statistics and give merits and demerits of $D^2$ analysis.
4. Define selection index and describe various types of selection indices with steps involved in computation of selection index.

# Unit 18 : Path, Content and Factor Analysis

***Index***



## 18.1 Introduction

In this chapter we shall study path, content and factor analysis. The main purpose of studying these techniques is given below. The concept of path analysis was originally developed by Wright in 1921, but the technique was first used for plants by Deway and Lu in 1959. Path coefficient analysis is simply a standardized partial regression coefficient, which split the correlation coefficient into the measures of direct and indirect effect. It measures the direct and indirect contribution of independent variables on dependent variable.

The content analysis consists of analyzing the contents of documentary materials such as books, magazines, newspapers and the contents of all other verbal materials which can be either spoken or printed.

The factor analysis is often used multivariate technique of research studies, specially pertaining to social and behavioral sciences. It is a technique applicable when there is a systematic interdependence among a set of observed or manifest variables and the researcher is interested in finding out something more fundamental or latent which creates this commonality.

**After the study of this unit, you will be able to know and understand :**

- Direct and indirect contribution of independent variable on dependent variable.
- Use of content analysis in analyzing the printed or spoken material.
- The analytical procedure of factor analysis and its applications.

## 18.2 Content

### 18.2.1 Path Analysis

The concept of path analysis was originally developed by Wright in 1921, but the technique was first used for plant selection by Dewey and Lu in 1959. Path coefficient analysis is simply a standardized partial regression coefficient which splits the correlation coefficient into the measures of direct and indirect effects. It measures the direct and indirect contribution of independent causal variables on dependent variable. For example in black gram, grain yield ($x_5$) is affected by number of primary branches ($X_1$), secondary branches ($X_2$) pods per plant ($X_3$) and seeds per pod ($X_4$). The path analysis reveals, whether the association of these characters with yield is due to their direct effect on yield or is a consequence of their indirect effects via other component characters.

**1. Types of Path Coefficients**

The path coefficient analysis is expressed in three ways, viz, (i) phenotypic (ii) genotypic (iii) environmental paths. The phenotypic path splits the phenotypic correlation coefficients into the measures of direct and indirect effects, while the genotypic path is estimated from the genotypic correlation coefficients. The environmental path is based on the estimates of environmental correlation coefficients. In plant breeding experiments, the first two types of paths are in common use.

## 2. Comparison of Correlation and Path Coefficient Analysis

| *Correlation Analysis* | *Path Coefficient Analysis* |
|---|---|
| 1. It measures the association between two or more variables. | 1. It measures the cause of association between two or more variable |
| 2. Analysis is based on variances and covariances | 2. Analysis is based on all possible simple correlations |
| 3. Does not provide information about direct and indirect effects of independent variable on the dependent variables | 3. It provide information about direct and indirect effects of independent variables on dependent variable. |
| 4. Does not provide estimate of residual effect. | 4. Provides estimate of residual effect |
| 5. Based on the assumptions of linearity and additivity. | 5. Also based on the assumptions of linearity and additivity |
| 6. Helps in determining yield components. | 6. Also helps in determining yield components. |

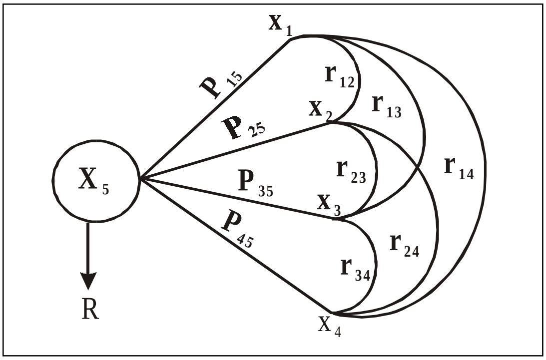
## 3. Computation of Path Coefficients

Path coefficient analysis is carried out using the estimates of correlation coefficients. Simple path can be worked out from both unreplicated and replicated data. But phenotypic, genotypic and environmental paths can be estimated from replicated data only. Computation of path coefficients from the replicated data involves the following steps.

1. Estimation of variance and covariances for all the characters and their combination, respectively.
2. Calculation of all possible simple correlation coefficients among independent and dependent variables, which is equal to $\frac{(n-1)}{2}$ where 'n' is the number of variables.
3. Then path analysis is carried out according to the procedure described by Dewey and Lu (1959). It consists of three steps, viz., calculation of (a) direct effects,(b) indirect effects, and (c) residual effects. For worked example please refer Singh and Chaudhary, 1985.

### (a) Direct Effects

Before estimation of direction of direct effects a line diagram known as path diagram is constructed. It is constructed with the help of simple correlation coefficients among various characters under study. The dependent variable, say yield is kept at one side and all the independent variables on the other side. This diagram is useful in several ways as given below.




**Path Diagram**

Path diagram constructed using the correlation coefficients among yield ($X_5$) and four of its components traits. 'r' denote the residual effects.

(1) It helps in setting up of simultaneous equations as given below, which are used for the estimation of direct effects.

(2) It depicts the cause and effect situation in a simple manner and makes the presentation of results more attractive.

(3) It helps in understanding the direct and indirect contribution of various independent variables towards a dependent variable.

(4) It also depicts the association between various characters.

$r_{15} = p_{15} + r_{12}\,p_{25} + r_{13}\,p_{35} + r_{14}\,p_{45}$

$r_{25} = r_{12}\,p_{15} + p_{25} + r_{23}\,p_{35} + r_{24}\,p_{45}$

$r_{35} = r_{13}\,p_{15} + r_{23}\,p_{25} + p_{35} + r_{34}\,p_{45}$

$r_{45} = r_{14}\,p_{15} + r_{24}\,p_{25} + r_{34}\,p_{35} + p_{45}$

Where $r_{12}$,$r_{13}$, $r_{14}$, etc. are the estimates of simple correlation coefficients between variables $x_1$ and $x_2$, $x_1$ and $x_3$, $x_1$ and $x_4$ etc. and $p_{15}$, $p_{25}$ $p_{35}$ and $p_{45}$ are the estimates of effect of variables $x_1$,$x_2$,$x_3$, and $x_4$ respectively on the dependent variable $x_5$ (in this case yield).

Every component character will have a direct effect on yield. In addition, it will also exert indirect effect via other component characters. The direct effect or contribution of various causal factors is estimated by solving the simultaneous equations, after putting the values of simple correlation coefficients. In other words, the estimates of direct effects, viz., the values of $p_{15}$, $p_{25}$, and $p_{45}$ are obtained. The above equations are solved by the process of elimination.

### (b) Indirect Effects

The effects of independent character on the dependent one via other independent traits are known as indirect effects. These effects are computed by putting the values of correlation coefficients and those of direct effects as follows.

Indirect effect of primary branches ($X_1$) via

Secondary branches ($X_2$) =$r_{12}$ $p_{25}$

Pods per plant ($X_3$) = $r_{13}$ $p_{25}$

Seed per pod ($X_4$) = $r_{14}$ $p_{45}$

Similarly, the indirect effects of secondary branches ($X_2$) will be as follows:

via primary branches ($X_{1)}$ = $r_{12}$ $p_{15}$

via pods per plant ($X_3$) = $r_{23}$ $p_{35}$

via seeds per pod($X_4$) = $r_{24}$ $p_{45}$

The indirect effects of other component trait, e.g. pods per plant ($X_3$) and seed per pod ($X_4$) may be computed in a similar fashion.

### (c) Residual Effect

In plant breeding, it is very difficult to have complete knowledge of all component traits of yield. The residual effect permits precise explanation about the pattern of interaction of other possible components of yield. The residual effect is estimated with the help of direct effects and simple correlation coefficients is given as follows:

$1 = P^2 R_5 + p_{15} r_{15} + p_{25} r_{25} + p_{35} r_{35} + p_{45} r_{45}$

Where, $P^2 R_5$ is the square of residual effect

**(d) Interpretation of Results**

(1) If the correlation between yield and a character is due to direct effect of a character, it reveals true relationship between them and direct selection for this trait will be rewarding for yield improvement.

(2) If the correlation is mainly due to indirect effects of the character through another component trait, indirect selection through such trait will be effective in yield improvement.

(3) If the direct effect is positive and high, but the correlation is negative, in such situation direct selection for such trait should be practiced to reduce the undesirable indirect effect.

(4) If the value of residual effect is moderate or high, it indicates that besides the characters studied, there are some other attributes which contribute for yield.

**4. Merits and Demerits**

***(i)*** ***Merits*** : (1) This technique provides information about the cause and effect situation and helps in understanding the cause of association between two variables. (2) It is quite possible that a trait showing positive direct effect on yield may have a negative indirect effect via other component traits. Path analysis permits the examination of direct effects of various characters on yield as well as their indirect effects via other component traits. Thus through the estimates of direct and indirect effects, it determines the yield components. (3) It provides basis for selection of superior genotypes from the diverse breeding populations.

***(ii)*** ***Demerits :*** (1) Its computation is difficult and inclusion of many variables makes the computation more complicated. (2) Path analysis is designed to deal with variables having additive effects. Its application to variables having non-additive effects may lead to wrong results.

In brief Path coefficient analysis simply a standardized regression coefficient which splits the correlation coefficient into the measures of direct and indirect effect.

It measurers the direct and indirect contribution of independent variables on dependent variable and thus helps breeder in determining the yield components. Path coefficients are of three types, viz., phenotypic, genotypic and environmental paths. The first two are in common use in plant breeding. The computation of path consists of three main steps, viz., calculation of direct effects, indirect effects and residual effect.

## 18.2.2 Content Analysis

Content analysis consists of analysing the contents of documentary materials such as books, magazines, newspapers and the contents of all other verbal materials which can be either spoken or printed. Content-analysis prior to 1940's was mostly quantitative analysis of documentary materials concerning certain characteristics that can be identified and counted. But since 1950 content-analysis is mostly qualitative analysis concerning the general importance or message of the existing documents. "The difference is somewhat like that between a casual interview and depth interviewing". **Bernard Berelson's** name is often associated with the latter type of content-analysis. Content analysis measures pervasiveness and that is sometimes an index of the intensity of the force.

The analysis of content is a central activity whenever one is concerned with the study of the nature of the verbal materials. A review of research in any area, for instance, involves the analysis of the contents of research articles that have been published. The analysis may be at a relatively simple level or may be a subtle one. It is at a simple level, when we pursue it on the basis of certain characteristics of the document or verbal materials that can be identified and counted (such as on the basis of major scientific concepts in a book). It is at a subtle level when researcher makes a study of the attitude, say of the press towards education by feature writers.

## 18.2.3 Factor Analysis

Factor analysis is by far the most often used multivariate technique of research studies, specially pertaining to social and behavioural sciences. It is a technique applicable when there is a systematic interdependence among a set of observed or manifest variables and the researcher is interested in finding out something more fundamental or latent which creates this commonality. For instance, we might have data, say, about an individual's income, education, occupation and residential area and want to infer from these some factor such as social class which summarizes the commonality of all the said four variables. The technique used for such purpose is generally described as factor analysis. Factor analysis, thus, seeks to resolve a large set of measured variables in terms of relatively few categories, known as factors. This technique allows the researcher to group variables into factors based on correlation between variables and the factors so derived may be treated as new variables often termed as latent variables and their values derived by summing the values of the original variables which have been grouped into the factor.

The meaning and names of such new variable are subjectively determined by the researcher since the factors happen to be linear combinations of data of each co-ordinates of each observation or variable is measured to obtain what are called factor loadings. Such factor loadings represent the correlation between the particular variable and the factor, and are usually placed in a matrix of correlations between the variable and the factors.

The mathematical basis of factor analysis concerns a data matrix also termed as score matrix, symbolized as S. The matrix contains the scores of N persons of k measures. Thus $a_1$ is the score of person 1 on measure $a_1$, $a_2$ is the score of person 2 on measure $a_1$ and $k_N$ is the score of person N on measure k. The score matrix then take the form as shown following :

Score Matrix OR S-Matrix

Measures (variables)

| | | a | $b_1$ | $c_1$ | k |
|---|---|---|---|---|---|
| Persons | 1 | $a_1$ | $a_2$ | $a_3$ | $k_1$ |
| (Objects) | 2 | $a_2$ | $b_2$ | $c_2$ | $k_2$ |
| | 3 | $a_3$ | $b_3$ | $c_3$ | $k_3$ |
| | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| | n | $a_n$ | $b_n$ | $c_n$ | $k_n$ |

It is assumed that scores on each measure are standardized, i.e. $x_i = (x_1 - \bar{x}_i)/\sigma i$ This being so, the sum of scores in any column of the matrix, S, is zero and the variance of scores in any column is 1. Then factors (a factor is any linear combination of the variable in a data matrix and can be stated in a general way like: $A=W_a a + W_b b+......+w_k k$) are obtained by any method of factoring. After this, we work out factor loading i.e., factor-variable correlations. Then communality, symbolized as $h^2$, the eigen values and the total sum of squares are obtained and the results interpreted. For realistic result, we resort to the technique of rotation, because such rotations reveal different structures in the data. Finally, factor scores are obtained which help in explaining what the factors mean. They also facilitate comparison among groups of items as groups. With factor scores, one can also perform several other multivariate analyses such as multiple regression, cluster analysis, multiple discriminant analysis, etc.

## 1. Important Methods of Factor Analysis

There are several methods of factor analysis, but they do not necessarily give same results. As such factor analysis is not a single unique method but a set of techniques. Important methods of factor analysis are:

(i) Centroid method (Thurstone)

(ii) Principal components method (Hotellings)

(iii) Maximum likelihood method.

## 2. Various Terms Used in Factor Analysis

**(i) Factor :** A factor is an underlying dimension that account for several observed variables. There can be one or more factors, depending upon the nature of the study and the number of variables involved in it.

**(ii) Factor Loadings** : Factor loadings are those values which explain how closely the variables are related to each one of the factors discovered. They are also known as factor variable correlations. In fact factor loading works as a key to understanding what the factors means. It is the absolute size rather than the signs, plus or minus of the loadings that is important in the interpretation of a factor.

**(iii) Communality** : It is symbolized as $h^2$, shows how much of each variable is accounted for by the underlying factor taken together. A high value of communality means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under: $h^2$ of $i^{th}$ variable =($i^{th}$ factor loading of factor A)$^2$ +( $i^{th}$ factor loading of factor B)$^2$.

**(iv) Eigen Value (or Latent Root) :** When we take the sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen value or latent root. Eigen value indicates the relative importance of each factor in accounting for the particular set of variables being analysed.

**(v) Total Sum of Squares :** When eigen values of all factors are totaled, the resulting value is termed as the total sum of square. This value, when divided by the number of variables (involved in a study), results in an index that shows how the particular solution accounts for all the variables taken together represent.

**(vi) Rotation :** Rotation, in the content of factor analysis, is something like staining a microscope slide. Just as different stains on it reveal different structures in the tissue, different rotation reveal different structures in the data. Though different rotations give results that appear to be entirely different, but from a statistical point of view, all results are taken as equal.

**(vii) Factor Scores :** Factor score represents the degree to which each respondent gets high scores on the group of items that load high on each factor. Factor scores can help explain what the factor mean. With such scores, several other multivariate analyses can be performed.

## 18.3 Glossary

**Biometrical Techniques :** Various statistical procedures employed in biometrical genetics.

**Path Analysis :** A standardized partial regression coefficient which splits the correlation coefficient into the measures of direct and indirect effect and also measures the direct and indirect contribution of each independent variable on the dependent variable.

**Direct Effect :** The effect of an independent character on the dependent character.

**Indirect Effect :** The effect of an independent character on the dependent character via. other independent traits.

**Correlation :** The ratio of appropriate co variance to the product of two standard deviations.

**Residual Effect :** The residual effect permits precise explanation about the pattern of interaction of other possible components of dependent variable.

**Partial Correlation :** The correlation coefficient between two variables eliminating the effect of other variables.

**Phenotypic Variation :** The total or observable variation in a population. It consists of genotypic and environmental component .

**Genetic Variation :** The inherent variability. It consist of additive dominance and epistemic components

**Factor :** A factor is an under lying dimensions that account for several observed variables. There can be one or more factors, depending upon the nature of the study and number of variables involved in it.

**Content Analysis :** It consists of analysis of the content of documentary materials.

## 18.4 Summary

In this chapter the path, content and factor analysis is described in brief. The analytical procedure is also given. These analytical techniques are very much useful to the researchers. By path analysis the direct and indirect contribution of independent variables on dependent variable can be estimated. This technique will help the researcher in determining the yield components or the components of dependent variable under study.

The content analysis technique is also useful for analyzing in documentary materials such as books, magazines, newspapers and the content of all other verbal materials which can be either spoken or printed.

The factor analysis will help the researcher in research studies related to behaviors sciences.

## 18.5 Exercises for Practice

**Answer the following question in 200 words each**.

1. Define path analysis and path coefficient and describe its merits and demerits.
2. Describe different types of path and the applications of path analysis.
3. Explain path diagram and its importance in path analysis.
4. What is content analysis and factor analysis ? Describe various basic terms used in factor analysis

# Unit 19 : Packages for Data Analysis

***Index***



## 19.1 Introduction

The computer has immense power of information and data processing and dissemination, storage and retrieval. Now a days, through computer networking with use of modem, computer has become extremely useful for communication. The tremendous speed of computer is another advantage. Initially mainframe computers were used in data analysis which required specially trained manpower and one has to depend on such trained personnel for the processing of data or information. During the last ten years computer technology has undergone a revolution. With the advent of powerful personal computers and user friendly software, it has become possible for everyone to work on computers sitting right on their desks. Thus, the opportunities for computer application in all the field have further increased.

The recent advances in computer and communication technology have made computer hardware and software more affordable and have resulted in faster movement of information and its utilization. Several Government and Non-Government agencies are engaged in development of database, software and packages. Software needs to be appraised on various criteria, which include statistical, computational and commercial considerations. Keeping these considerations, new and revised packages, for data analysis are being released continuously and it is difficult to provide and upto date review of them all here. However, some of the established packages are described in this chapter.

**After the study of this unit, you will be able to know and understand :**

- λ Important software packages used for data analysis
- λ How to select the software packages for data analysis.
- λ Details of SPSS package.

## 19.2 Content

### 19.2.1 Need of Software Packages for Data Analysis

A first requirement for any computer program for statistical analysis is that it should give an analysis that conforms with good statistical practice and that results should be correct. Also important are ease to use and the presentation of results in a manner that allows ready assimilation. The writing of programmes with these qualities requires both a through understanding of the analysis to be performed and a high level of computing skill. The necessary combination of skills is most likely to be found when groups of individuals with a high level of statistical and computing expertise co-operate to write programs intended for wide use. For the case of single individual, the need to write special programs, or even use algorithms is slowly diminishing, as the wider and wider range of software packages are becoming available in the market.

Statistical packages now provide the most popular way of handling statistical analyses. A package is distinguished from a library of subroutines because it is complete in itself. It has its own command language for specifying the form of data input, the manipulations required in the data and the analysis or analyses required. An ideal is that the command language should allow the statistician to describe the problem in appropriate statistical terms, leaving to the computer package, the task of interpreting and acting upon this description.

Packages vary widely both in quality and in what they will do, and the choice between them may not be easy. Some have been written by expert and others have not. Some print out plenty of warning, error and help messages, but others do not and may indeed go ’happily’ producing meaningless results. Some packages are written for one small area of methodology while others are more general. Some are written for expert users, while others are intended for statistical novices. Some packages allow completely automatic analysis of data, where the analyst abdicates all responsibility to the computer, but interactive analyses are usually preferable.

Software needs to be appraised on various criteria which include statistical, computational and commercial considerations. Keeping these considerations, new and revised packages are being released continuously and it is very difficult to provide an up-to-date review of them all here. Now a days the well established packages, written originally for mainframe computers, have mostly become available for microcomputers as well. Note that most packages are now available in a Windows version which is generally easier to use. Also note that the comments made below at the time of writing may become outdated as packages are improved. In this regard it is necessary that one should keep up-to-date information buy referring to reviews of computer software in various journals such as American statistician and selected computing journals, the need has arisen for objective evaluation and comparison of these systems.

### 19.2.2 Software Packages Used for Data Analysis

Choosing of the appropriate hardware and software for any individual /office/ organization is critical. Software requirement must be assessed on the basis of the nature of applications. Some of the important packages which are in common use are given below.

**(1)** **MINITAB** : It is an interactive, command- driven package which covers such topics as exploratory data analysis, Significance tests, regression and time-series analysis. It is very easy to use and is widely employed by both commercial, academic and teaching institutions. MINITAB is patterned in part on the OMNITAB statistical system, which keeps a worksheet, where data are laidout in columns. In MINITAB a column is just as appropriately viewed as a vector of values of one variate. Data that are to be placed in the worksheet may be included preceded by a READ or SET command which gives details of where the data are to be placed, among the MINITAB commands. Commands appear one in a line. A sequence of MINITAB commands may be stored together, under a single name, as a macro. EXEC, followed by the macro name then causes execution of the whole sequence of commands.

**(2) GENSTAT :** It is a statistical programming language which allows the user to write programs for a wide variety of purposes. It is particularly useful for the analysis of design of experiments, the fitting of linear models including regression and generalised linear models. It also covers most multivariate techniques, time-series analysis and optimization. It allows much flexibility in inputing and manipulating data. Programs can be constructed from macros which are blocks of statements for a specific task. A library of macros is supplied. The package is designed for the expert statistician and is not user-friendly.

**(3) BMDP :** It is a suite of programmes which covers most statistical analysis from simple data and display to multivariate analysis. BMDP function as a collection of separate analysis programmes but with a command language which is similar across all programmes. A wide range of output options are in most cases available and output is admirably well laid out and input facilities are less flexible. If extensive manipulation of data is required prior to an analysis, a separate package must be used. BMDP is well suited to use in an interactive computing environment where BMDP programmes form part of a collection of utilities for data manipulation and editing, with easy means for using the output from one utility as input to another. This comprehensive package is good for the professional statistician and has good user support but can be rather difficult to learn.

**(4) SPSS** (**Statistical Package for the Social Sciences) :**. This is a comprehensive package, is probably used more than any other package, but mainly by non-expert users. Some statistician view the package with some reserve. This is comprehensive package for Social Sciences.

**(5) GLIM (Generalised Linear Interactive Modelling**) : This software packages has been developed in U.K. by the Royal Statistical Society Working Party. It is an interactive command-driven package which is primarily concerned with fitting generalized linear models. This means that it covers regression, ANOVA, probit and logit analysis and log-linear models. The user must specify the error distribution, the form of the linear predictor and the link function, the package is powerful but requires considerable statistical expertise, is not user-friendly, and can produce output which is difficult to interpret (e.g. estimates of main effects are given in a non-standard form). It can do things which are difficult to perform using other packages. Macros are also permissible here. The names of variates or factors in GLIM may be of arbitrary length, but characters after the fourth are ignored.

**(6) SAS (Statistical Analysis System)** : This programming language is widely used, both in the USA and UK. It is suitable for the more advanced statistician. It is composed of basically three sets of statements which can create data sets, run statistical analysis and print data. The first type of statement are called DATA statements while the other types are called PROC's (short for procedures). The interesting feature of SAS is that it has its own IML (Interactive Matrix Language) which includes many functions and subroutines.

**(7) Spar 1 :** Spar 1 is a statistical package for agricultural research, developed at Indian Agricultural Statistical Research Institute (IASRI), New Delhi. The package has been developed for the statistical analysis of experimental data in plant breeding and genetics. The package is menu driven and works in user friendly manner. Help is available at all the stages while navigating through the system.

**(8) Other Packages :** There are numerous other general and more specialized packages are available and all of them can not be covered here. To name a few MSTATC, STATGRAPHICS, SX, MICROSTAT, STRATA, PC-CARP,EXACT, STATLAB, SIGMAPLOT and FREELANCE are some useful specialist graphics packages.

There are number of packages which are mathematically oriented. They include MATLAB, MAPLE and MATHEMATICA. As well as some data-analysis procedures, they enable the user to work with a variety of mathematical functions, carry out numerical analysis and matrix algebra, plot 2-D and 3-D functions and graphs, etc.

One class of packages not mentioned so far are spreadsheet packages. They allow easy manipulation of data in the form of two-or three-way table of numbers and also allow some simple statistical analysis to be carried out on them. They are widely used, and the best known packages are LOTUS 1-2-3, EXCEL, FOXRPO, REFLEX and QUATTRO PRO.

## 19.2.3 SPSS (Statistical Package for Social Science )

### 1. Important Features of SPSS

Almost all the software's prepared for data analysis are MENU driven and can be operated easily with the instructions given or manuals supplied with it.

Statistical packages for the Social Science requires some training. The Important features of SPSS are given below for ready reference. SPSS Inc. offers a comprehensive and flexible set of software tools for data entry, data management, statistical analysis and presentation. SPSS software operates on models of major computers under a wide variety of operating systems. SPSS integrates complex data and file management, statistical analysis and reporting functions. With menu interface and straight forward English syntax commands, one does not need any prior programming experience. The package includes facilities for adding user defined specialized procedures and a variety of seamless interfaces to leading data base management systems. Other modules are

(i) SPSS Graphic (ii) SPSS Table
(iii) SPSS Categories (iv) SPSS Trends
(v) SPSS Advance Statistics (vi) SPSS Data Entry

### 2. Goals and Objectives

The information given here will help the user how to run SPSS on one of the operating system for which the software is available. It shows how to use SPSS commands to manage and analyze data with few examples. It is restricted to brief operational instruction for various commands. For detailed study users are requested to refer the manual supplied with SPSS package.

## 3. System Requirements

(i) Pentium IV PC
(ii) WINDOW or UNIX operating system
(iii) Memory 256 MB
(iv) Hard disc with atleast 40 GB
(v) Colour Monitor

## 4. Importance of SPSS

(i) It is easy to learn and use
(ii) It includes a full range of data management system and editing tools
(iii) It provides in depth statistical capabilities
(iv) It offers complete plotting, reporting and presentation on features.

## 5. Brief Descriptions of SPSS Under MS-DOS

For SPSS package under MS operating system, after switching ON a PC system, the display desktop icons.

### *Executing SPSS System*

Click ‘Start’ then ‘Programe’ and then select ‘SPSS’ icon.

It executes SPSS system and loads main menu. Now, press F1 Key, menu will be displayed at bottom of screen as mentioned below

| Info | Review | Help | Varlist | File List | Glossary | Menu | Hlp Off |

Using arrow key, appropriate module can be executed according to the requirement of user. Following mentioned command should be noted.

| *Press this key* | *To do this* |
|---|---|
| Esc | Remove Last temporary paste and move up one level |
| Alt M | Remove Menus |
| Alt E | Switch  Edit Menu |
| Alt Esc | Jump to Main Menu |
| Alt T | Get Typing Window |
| Alt X | Switch between standard and extended menus |
| F1 | Manager help and Menus |
| F2 | Switch, change size, zoom |
| F3 | Insert file, edit different file |
| F4 | Insert, delete, undelete |
| F5 | Find text, replace text |
| F6 | Area, output page, after last line executed |
| F7 | Mark/unmark lines, rectangle |
| F8 | Copy, move, delete, round numbers |
| F9 | Write area on file, delete file |
| F10 | Run command from cursor or marked area, exit from SPSS |

**How to use SPSS editor**

After executing SPSS switch to edit insert window and type instruction or data using following commands.

**Cursor Movement Commands**

using arrow key, user can move cursor

Ctrl Home

Ctrl Pgup

Ctrl Pgdn

Ctrl End

**Delete Commands**

Del Key Delete character at cursor or join lines

**Insert Command**

Ins key Switch between insert and over type modes

Enter key Split line at cursor or go to next line

For further analysis anddetail of processing information will be available in SPSS reference manual.

**SPSS Session :** It is a session when one execute a sequence of SPSS commands.

Data definition commands provide information about the variables and their location in the data file.

Data transformation commands are used to create new variables, modify existing ones and select subsets of cases for analysis. Procedure commands indicates what statistics, reports or tables are to be produced.

Data definition and transformation commands are generally not executed until a command that reads SPSS are not specified.

An SPSS session must begin with at least one data definition command that either defines the data of SPSS or specifies an SPSS system file, which already contain data definition.

SPSS commands can be processed in several different ways :

One can create a file of SPSS commands.

One can run commands one at a time.

One can build command file with the help of a command generator.

One can run command within the file during the SPSS session.

## 6. Illustrations of SPSS Package

Examples with solutions are provided in this short note to process different kinds of statistical analysis using SPSS Packages. Users are requested to use the reference manual supplied by SPSS.

### *Example 1*

The following marks have been obtained in three papers of Agricultural Communications in examination by 12 students. Compute descriptive statistics using SPSS

Marks obtained by twelve students. (100)

Paper A – 36,56,41,46,54,59,55,51,52,44,37,59

Paper B – 58,54,21,51,59,46,65,31,68,41,70,36

Paper C – 65,55,26,40,30,74,45,29,85,32,80,39

also calculate and list marks secured by each students, when using relation

Y=3*A + 2*B +1.5*C

***Solution :***

In this example, there are 12 cases and three variables named as A,B and C.

Syntax : Descriptives (Variables=) varname/statistics all.

As explained above execute spss, diplays input window and switch to edit mode by pressing key "Alt E" and start typing following lines

```
set list = 'decrip.res'.
data list free /ABC.
begin data
36 58 65
56 54 55
41 21 26
46 51 40
54 59 30
59 46 74
55 65 45
51 31 29
52 68 85
44 41 32
37 70 80
59 36 39
end data.
descriptive/variables all / statistics all.
compute Y=3*A+2*B+1.5*C
list A B C Y
```

After typing above mentioned lines, save by pressing F9 key, writing unique and meaningful file name. Bring cursor at the beginning and press F10 key and run. The displayed results can be stored in the file which can be printed after exit from SPSS.

***Example-2***

Compute a Simple Linear correlation Coefficient between Soluble Protein Nitrogen (X2) and Total Chlorophyll (X2) in the Leaves of Rice Variety IR8

| Sample Number | Soluble Protein N (mg/leaf) (X1) | Total Chlorophy (mg/leaf) (X2) |
|---|---|---|
| 1 | 0.84 | 0.55 |
| 2 | 1.24 | 1.24 |
| 3 | 2.10 | 1.56 |
| 4 | 2.64 | 2.52 |
| 5 | 1.31 | 1.64 |
| 6 | 1.22 | 1.17 |
| 7 | 0.19 | 0.04 |

***Solution :***

In this example there are 7 cases, two variables named as X1 and X2.

Syntax : Correlation (variables=) varname WITH Varname.

Switch on to edit mode and after executing SPSS as mentioned above and type following lines.

data list free /X1 X2.

0.84 0.55

1.24 1.24

2.10 1.56

2.64 2.52

1.31 1.64

1.22 1.17

0.19 0. 04

end data.

list

**correlation/variable X1 with X2.**

**Run and store these statement lines as explained earlier in example one.**

## 19.3 Glossary

**Program** : A set of instructions written in the specific sequence for the computer to accomplish a given task.

**Computer :** An electronic device that can transmit, store and process information or data. It is popular because it has speed, accuracy and large storage.

**Floppy Disc** : A thin plastic disk coated with magnetic oxide and used for information storage.

**SPSS :** Statistical Package for Social Sciences, used extensively for data analysis.

**MS DOS :** Microsoft Disk operating system.

**Package :** Software program written in a computer language with specific purpose.

**Hardware :** Comprises of all physical components of the computer system i.e. key board, monitor, processor, hard disk floppies etc.

**Software :** Comprises of various computer instructions which controls the operation of computer system and help the user e.g. operating systems, compilers, assemblers, utility programme and other application packages.

## 19.4 Summary

Computer technology is changing very fast with the advent of powerful personal computers and user friendly software's, it has become possible for everyone to work on computer and to analyse the complex data and formulae. In this chapter the some of the commonly used software packages are given along with their specific applications. The ready made packages can serve the purpose, however some times it is necessary to develop software packages according to our own need of data analysis.

The more sophisticated and costly software like SAS (Statistical Analysis System) which are widely used in USA and UK are now available in India. The SPSS (Statistical Package for Social Science) is also available which is quite comprehensive statistical package used for data analysis.

The packages like SAS and SPSS require some training for its fullest use. However, most of the software packages are user friendly and are menu drive.

## 19.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. What is software packages and explain various software packages used in statistical data analysis.
2. Describe SPSS package for statistical analysis of data in social sciences.
3. What is PC software and hardware ? How they are helpful in research investigation in agri-business management ?
4. Describe in detail how to select software packages for data analysis ?

# Unit 20 : Scientific Report Writing

***Index***



## 20.1 Introduction

The post graduate and research students are often required to present scientific assignments, case studies, seminars, dissertations and thesis in the form of written reports as a part of their degree programmes. The object of the presenter, a student or a researcher is to communicate such information to the targeted audience in a manner that the audience or reader understands perfectly. Research is a continuous process and through this process, several inventions, discoveries, ideologies, concepts are generated by scientists and research students. It is, necessary to publish such material in the form of research article, books or thesis etc. However, such valuable material remains in the laboratory or with the researcher, due to lack of knowledge of scientific report writing. The aim of this topic is to help the students in presenting their research material in scientific manner. Thesis and Dissertation, research papers, seminars, reports, assignments, symposium, project reports, scientific articles are most common scientific reports and hence emphasis has been given on them. It may be worth to mention here that knowledge of scientific and technical report writing is not just for completion of thesis or dissertation but is also useful in future professional development of a person.

**After the study of this unit, you will be able to know and understand :**

- Importance of scientific report writing.
- Guidelines for writing scientific reports like thesis, dissertation, research assignments and papers etc.
- Effective presentation of research material

## 20.2 Content

### 20.2.1 Scientific Writing

Scientific writing is writing about scientific subjects and about the subjects associated with it. It is planning the scientific information systematically, organizing skillfully and putting it in a logical sequence by following an effective style of communication and presentation.

A report which contains relevant, adequate and logically arranged scientific information presented in a specific format is in true sense a scientific report. The thesis is a scientific report and the efforts, skills put in preparing it, is a process of scientific writing. However, technical report writing is, writing about technical subjects and the associated areas.

The writing procedure and style of reporting however do not have significant differentiation between scientific and technical report. Further, it is important to note that the scientific report writing is presentation of scientific truth and not the opinion, views or personal feelings of the author. The scientific report writing is normally according to standard national and international norms of writing which we may not find in case of literally type or novels, stories etc.

The importance of scientific reports is relatively more, from further research point of view.

### 20.2.2 Types of Scientific Reports

Post graduate and research students and research scientists are generally come across with following types of reports.

(1) Thesis/Dissertation/Comprehension report

(2) Research papers/articles

(3) Review Papers

(4) Case studies

(5) Outline of Research Work (ORW)

(6) Synopsis of Research Work (SRW)

(7) Project reports

(8) Annual research reports

(9) Popular articles

Thesis or dissertation, outline or synopsis of research work and seminar reports are some of the common reports the students have to deal with. Therefore, emphasis in this topic has been given on these reports.

### 20.2.3 Thesis and Dissertation Writings

The post graduate students of most of the disciplines of Science are required to submit thesis in partial fulfillment of post graduate degree programme. Therefore, it is necessary for all the students to know the scientific way of thesis writing. When a students and his advisory committee are satisfied with the quality and amount of research work done, then the student should make a plan for writing his thesis. It is, therefore, necessary to prepare the outline of thesis. The format of thesis is given below :

#### 1. Format of Thesis

The researcher should present his thesis/dissertation in following format.

(i) Preliminary pages

(ii) Main Text

(iii) Bibliography

(iv) Appendices

In the above format some changes may be there according to provisions made by different Institutes or Universities.

**(a) Preliminary Pages**

The sequence of preliminary pages starting from the cover page should be as follows :

***(i) Cover :*** This is a cover of the bound thesis. This should include in the sequence the title of thesis, author, name of the department or faculty, name of the institute/University, place and the year in which the thesis is submitted.

***(ii) First Title Page :*** This should also include the Title, Purpose, Author, Department, Institute, Year of submission.

***(iii) Certificates :*** This should include (i) Candidate's Declaration, (ii) Certificate of Research Guide

***(iv)*** Acknowledgement by Student

***(v)*** Table of Contents

***(vi)*** List of Tables

***(vii)*** List of Figures

***(viii)*** List of Abbreviations

***(ix)*** Abstract

**(b) Main Text**

The main text may include following chapters and the details of these chapters are given below.

***(i) Introduction :*** The introduction should cover the importance of timeliness of the subject. The scope and nature of the material covered should be indicated. Attempts should be made to justify the selection of particular problems, its scientific and practical utility. It should be ended with list of major objectives. The introduction should generally be kept brief. The title 1. INTRODUCTION preferably be in the capital letters. The objectives of the study should be at the end with numbers 1, 2, 3,….,5

The general format is given below.

(1) Background information substantiated with relevant review of literature.

(2) Importance of study.

(3) Objectives of the study.

(4) Hypothesis or assumptions if any.

(5) Scope and limitations of the study.

**2. Review of Literature :** The title should be "REVIEW OF LITERATURE" on a new page. The purpose of review of literature is to get to know the reader, with earlier work done on the subjects. The Review should be exhaustive and complete. The extent and details of the review must be determined. The name of the author and year in bracket is commonly used for referring to the literature. The system of referring to the literature should be consistent for all main text. Now a days, with the *Internet* facility, one can get easily the upto date review of literature.

**Review of literature is useful in the following ways :** (1) Identification of areas of research already covered in a particular field of science. It helps in avoiding duplication and repetition of research. (2) The depth of the research that has been reached in the past. (3) Understanding the methodology used by the past researcher. (4) The findings or observations recorded by the past researchers with regard to a particular field of science.

**Review of literature involves following aspects :** (1) Identification of the sources of literature and reading. (2) Taking notes on the reference card. (3) Writing review of literature and theoretical orientation. (4) Citing references in the text at appropriate places. (5) Preparing the bibliography or list of references cited or literature cited

**3. Material and Methods :** This section can be divided into two sub-divisions. First describe the Material and their sources, it should be followed by Methods. Describe the procedure used in brief with sufficient details. State various methods in the sequences as they appear in results. The reference should be made to the literature when the method has been previously reported or standardized by other person.

The value and usefulness of the results or findings depends on approximate methodology and hence care should be taken to use appropriate material and methods for the study. The methodology chapter mainly deals with material and methods as follows :

***Material :*** The adequate information about the following aspects is necessary in these regards. Use of scientific and standard terminology is absolutely necessary.

(1) Input used for research such as chemicals, cultures media, seeds, fertilizers, insecticides, reagents, medicines, plants, primary data, secondary data etc.

(2) Machines, equipments, instruments etc. used

***Methods :*** While describing the methods, the details regarding the different methods, procedures, statistical techniques of analysis and design used for the study should be given in a logical sequence. It usually include the following aspects.

λ Variables, their definitions, measurement and their scoring procedures adopted.

λ The standard procedures, techniques and conditions used for investigation or experimentation and methods of analysis etc.

λ Sampling procedure, techniques of data collection used.

λ Statistical methods and formulae used for analysis of data

**4. Results and Discussion** : This chapter (Section) can also be divided into two sub-sections as Results and Discussions

These two sections are interrelated. This is the most important chapter in the thesis/research paper/research reports.

***Results :*** Results is a statement of facts based on the systematically analysed and tabulated data. Whereas the discussion is an explanation or justification to the results obtained on the basis of experiences and past research. The results provide factual statement of observations supported by statistical analysis. The results are presented in logical order in light of objectives of the study. For proper communication and understanding of the problem the results may be presented in the following way.

(i) Tabulation of data (Tables)

(ii) Interpretation based on tables and analysis

(iii) Supporting the results by graphical representation.

***Discussion :*** The discussion chapter deals with discussion on the findings of the study . Discussion in other words means finding the justification, probable reason to the observed results and supporting them with reviewed literature. Discussion mainly answers "Why ?" of the findings. It is, therefore, thoughtfully written and requires in depth knowledge, experiences, observations and past literature related to the subject of study.

The discussion is also a vital part of the study. It is, therefore, recommended to give separate identity for this chapter. However, students have choice to give it with results, if it is confident and appropriate. This chapter should include comments on the methodology, overall interpretations of the findings and predictions on scientific basis. This chapter should not repeat results in to-to but should discuss and justify if the results agree or dis-agree with those of previous workers. The background knowledge and the present findings should be used to predict the future line of work and modifications if any. The lacunae and suggestions in the work should be clearly indicated. The title of this topic should be DISCUSSION or RESULTS AND DISCUSSION

**5. Summary and Conclusion :** Summary should include importance of the study or purpose, the objectives in brief, and a brief summary of the findings of the study. In conclusion, a list of conclusions drawn from the study should be given in brief. The general format may be (i) Purpose (importance of study), (ii) Objectives of the study on brief, (iii) Methodology used in brief and (iv) Findings of the study with respect to objectives and broad Conclusions.

Summary is an abridged form of the main report. Therefore, it should be given thoughtfully in thesis/research papers. It may be noted that conclusions is not the repetition of the results. Tables, diagrams and figures should not be given in the conclusions. Summary should not be too lengthy and it should not exceed more than 500 words.

***Implications :*** In social sciences, the implications of the study are very important. Therefore, it should be included at the end in the chapter of summary and conclusions or a separate chapter may be prepared. The implications should be based on the significant findings of the study. They indicate probable usefulness of the findings for planning future work related to the study. The implications of various findings may be narrated with a view to promote future research work in the related scientific field and the future extension work for transfer of technology.

**(c) Bibliography (Literature Cited)**

This is very important topic which includes all the references cited in the text. Citation should contain all the data necessary to locate the source easily in the library. The references should be complete, consistent and accurate. The original papers must be referred while giving details about year, volume numbers, pages and correct initials of all the authors. All the references cited must be arranged alphabetically for better searching.

There are different sources of information such as textbooks, research journals, thesis, annual reports, magazines, newspapers, abstracts, encyclopedia etc. The style of writing bibliographical details changes with source. This is illustrated below.

**(i) Text book :** In case of text book following format may be used.

(1) Surname of authors followed by their initials.

(2) Year of publication.

(3) Title of the book with name of publisher.

(4) Pages referred for the study.

For example

**Rangaswamy R. (1967) :** *A Text Book of Agricultural Statistics*, Newage International Publisher Ltd., New Delhi, pp 100-110

**(ii) Research Journal :** The style of writing the research article is same as that of books with minor changes. The general format is as follows.

(1) Surnames of Authors followed by their initials.

(2) Year of Publication

(3) Name of Research Article referred with Name of Journal

(4) Volume No. and Issue No. of the Journal.

(5) Page Numbers referred.

Example :

**Gunjal S.S.**, *My Corrhizal Control of wilt in Casuarina Equisetifolia,* Agroforestry Today, Vol 22(3) : 11--116

**(iii) Thesis :** In case of thesis the general format is

(1) Surname of authors followed by his initials.

(2) Title of thesis with degree for which it was submitted.

(3) Name of the Institute.

(4) Pages referred.

**(iv) Reports :** In case of Government report or any other reports, they are published by the department, hence instead of author write the word "Anonymous" then the year of publications followed by the name of the report, name of Institute by which it is published and pages referred.

e.g.

**Anonymous**, 2001. *Epitome of Maharashtra*, Department of Agriculture, Government of Maharashtra, p.p.25-27.

The style of writing described above is a general format. However, minor changes are observed in the style of writing references. Therefore readers are requested to refer the particular journal/thesis of the concerned University/ Institute before submitting thesis, article, research paper for publication.

Due to computer facility browsing of references and its standard forms are available on web.

## (d) Appendices

The appendix should be typed in capitals and centered on a page preceding the appendix contents. The appendix provides a place for material that is complimentary to the text. Tables with too many details for text presentation, forms and questionnaires used for collection of data, analysis of variance tables and case studied too long to accommodate in text are included in Appendix.

The appendices may be numbered as Appendix-I, II, III and so on. A separate title to each appendix be given.

## 20.3 Glossary

**Scientific Report** : A scientific report is a report, which contains relevant, adequate and logically arranged scientific information presented in a specific format.

**Lay Audience**: Lay audience means general public, farmers and so on which do not possess formal education in a particular field.

**Thesis** : Thesis is a statement, projection or proposition which the researcher advocates and is prepared to maintain it.

**Dissertation :** It is a written presentation advanced by candidate for higher studies. Thesis/dissertation report of bonafide research conducted by the researcher on a given topic.

**Popular Article** : Popular article related to scientific subject is meant for lay audience and it is relatively less technical and written in local and simple language.

**Research Paper** : Research paper or an research article is a compressed and selective form of comprehensive report/research carried out by researcher.

**Synopsis** : Synopsis is the presentation of final research work carried out as per the plan and details of Outline of Research Work (ORW).

**Results** : Result is a statement of facts based on the systematically analyzed and tabulated data.

**Bibliography** : Bibliography is comprehensive listing of the references, it includes references cited in a text as well as other related references on the selected topic.

## 20.4 Summary

A thesis or dissertation is the final outcome of the research for the use of readers. Therefore, its presentation should be neat, correct and attractive. Research conducted by the researcher should be presented in a scientific way. Therefore, in the present topic, the guidelines for writing Introduction, Literature Cited, Materials and Methods, Results and Discussions, Bibliography etc. are given which will be helpful to the student in writing their thesis and research articles. For details, it is suggested to refer the books on scientific report writing.

## 20.5 Exercises for Practice

**Answer the following questions in 200 words each.**

1. Explain the need and importance of Scientific Report Writing.
2. What is review of literature ? Explain its importance in scientific research investigation.
3. Explain the need of materials and methods in scientific report writing.
4. Describe format of thesis, disserrtation, research paper and review articles.

## Reference Books


01. **Gupta S. P. (1985),** *Statistical Methods,* Sultan Chand & Sons, New Delhi-2.
02. **Nageswara Rao, G. (1983),** *Statistics for Agricultural Sciences,* OXFORD & IBH Publishing Co. New Delhi – 6.
03. **Rangaswamy. R (1998),** *A Text Book of Agricultural Statistics*, New Age International Publishers Ltd., New Delhi –2.
04. **Gomez K. A. and Gomez A. A.,** *Statistical Procedures in Agricultural Research*, IRRI, Philippines.
05. **Donald P. Warwick, Chales A. Lininger,** *The Sample Survey : Theory and Practice,* McGraw Hill Book Company, New York.
06. **Goode & Hatt**, *Methods in Social Research*, McGraw Hill Book Company, New York.
07. **Dhondyal S. P.,** *Research Methodology in Social Sciences & Essentials of Thesis Writing*, Aman Publishing House, Meerut.
08. **Kuppuswamy B.,** *Communication and Social Development in India,* Sterling Publishing Pvt. Ltd.
09. **Singh S. N., Bansal M. L., Singh T., Rakesh Kumar**, *Statistical Methods for Research Workers,* Kalyani Publisher, New Delhi.
10. **Gupta S. C. & . Kapoor V. K**, *Fundamentals of Mathematical Statistics,* Sultan Chand & Sons.
11. **Goulden C. H.,** *Methods of Statistical Analysis,* Johan Kiley & Sons.
12. **Paterson D. D.,** *Statistical Techniques in Agricultural Research,* McGraw – Hill Book company, New York.
13. **Love Harry. H.,** *Application of Statistical Methods of Agril. Research*, The Commercial press Ltd.
14. **Kempthorne Oscor,** *The Design & Analysis of Experiments,* John Wiley & Sons.
15. **Snedeeor & Cochron**, *Ststistical Methods,* Oxford & IBH Publishing Co., Bombay, Calcutta.
16. **Guilford**, *Psychometric Methods,* THM Edition Published by Tata Mc Graw Hill Publishing Company Faridabad Haryana.
17. **Henry E. Garret and R.S. Wood Worth,** *"Statistics in Psychology and Education"* 10th edition published by Vakils Feffer and Simons Ltd. Mumbai
18. **Sidney Siegel,** *"Non parametric Statistics for the behavioral Sciences"* International student edition by Tata Mc. Graw Hill New Delhi.
19. **George W. Snedecor and W. G. Cochrum,** *"Statistical Methods"* , 6th edition oxford and IBM publishing company New Delhi.
20. **Kumar Arvind (2002),** *Research Methodology in Social Science,* Swarup and Sons, New Delhi.