PREFACE

In the curricular structure introduced by this University for students of Post-Graduate degree programme, the opportunity to pursue Post-Graduate course in Subject introduced by this University is equally available to all learners. Instead of being guided by any presumption about ability level, it would perhaps stand to reason if receptivity of a learner is judged in the course of the learning process. That would be entirely in keeping with the objectives of open education which does not believe in artificial differentiation.

Keeping this in view, study materials of the Post-Graduate level in different subjects are being prepared on the basis of a well laid-out syllabus. The course structure combines the best elements in the approved syllabi of Central and State Universities in respective subjects. It has been so designed as to be upgradable with the addition of new information as well as results of fresh thinking and analysis.

The accepted methodology of distance education has been followed in the preparation of these study materials. Co-operation in every form of experienced scholars is indispensable for a work of this kind. We, therefore, owe an enormous debt of gratitude to everyone whose tireless efforts went into the writing, editing and devising of proper layout of the meterials. Practically speaking, their role amounts to an involvement in invisible teaching. For, whoever makes use of these study materials would virtually derive the benefit of learning under their collective care without each being seen by the other.

The more a learner would seriously pursue these study materials the easier it will be for him or her to reach out to larger horizons of a subject. Care has also been taken to make the language lucid and presentation attractive so that it may be rated as quality self-learning materials. If anything remains still obscure or difficult to follow, arrangements are there to come to terms with them through the counselling sessions regularly available at the network of study centres set up by the University.

Needless to add, a great part of these efforts is still experimental–in fact, pioneering in certain areas. Naturally, there is every possibility of some lapse or deficiency here and there. However, these to admit of rectification and further improvement in due course. On the whole, therefore, these study materials are expected to evoke wider appreciation the more they receive serious attention of all concerned.

Professor (Dr.) Ranjan Chakrabarti
Vice-Chancellor

Netaji Subhas Open University
Post Graduate Degree Programme
Master of Business Administration (MBA)
Course Code : CP-205
Course : Research Methodology

First Print : March, 2023

Printed in accordance with the regulations of the Distance Education Bureau of the University Grants Commission.

Netaji Subhas Open University
Post Graduate Degree Programme
Master of Business Administration (MBA)
Course Code : CP-205
Course : Research Methodology

: Course Writer :

Prof. Dilip Roy
Prof. BU (Retd.)

: Editing:

Prof. Ratan Khasnabis
Retd. Prof., CU

: Format Editor :

Prof. (Dr.) Anirban Ghosh
NSOU

Netaji Subhas Open University

Master of Business Administration (MBA)

Course : Research Methodology
Course Code : CP-205

Unit 1 □ Nature and scope of Research Methodology

Structure

1.1. Research

It is the search for knowledge that has shaped the human civilization. This search for Knowledge has led to scientific development and business progress through stages of invention, innovation and diffusion. This is an unending process too. What we establish today may become insufficient to explain the future experiences, For ex-ample, the concept that light travels in a straight line was a sort of knowledge with which experiences of reflection and refraction could be explained. But the presence of minute dark and light bands at the end of a shadow could not be explained with that knowledge. This called for a reexamination of the entire concept and the knowl-edge that came out of the new search was the wave theory for the movement of light. In fact, light is considered variously as a wave, corpuscular or quantum phenomenon. This process of examination and re-examination of different issues based on obser-vations and or experiences is aimed at expanding the domain of knowledge. This is a continuous process. It starts from where it ends. It is a movement from known to unknown. This movement is called Research.

This movement cannot be an aimless one or a haphazard one. There must be a systematized effort to gain new knowledge based on the existing knowledge and additional observations / experiences. Thus, one may, formally, define research as a scientific and systematic search for pertinent information on a given topic. For example, if we are interested in the field of marketing then marketing research will be defined as the systematic and objective search for and analysis of information relevant for the identification and solution of any problem in the field of marketing.

Question may arise about the emphasis given on the terms 'scientific' or 'objective', and 'pertinent or relevant information'. Actually, if research is allowed to be guided by personal views and personal considerations the utility of the research outcome will get significantly reduced. This type of biased conclusion is also very risky. Sometimes, risks are so high that one may not be able to use the research result at all. Research in the field of social science is more prone to subjective search than objective search.

Researcher must take adequate guard against these subjective influences and make the searching process an objective one.

It has also been observed that quite often irrelevant information is collected because of unclear idea about the information needs. Collection of irrelevant data not only increases the cost of the project and duration of the work but also creates difficulties during the analysis stage. It is, therefore, absolutely necessary to diagnose the information needs of a research project to plan the process of data collection. If properly planned, one can restrict the flood of irrelevant information and collect pertinent information only.

Information needs depend on the research problem itself. A clear understanding about the research problem eliminates much of the hurdles. This is because the search process can then be made more systematic and more scientific. The research methods and techniques to be used at different stages of the search process can be outlined in a better way. These, in turn, can give a clear idea about the information requirement. Researcher will know which information is pertinent and which the tion is not. Thus, formulation of the research problem and development of tte search process are two very important aspects of research. The later one is often referred to as research methodology.

1.2. Research Methodology

Research Methodology can, formally be, defined as a way to systematically and objectively solve the research problem. It tfis many dimensions to address. It not only describes the research methods to be adopted but also explains the suitability of one method over the other. It deals with the assumptions in sequence, their feasibilities. It describes as a whole how the researcher wants to arrive at a rational solution to his research problem. In this sense, research methodology is problem-specific. It starts by indicating why a research project has been selected. It explains how the research problem has been formulated. It presents the reasons behind the choice of hypothesis, choice of data sources. It presents the methods of data collection and describes the importance of data. It defends the choices of analytical techniques for extraction of information from the collected data. It covers many other aspects of the research activities to ultimately make the research results acceptable to others.

Research methodology is, sometimes, mixed up with the term ‘Research Methods’. But, conceptually, these two are different concepts. We have already explained what research methodology is along with its importance. Let us define the other term, i.e., research methods. Research methods are tools and techniques the researchers make use of while performing their research activities. Thus, methodology is problem-specific and methods are problem-independent. Research methods provide a set of different tools and techniques but in research methodology we make specific choices of different methods along with the reasons for such choices. Further, research meth-odology addresses the complete problem under study but research methods are of use for sub problems of the entire problem.

1.3. Scope of research and research methodology

Scope of research methodology is related with the scope of research itself and the way it is to be carried out. Be it scientific field or economic field or administrative field, research and research methodology have very important roles to play. In fact, in the recent past the role of research in the field of macro and micro economic activities has assumed a significant position widening the scope of both research and economic activities. This change in the role of research can be attributed mainly to growing complexities in the economic environment and to discrete changes in the business environment throughout the globe. Now-a-days nearly all government policies on the economic system are having an in-depth research backdrop. Government machineries carry out detailed analysis on the needs and desires of people at large and societal segments in particular to arrive at the future plan of the Government. No government can survive for long without meeting the social requirements. In fact, cost of carrying out research is much less than the cost of repairing measures if policies move in the wrong direction due to incomplete analytical studies. These remarks are equally applicable for corporate planning or operational planning of a business organization. If a business organization does not carry out adequate analysis of the business environment, it may have no other way out but to go for early retrenchment strategy.

In our country, research in government has been playing an important role by facilitating the decision making process and providing with pertinent information to the policy makers. Since there is region heterogeneity not only in respect of language and culture but also in respect of the level of development and economic growth, there is a basic need to understand the regional requirements and make allocation of nation's resources in a judicial manner so as to eliminate the disparities and march together for a better future. Our five-year plans are aimed at a balanced development, which is more than a mere growth in the national income of our country. To become effective in respect of targets and efficient in respect of utilization of human and nonhuman resources, planning commission is engaged in applied economic research with'central statistical organization as the main source for pertinent and detailed information. In fact, research as a tool for economic development and formulation of economic policy has three distinct phases of operation. The first phase of operation is known as investigation of economic structure. This phase involves continuous compilation of facts and figures for depicting the structure of the economy. In the second phase one diagnose the events along with their root causes of occurrence. Once the underlying causes are known it will be easier for the researcher to regulate them or take preventive measures to neutralize them. Next comes the obvious problem of prediction, the corresponding phase being known as prognosis phase. Keeping in mind the prediction requirement one may work out for future forecasting and arrive at an appropriate economic policy has three distinct phases of operation. The first phase of operation is known as investigation of economic structure. This phase involves continuous compilation of facts and figures for depicting the structure of the economy. In the second phase one diagnose the events along with their root causes of occurrence. Once the underlying causes are known it will be easier for the researcher to regulate them or take preventive measures to neutralize them. Next comes the obvious problem of prediction, the corresponding phase being known as prognosis phase. Keeping in mind the prediction requirement one may work out for future forecasting and arrive at an appropriate economic policy.

In view of the growing environmental complexity and uncertainty and global competition, research is also having an important role to play for. business planning to efficiently handle the functional activities and draw the future strategy of the economic unit. Basic functional activities cover marketing, production, logistics, human resource development and research & development. Resources are allocated to different func-tional areas depending upon their resource requirements, their past performances and achievements. In the field of production and operations management common research problems are the problems of cost minimization, assembly line balancing job-sequencing, transportation of items from plants to warehouses and from warehouses to markets. In the field of marketing management common research problems are the problems of sales analysis and forecasting of demand, new product development and test marketing, advertising and media selection, brand switching, consumer behaviour etc. Besides, research is being carried out in the fields of man-power planning, labour turnover-, job satisfaction, absenteeism etc.

1.4. Summary

Research is a movement from known to unknown, it involves systematic and scientific search for pertinent information on a specific topic. It is a continuous process that expands the domain of knowledge.

Research methodology is a way to carry out research. It describes es a whole how the researcher wants to arrive at a rational solution to research problem. It explains the choices of research problem, hypothesis, data sources, methods of data collection, analytical techniques etc. to ensure acceptability of the research outcomes.

Research methodology is different from research methods because the later is a collection of different tools and techniques, a researcher may make use of. Scope of research methodology is related with the scope of research and scope goes beyond scientific field. Research is increasingly in use to shape the governmental policies and the national economy. It is a must for business units for corporate and functional planning in the face of complex environmental changes.

1.5. Questions

Long answer type Questions :

1. What do you mean by research? Explain with examples from any field of importance..
2. Explain the needs for scientific and systematic search.
3. Research Methodology and Research Methods are two distinct concepts. Do you agree? Give reasons.
4. Describe the scope of research in economic and administrative activities.

Short answer type questions :

1. Why do we need pertinent information?
2. What happens if the search is haphazard?
3. Starting from the definition of research, define marketing research.
4. What do you mean by phases of research related to economic policy?

Objective type questions :

Indicate whether the following statements are true or false.

1. Research is a movement from unknown to known.

   True □ False □

2. Search for pertinent information increases the cost of search.

   True □ False □

3. Research is more meaningful when there are discrete changes in the business environment.

   True □ False □

4. Research methodology is problem specific.

   True □ False □

5. Research as a tool for economic development has two distinct phases.

   True □ False □

6. Business and research are antithetical.

   True □ False □

7. Arriving at a decision without research is less costly in the long run.

   True □ False □

Unit 2 □ Different aspects of a Research Process

Structure

2.1. Research process

To have a clear idea about the research methodology and research techniques one has to understand the generic process of research.-lt involves different steps which are not necessarily separable and distinct. However, for the sake of a general presentation we shall indicate different steps of thJ° research process in order to help the researcher in developing disciplined thinking and a rational bent of mind.

The first step is to define the research problem. If the problem is not well defined the search process cannot be scientific and complete. Incomplete search may lead to an incomplete answer. Further, a non-scientific approach may result in subjective conclusions, which cannot be of use in future research investigations. Sometimes, wrongly carried out research may mislead the future studies, resulting in wastage of intellectual energy. A well-defined research problem, on the other hand, can be instrumental in guiding the search process by giving a clear understanding about what is to be done and why it is to be dono. There are basically two types of research problem. One type of research problem attempts to examine the states of the nature. The other type of research problem wants to establish relationships among different variables of interest. These two are very broad areas. But one should start from one of these areas. Gradually, one should converge to the specific problem by reducing the ambiguities in the formulation of the problem. In this process, domain of study becomes properly defined. This, itself, is a type of scientific work and is the key step of the research process.

The second step is to review the concepts and theories and the previous reseach findings. This step is in line with our stated philosophy of research - a movement from known to unknown. This means we need to develop a clear idea about the existing knowledge. Existing knowledge being the starting point of any research work, one has to extensively study the literature in the field of interest. Sometimes, studies in the related field may be of immense help. One may start with the current journals, study the relevant research articles and examine the cited references and more backward to the extent needed. Current articles will give a list of references. Each of those references will give a list of earlier references. By keeping a track of all those references, one may get a clear idea about the current concepts and theories and the trend of work. Seminar volumes, conference proceedings, research reports, govern-ment publications, websites, current books and even popular magazines may act as sources of existing knowledge. Published and unpublished theses are also of great help for this purpose. All these references are to be arranged alphabetically along with their summary of findings. This entire exercise is a systematic search that will help the researcher to move in the right direction.

The third Step is to formulate the working hypothesis. It is the tentative assumption about the focal point of research; tentative in the sense that extensive study of the literature leads to this assumption but it is to be tested in the light of the researcher's experiences and or observations. Once the hypothesis is framed, analysis gets streamlined and he area of research gets limited. It goes on keeping the researcher on the right path by reminding him of the local point of work. It also helps the researchers to become clear about the data requirements and information needs. Mostly, working hypothesis fixes a value for a measure of the population under study. This measure may be a measure of location or a measure of dispersion. Sometimes, working hypothesis deals with interrelationship amongst a group of variables or attributes, one may be interested to know whether sets of variables or attributes are independent or not. In case of dependence one may be interested to know the extent of correlation. Occasionally, one may face a situation where a researcher does not need to make any tentative assumption in the form of a hypothesis. Mostly in situations where the researcher wants to gain familiarity with a phenomenon, formulation of hypothesis is not needed. Nevertheless, it is an important step to be followed wherever necessary.

The fourth step is to prepare the research design. By research design we mean a conceptual framework within which the investigation is to be carried out. Basic purpose of such a design is to induce efficiency in the research work while remaining effective at the end. The term efficiency is related to utilization of human and non-human resources in an optimum manner. In research, effort, time and money are the key resources, consumptions of which are to be minimized. This is, no doubt, a difficult task and it depends on the purpose and nature of research work. Generally, a researcher one' should address the following issues while developing the conceptual framework of research:

How the sampling units are to be drawn?

How the selected units are to be observed?

How the analysis is to be carried out?

How the entire process is to be operated?

While addressing these issues the researcher must take into consideration the cotnstraints he/she will be facing. These constraints mainly arise out of the availability and skills of the supporting staff, allocated time for this purpose and availability of funds, for research.

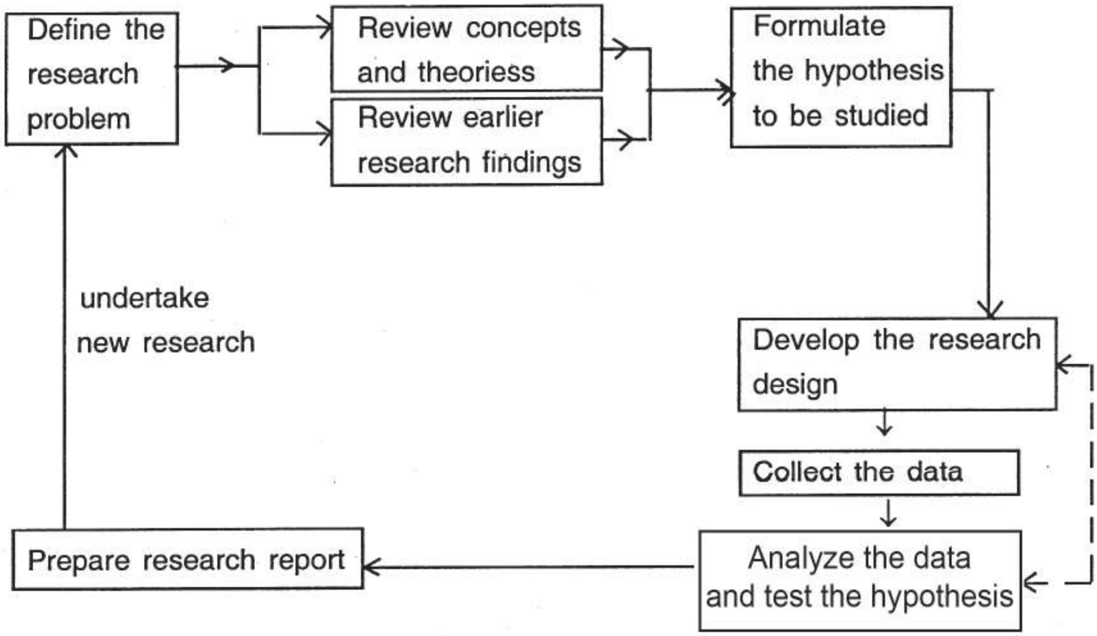
The fifth step is to collect the data. There are different ways of collecting the data, each way having its own merits and demerits, Here again, cost of collection and time needed for collection are two important issues that are to be kept in mind while deciding about the way of collecting the data. For example, primary data can be collected through survey work. These can also be obtained as the outcome of an experiment. Secondary data can be collected from statistical publications of governmental and non-governmental organizations, annual reports of business units, periodicals of different associations and technical bodies. Earlier research studies may also act as sources for secondary data.

The sixth step is to analyze data and test hypothesis. This involves consistency check at the very outset. If the collected data fail to ensure consistency either fresh collection of data is to be undertaken or the collected data is to be cleaned or dressed up so that the resultant data set becomes ready for further analysis. Codification is a several procedure to be adopted for ease of tabulation. The objective of codification is to transform categories of data into symbols that can be tabulated and counted. Once coding operation is over data is to be summarized initially in a tabular form. In fact, number of tables can be formed out of a given data set following different classification system. For example, consumption data may be classified according to age, sex, income and educational background of the consumers giving rise to four tables from the same data set. There may be cross-classifications to arrive at bivariate tables. For example, one may classify according to age and sex or age and in come. Computers can be of help in tabulating voluminous data with increased accuracy and reduced time. After the tabulation work, a researcher takes up the task of analysis. Analytical studies mostly make use of tfie tables for ease of derivation. There various statistical measures to describe the central tendency and dispersion of data. These can be, easily, calculated from the frequency table or from the raw data. In case of bivariate or multivariate data, one may examine the nature of dependence among the variables, differences among the marginal characteristics, or a multivariate measure based on the joint distribution. After the analysis of data one may go ahead with the testing problem and examine the tenability of the working hypothesis framed at the beginning of the research process. In the light of the given information derived from the collectgd data one may draw the necessary conclusion either in the form of the hypothesis or against the hypothesis. Either the null hypothesis, i.e. the working hypothesis will be accepted or rejected. In case there is a very strong evidence on the tenability of the hypothesis, the researcher may generalize the related concept into a theory. This generalization is the main aim of any research work.

The seventh step is to prepare the report of the thesis. This report should describe the work of the researcher by presenting an introduction on the problem, summary or preview of the work, main work and conclusions. References are to be added at the end. While writing the report every care should be taken· by the researcher to avoid vague expressions, redundancy, repetition and incompleteness. Each and very conclusion should be supported by facts and figures. Every statistical test should be accompanied by a statement on level of significance and distributional assumptions. It is also desirable that the limitations of the work be cleanly stated so as to help the future research in that area.

The following is a schematic presentation of the research process as described above.

FLOWCHART : RESEARCH PROCESS


Define the research problem
Review concepts and theoriess
Review earlier research findings
Formulate the hypothesis to be studied
Develop the research design
Collect the data
Analyze the data and test the hypothesis
Prepare research report
undertake new research

It may be pointed out in this context the generic nature of the above mentioned research process. Depending on a specific problem one may expand a step into multiple steps and may skip a few steps stated here in. We also propose to present a more detailed discussion on some of the steps and methods.

2.2. Problem formulation

Let us start with a symbolic presentation to define the research problem. Let
I be an individual or a group,
E be the environment in which I is there,
C_i, be the i-th course of action, $i \geq 2$, and
O_j, be the j-th outcome, $j \geq 2$.

Here, E needs to be defined in terms of the values of the uncontrollable variables or parameters. A course of action, on the other hand, can be described in terms of one or more controllable variables.

Let P $\{O_i[I,C_i, E\}$ be the chance of occurrence of an outcome O_i if I opts for the course of action C_i in the environment E. To make the problem interesting and meaningful at least one of the above-mentioned chances, must be different from the rest. In case this is not, all the courses of action are takrie to be equally good and there is no meaningful problem under the given setup. When chances are not all equal, the researcher's job will be to find the best course of action so that the researcher can achieve his/her objective in terms of the preferred outcome. This means, out of at least two outcomes of a course of action one outcome should be preferable.

Given the above setup, if the individual I does not know what will be. the best course of action then there is some doubt about the choice of the course of action. In that case, there exists a problem, which needs further investigation. Thus, we can enlist the components of a research problem as

1. existence of an individual or a group having some doubt about the best course of action,
2. existence of some objective, to be achieved by undertaking the best course of action,
3. existence of multiple courses of action for obtaining the desired objective.
4. existence of problem regarding relative effectiveness of the alternative courses of action,
5. existence of an environment that perturbs the deterministic nature of the out- comes, and
6. desire of the resehrcher to find out the best solution for this given problem in the context of the stated environment.

A problem becomes complicated when the environment is unstable affecting the values of the outcomes. Complicacies also arise in case there are large number of alternative courses of action. Sometimes, if person not involved in making decision, get affected by the outcome, additional difficulties may creep into the system.

Selection of the problem is, therefore, very important and requires utmost attention. As a thumb rule we may state that

1. a subject which has received limited attention in the past is a fit case for selection provided it has a wider scope of application,
2. a subrect-about which existing knowledge is vague and concrete formulation of the problem is difficult, is to be avoided,
3. a subject which does not meet the feasibility requirement should preferably be avoided,
4. a subject which is familiar but not overdone may be selected for study and
5. a subject which is not too familiar may be selected for study provided the preliminary study that precedes the main study supports such a choice in an unfamiliar area.

Once the selection of the problem is ove, the researcher must define the same in a formal way and this definition should b as clear and concrete as possible to eliminate the future hurdles, Basically defining a problem means indicating the objective of the study along with the domain of investigation. This definition is to be arrived at in a systematic and sequential manner. Let us describe that sequence-which is prescriptive in nature involving five steps,

Step 1. Start with a broad area of interest either from the scientific point of view or from the intellectual point of view or from the practical point of vew undertake a pilot survey to throw light on the feasibility of the study and then seek view of the experts and guides to state the problem in a general way.

Step 2. Try to understand the origin of the problem so that the thinking and rethinking process over the problem can be streamlined, focused and accelerated. Simultaneously, study the other nature of the problem to get an idea about the work boundaries.

Step 3. Scan the available sources of information and conduct an experience survey to know more about the problem under study. Select the specific areas, based on such discussions, so as to limit the field of investigation and know the techniques that can be used for analysis. Further, examine gaps in earlier works, weaknesses of earlier assumptions, limitations of earlier analyses etc.

Step 4. Rewrite the research problem in a focused manner by incorporating the initial general statement and the subsequent information gathered under steps 2 and 3. Rephrase the words wt1erever needed to make the proposition operationally viable and suitable for the formulation of working hypothesis.

Step 5. Include in the statement of the problem definitions of the technical terms used there in, assumptions and postulations made for further analytical treatment or for qualitative analysis. Also indicate the period of reference, place of interest if any and the value and the scope of the investigation.

Let us explain through an example how a general statement of the research problem can be made more specific and operational. Let us suppose that the broad statement of the research problem be as follows “To study the application of image processing techniques in business activities.”

This is a statement where practical concern relates to business activities, this area of interest is not a specific one and cannot be made operational without further specification. If we propose the banking sector it will narrow down the domain of study. Further, based on the earlier works and experience survey we may consider the issue of security solution as the more specific domain of study. Throughout the globe failure in security causes financial loss of a large amount and more amount of money is being spent to investigate into security lapses. Thus, if a security solution can be provided for, the banking sector will save enormous amount of money and the loss of face will be reduced or eliminated. This study being a technical solution for security problem specification of a time period may not be. necessary. However, place of study may have direct impact on the choice of biometrics and hence the place is to be specified. Let us propose to carry out the study for the Indian banking sector only. So far as image processing is concerned there are basically four generic approaches. These are respectively template or prototype matching, syntactic approac,h of matching, matching through neural network and statistical approach of matching. We may consider the last approach, i.e. the statistical approach. Then the new tittle of the research problem may be read as “To study the application of the statistical approach of biometrics matching as a security solution for the Indian banking sector.”

This tittle be rephrased to highlight on the application aspect of the study and hence the research problem may be described by the following title “In search of a security polution for the Indian banking sector using the statistical approach of biometrics matching.”

2.3. Summary

Research process is a systematic stepwise procedure that can help the researcher in developing disciplined thinking and a rational bent of mind. The first step is to define the research problem, which may be either an attempt to examine the states at the nature or an attempt to establish relationships among different variables. The second step is to review the concepts and theories and earlier research finding so as acquire a clear idea about the existing state of knowledge. The third step involves formulation of the working hypothesis, which is to be subsequently tested in the light of the researcher’s experiences and or observations. There are situations where formulation of hypothesis is not needed especially when the researcher wants tc gain familiarity with a phenomenon. The fourth step is to develop the research design that provides with a conceptual structure for carrying out the detailed investigation. The basic objective is to become effective with efficient utilization of human and nonhuman resources. Collection of data is the fifth step of the research process and the researcher has to decide about the way of data collection keeping is find the time and cost constraints. The sixth step involves planning for analysis of data and testing of hypothesis. The raw data is to be cleaned, coded and tabulated. Based on the looks generated in this process the researcher can analysie differerlt features of the data set. Thereafter statistical tests can be employed to arrive at the conclusion part of the work. The seventh step is to prepare the report, describing the entire work, its applications and limitations.

Problem formulation is the starting point of the research process. It depends on the objective, different courses of actions with probabilistic outcomes and an environment with reference to which the problem is to be solved. The problem becomes meaningful only when the chances of achieving objective through outcomes are unequal with doubts about the relative effectiveness of different courses of actions. There are some thumb rules for the selection of a problem. They deal with familiarity, feasibility, applicability, and whether a subject is overdone or not. Generally· development of a formal definition of a problem starts with a general statement and converges to the specific statement after preliminary literature survey and experience survey covering views of experts in the field.

2.4. Questions

Long answer type questions :

1. Describe, in brief, different steps involved in a research process.
2. Draw the flow chart for research process and explain the step you consider as the most important one.
3. ‘‘If the individual does not know what will be the best course of action in that case there exists a problem.’’ Examine this statement explaining the roles of individual and courses of actions. Explain the reasons for not knowing the best course of action.
4. Describe how a research problem is to be selected from a set research problems.
5. Explain with an example the steps involved in defining a research problem.

Short answer type questions :

1. How does a researcher carry out literature survey?
2. What do you mean by working hypothesis? Is it a must to have a working hypothesis? Give reasons.
3. Why do we call research design as a conceptual framework?

4. What are the aims of data analysis and testing of hypothesis?

5. Enlist the components of a research problem. Can there be a single course of action or a single output?

6. Do you agree with the statement ‘‘A subject which is overdone should not be normally chosen’’? If agreed, give reasons.

7. Why do you think that rephrasing is needed for defining a research problem?

Objective type questions :

Indicate whether the following statements are true or false.

1. Research design is a support of research process.

True □ False □

2. Working hypothesis describes the underlying assumption in a research work and not the idea to be tested.

True □ False □

3. Primary data can be collected from the government publications.

True □ False □

4. Cleaning or dressing up of data is needed before the tabulation of data.

True □ False □

5. A research problem should be rigidly defined so as to induce flexibility in the study.

True □ False □

6. There are basically two types of research problem,

True □ False □

7. Generalization of a concept into a theory is the ultimate aim of any research work.

True □ False □

8. Chances of occurrence of all the outcorns for each course of action must be equal to make the research problem interesting and meaningful.

True □ False □

9. In a problem there must exist multiple courses of action for obtaining the desired objective.

True □ False □

10. A problem becomes easier when the environment becomes stable.

True □ False □

Unit 3 □ Research Design

Structure

3.1. Research design as a conceptual framework

The research design provides with a conceptual framework within which the research activity is to be carried out. The basic objective is to minimize the bias and maximize the reliability of, the collected data and the subsequent analysis. It also addresses the issue of the economy in the procedure so as to carry out the research works with minimum cost. Major components of a research design are sampling design, observational design, statistical design and operational design. Under sampiing design one has to plan for the method of selecting the items which are to be studied under the research study. Observational design describes the conditions under which the selected units are to be observed. Statistical design prescribes the desired sample size and the analytical- procedure for extracting information from the collected data. The last design is operational design that deals with actual execution of sampling design, observational design and statistical design. It ties together these three designs to present a unified system.

Research design is a preplan for collection and analysis of data. Hence, one may consider research design as a strategy on the part of the researcher for gathering information and undertaking analytical studies. In all types of planning activities one has to take into consideration the available and mobilize-able resources and make use of these resources in an optimum way. Preparation of research design is no exception to this general rule. One has to take into consideration the availability and skills of research associates, availability of time and fund for completion of the. research work. In a planning process the starting point is the objective of the organization and given the goals to be achieved means are worked out under resource constraints. While framing the research design we have to follow an identical path. The objectives of the research will be the starting point for developing the research design. Once the well thought research design has been made, remaining task is to strictly adhere to the design, and implement the work according to plan.

Unfortunately, researchers sometimes undertake research activities without such proper planning, i.e. without a research design. Mostly they land up with unproductive work or with very limited application. Even if it is possible to come out with a solution to the research problem, the unplanned approach may lead to high cost and considerably high penalty. In today's competitive world any unwanted delay in the research work may deprive the researcher of the recognition as others may get their works published and recorded in the scientific literature ahead of haphazard research planner. Keeping in mind the above problems we may claim that the development" of research design is a must for completing the research work in right time with right cost and in a right manner.

3.2. Research design according to types of research

The nature of the research design will very from problem to problem. One cannot think of a standard remedial measure that fits for all types of research. Let us initiate our discussion by classifying different types of research and examining their planning needs. Only then one can suggest the appropriate designs and means for arriving at these designs. There are four broad groups under which one can classify different research problems. These groups are made according to !research objectives.I If the objective is to become more familiar with a phenomenon or to throw additional light on· it, the concerned research work may be termed as exploratory. or formu/ative research. If the object is to describe the state of affair with respect to a unit or a group, the concerned research work is referred as descriptive research. If the frequency of occurrence of an event or its occurrence in conjunction with another is the subject of interest of any research study, the same is known as diagnostic research. The research for studying the relationship among different variables is known as causal research which is often described as hypothesis testing. Research objective may be different combinations of the. above-mentioned four objectives. For such cases the research activities may be termed as mixed research. Research design will thus be of four specific types, namely, exploratory, descriptive, diagnostic and causal research. For a mixed research, the research design will be a combination of some or all of these specific research designs.

The features of a research design depend. on the type of research one proposes to do. An exploratory or a formulative research must be supported by a flexible design to permit changes in the exploration activity as and when needed. In case of a descriptive or diagnostic research accuracy of the portrait is very important and calls for minimum bias and maximum reliability in collected data. Let us explain the different research designs according to the type of research, i.e. exploratory or formulative research, descriptive and diagnostic research and hypothesis testing research.

3.2.1. Research design for exploratory or formulative research

Since the objective of this type of research is to formulate the problem for more precise enquiry the major emphasis is given on greater insight and on generation of ideas. Flexibility is the key word for transforming a broadly defined research problem to a precisely defined one through survey of related references, experience survey and case analysis.

Experience survey is the survey of experienced people who have direct knowledge on that research problem under study. They can help to have greater insight and may share new ideas with the researcher. These persons are to be selected based on the criteria of experience and competence and not based on a random mechanism. They are to be interviewed mostly with probing questions. Even on occasions they may be permitted to ask counter questions or raise issues not apprehended by the researchers. It is the flexibility in the collection of information that may help the researcher in noting the unnoticed issues, focusing the process of search and rephrasing the research hypothesis to make it more meaningful and precise.

Case analysis involves study of selected instances of the phenomenon to stimulate insight into the research topic. Because of this fact, case analysis is also known as analysis of insight-stimulating situations. These cases are to be so selected as to cove the extreme situations and become aware of the different features or dimensions along which the hypothesis can be reformulated. Attempts should be made to cover the views of heterogeneous groups or individuals, if needed on those cases so that analysis can be multidirectional.

We may conclude that for exploratory or formulative research the overall design must be flexible to capture different aspects of the research problem as and whep the additional information creeps in. For experience survey, one should take care of purposive or judgment sampling in place of probability sampling. Selection of cases should also be made according to the research purpose to project the extreme situ-ations. No preplanned procedure can be suggested for making observations. Unstructured instruments are the best choices for collection of data. Also, one cannot prep Ian the statistical or other analyses. In view of all these flexibilities needed at every stage, starting from sampling design going up to statistical design, one can hardly fix the operational design.

3.2.2. Research design for descriptive. and diagnostic research

Research design for descriptive and diagnostic researches must be rigid and categorically defined. It must address the basic issues relating to collection of data, processing of data and reporting of the findings. Jf the rigidity is relaxed, the collected data may not be sufficient enough to provide with reliable information and unbiased measures. It is desirable that the instrument for data collection be pre-tested. Different instruments/ which are mainly used for descriptive and diagnostic research, are ob-servation, questionnaire, interview and earlier reports. Units are to be selected based on a rigidly defined sampling design. Probability sampling is recommended to eliminate the bias, if any, on the part of the researcher. The observational design must also appreciate structured instruments. Preplanning is needed for carrying out statistical analysis and other analytical studies. All these rigidities stress the need for advance decisions on the administrative procedures to make the above designs operational.

3.2.3. Causal Research

In hypothesis testing research, one has to infer on causal relationship. We shall now discuss the framework under which such studies are done.

The framework is known as Causal Research. The necessity of having such a framework can be understood if one considers a limitation of descriptive studies. Descriptive studies can capture the features of a phenomenon with a measure of success. But they do not aim at attaining precision in prediction. Usually, the researcher tries to attain a precision in prediction by identifying the cause and effect relations among the factors that describe a phenomenon. Precision,in quantitative terms,can be achieved by describing the factors as variables and then identifying the underlying relations among the variables. Usually, this is done by understanding the relation of causation in the phenomenon under study. The study is therefore, called a causal study and research framework is often noted as Causal Research.

The basis of causal study is the understanding that there exists a cause and effect relation that explains the feature and the behaviour of the object under study. The philosophical root of causality as a method of study is positivism, which is usually applied in the study of the natural phenomena. Ordinarily, the causation is studied in terms of variables, which are· usually taken to be quantifiable. However, the logic of causation is applied also in such cases where the variables are qualitative in nature. In empirical research, usually, the causality is first conceived in a theory that describes a phenomenon in terms of cause and effect relation. The relation is then tested empirically by having the data with respect to variables under study. Some-times, the theory is sound to be inadequate for developing a working hypothesis for empirical research. In such a situation, the researcher takes recourse to experimental study for unfolding the causality that might be hidden in the phenomena. (See Experi-mental Technique and Experimental Design in Unit 5).

In the causal studies, the variables are first understood as explained and explanatory set of variables. The casual variables are known as explanatory variables and outcome or the effect of the causal variables are noted as explained variables. For example, the quantity demanded with respect to a particular product (assumed to be captured in sales figures) might be taken as explained variable. From theory, one understands that the sales values of the product is explained by several factors such as the price of the product (P_x), the price of the related products, (P_y P_z), the purchasing power of the customers (captured through the disposable income (Y) of the target group of customers) and also the residual factors, i.e., the factors that have not been taken care of in the theoretical construction. If the sales data describe the amount of the product sold and we note it as quantity demanded (X_d), then X_d, is the explained variable and according to the theory, X_d is explained 'by a set of explanatory variables such as P_X, P_y, P_z,Y and R which is the symbol for the residual term. In simple mathematics this is a description which is written as $X_d = f(P_X, P_y, P_z, Y, R)$. In empirical research the researcher then specifies the exact form of the function and tests the validity of the relation by taking recourse to the econometric analysis.

Given this broad outline, we shall now discuss the types of causation that we may conceptualise, while taking up the empirical research.

Deterministic Causation

. The causation is considered to be deterministic, if the identified explanatory variable can explain the behaviour of the explained variable completely. Consider for example, the previous example of the behaviour of the sales data that is supposed to describe the demand begaviour of the product. If it is assumed that the demand of the product described through the quantitative information of the sales data and the quantity sold is explained by four explanatory variables, namely price of the product (P_x), price of related products (P_y and P_z) and the level of income of consumer (Y), then, in effect, it is being suggested that the volume of sales is completely determined by these four factors. The implication is to be noted carefully. If we know the exact relation by which these four explanatory variables are related to the explained variable, then for various given values of P_x, P_y, P_z and Y we can determine the quantity demanded for the product (X_d)with all the precision.

A causal study which is performed under this framework is known as a study under deterministic causation.[1]

Probabilistic Causation

Let us try to enter deeper into the problem. How one knows that X_1, X_2, and X_3, are three variables, which are necessary as well as sufficient to capture the. movement of the variable Y? The answer could be that we get it from the theory... in this case from the theory of demand for the product under consideration. How the theory is developed is a different issue and does not belong to the purview of the present discussion. The researcher accepts the theory as a robust one so much so that according. to his belief the theory has the power to explain fully the behaviour of X_d , from the given information on P_x, P_y, P_z and Y.

No theory, however, is that much robust at heart from the empirical point of view. If we test the theory for empirical verification, one will observe that even with an exhaustive set of the explanatory variables, one cannot explain the behaviour of the explained variable in totality. A gap will still exist-the gap between what the theory explains and what occurs in reality. The gap will, of course, below; given that the theory is sufficiently robust. A robust theory will minimise the gap, but the gap is still expected to exist. *The reason is that there is basic randomeness in the behaviour of the variables under the study.* The root of this randomness is, of course, the incompleteness in knowledge. But this incompleteness, one suspects, will exist forever. The researcher in the empirical area of the subject needs a tool to tackle such a situation. A probabilistic model that contains/captures the chances of occurrence of the explained variable, given the occurrence of one or more explanatory variables, can be suggested as the effective way of handling such a situation. Instead of taking a deterministic model one may gain by replacing this with a probabilistic model where the researcher enters a disturbance terr in the proposed regression. The disturbance tern is taken as a random variable with a given probability distributior..

3.2.4. Concluding Observations on Research Designs

While setting the research design, a practitioner should note that a particular research design is drawn on the basis of the suitability of the proposed design in the context of the research problem. The researcher shall have to apply his own judgement as regards the suitability of a particular design in the context of a given research problem. It is not

[1] When we run a regression wth X_d as the explained variable and P_x, P_y, P_z and Y as explanatory variables, it may so happen that the regression cannot explain X_d completely in terms of the four explanastartstory variables. Still the model will be considered as a deterministic model because the model starts with certainty, which part remains unexplained by the selected explanatory variables.

necessarily true that the causal resarch design would be the first choice for a researcher because it contains more mathematical rigour. For a particular research problem the most suitable design might be the exploratory research design. In some cases a descriptive research design might be tackled best by drawing a research design which contains the elements of various designs that have been outlined above. In other words, flexibility in research design is key to a good research. The researcher will be rewarded best if he draws his research design according to his perception of the research problem keeping in view that a part of the research might be exploratory in nature ; but at the same time some parts· contain the research issues that could be addressed in a better way if the design is causal in nature. An understanding about the alternative research designs helps the researcher understand the nature of options that he may apply while pursuing a research problem.

3.2.5. More on Hypothesis Testing

In hypothesis-testing research one has to infer on causal relationship. To do so, one may have to conduct experiments not only to ensure minimum bias and maximum reliability but also to provide information on marginal and joint effects of the causes under study. The research design for such studies is also known as design of experiments. Though experimental design made a humble beginning in the field of agricultural research, it is now a subject by its own merit. However, in view of its agricultural origin several terms such as treatment, plot, block, yield etc. will be in use in the description of experimental design irrespective of the field of study. There are three basic principles, which are of use in making the inference sound and powerfulel. These are respectively the principle of replication, the principle of randomization and the principle of local control. Let us explain these terms in sequence.

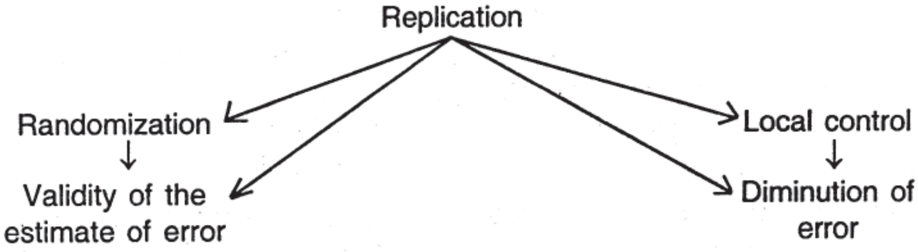
The principle of replication underlines the need for repeating the experiment more than once. This means each treatment is to be applied on more than one experimental unit. During the stage of analysis, the estimation of treatment effect can be done with increased statistical accuracy with the standard error inversely proportional to the square root of the number of replication. Sometimes the entire experiment may be repeated several times to attain a targeted level of accuracy. It may, however, be kept in mind that greater replication means higher cost and greater computational work. It is therefore, the task of a researcher to design the experiment in such a way that there can be an optimum balance between accuracy and cost.

The principle of randomization provides a guard against effects of extraneous factors. Extraneous factors are those factors that are not related with the research study but may affect the dependent variables, if not properly planned. The effects of extraneous variables on dependent variables are technically referred as experimental error. The purpose of the experimental design is to see that the effects on the dependent variable can be attributed solely to the independent variables and not to the extraneous variables. Then, variations caused by the extraneous variables can all be combined under the chance category. The principle of randomization in its purest form means random assignment of treatments to experimental units or random selection of experimental units for assignment of a given treatment.

The principle of local control in also known as error control. In the simplest case, division of experimental units into homogeneous groups or blocks can eliminate the variation among the groups or blocks from the error and increase the efficiency of the experiment. Sometimes, restricted random allocation of the treatments to experimental units may ensure reduction in error and increase in efficiency of the design. There are other means of controlling error. Size and shape of the experimental units are having effects on the error. If we consider agricultural plots of land as experimental units, plots of bigger size may tend to be fertility-wise heterogeneous and invite more error. Another interesting way of controlling error is the use of confounding approach. When the dependent variable gets affected by the extraneous variable the relationship between the dependent and independent variable becomes confounded by that extraneous variable. In experiments involving large number of factors or treatments. the blocks of experimental units become large in size. To avoid this problem of lar:ge block size and the resulting problem of high degree of error due to heterogeneity, one may divide the block into multiple parts confounding he effects of some treatment combinations with the block effects.

In general, all these three principles can be used simultaneously while designing an experiment. While replication and randomization can be applied together to validate the error estimate, replication and local control can be combined for diminution of the error component. Let us reproduce below Fisher's diagram to explain how these three principles act on the error component to ensure validity of the estimate of error and control the extent of error.

Fisher's diagram



Regarding the statistical design for such experimental research, the analysis of variance (ANOVA) technique is very useful to provide information on total variance, error variance, block variance (wherever applicable) and treatment variance. Treatment variance may be subdivided into multiple components so as to examine the main effects of the treatments and joint effects of their combinations. Effects of combinations of treatments are also known as interactions. First order interactions are those effects where two treatments get combined. Higher order interactions take into consideration joint effects of more than two treatments. However, for informal experimental designs which are frequently used in business research and research in social science, comparison of proportions or means can be used to arrive at an overall conclusion about the effectiveness of a treatment.

The differences among the research designs for the four types of research studies (exploratory or formulative, descriptive and diagnostic and hypothesis-testing) can be summarized as follows :

Research design	Type of study		
	Exploratory or formulative research	Descriptive research and diagnostic research	Causal research
Overall design	Flexible design to keep scope for considering different additional aspects of the problem as the research work progresses.	Rigid design to ensure minimum bias and maximum factuaf reliability under cost. and time constraints.	Rigid design of experimental type to ensure minimum bias, maximum reliability and valid information on causal relationship.
Specific issues: (a) Sampling design	Non-probabilistic sampling design is preferred for selecting respondents based on purposive or indgement sampling	Probabilistic sampling design is needed to minimize sampiing bias and obtain valid estimators.	Experimental design is preferred based on the concept of randomization, replication and focal control.
(b) Observational design	Unstruetured instruments are used for the collection of in information or data.	structured instruments are used to collect information and data.	Instruments are needed to 'measure or observe the effects of experimentalion .
(c) Statistical design	No greplanned analytical procedure Analysis depends on the actual information collected.	Preplanning is needed to design the analysis of information and data.	Mostly ANOVA technique is being used for formal experimental design. Other statistical methods be applied too.
(d) Operational design	Preplanning is not possible about the operational procedure due to flexibility of the system.	Prepfanning is needed to make the rigid design operational in practice.	Preplanning is needed to conduct the experiment to observe the expeimental results.

3.3. More about experimental design

Experimental designs are of two types; informal experimental design and formal experimental design. Informal experimental designs are less sophisticated and easy to implement. Formal experimental design employs sophisticated procedure to ensure greater applicability of the research results. Informal experimental design is of three types. These are respectively Before and After without control, Before and After with control and After-only with control.

As their names stand the first one is made of only one group of experimental units on which the treatment is to be applied. Observations are to be taken for each experimental unit before the application of the treatment and after the application of the treatment in respect of the characteristic of importance. Thus, each experimental unit will provide at the end of the experiment two observations. Let us denote by X_o, the observation before the application of the treatment and by X, the observation after the application of the treatment. Then, for that unit of study the treatment effect, T, can be expressed as

$T = X_z - X_o$

which is the net difference between the levels of the phenomenon under study before and after the treatment. Though we are having only one group of experimental units for this study yet for the sake of comparison with other designs we like to refer this group as test group because all the experimental units belonging to this group receive the treatment. The average T value will give the overall effect of the treatment.

In case of Before and After with control type, of experimental design we consider two groups of experimental units. Units belonging to the first group, known as test group, or experimental group, receive the. treatment under study. But units belonging to the second group, known as control group, do not receive the treatment. Let us denote by X_o, the observation on the characteristic X, for an experimental unit taken before start of the treatment. For the same unit, let X_1 be the observation taken after the treatment. Then the difference $D = (X_1 - X_o)$ will measure the change in the characteristic. The average value of D, averaged over all experimental units, will give the average change in the level of the phenomenon. This change may take place due to two reasons. one due to treatment effect and the other due to change in time. To get a precise idea about the treatment effect we need to subtract from the average D value the second component, i.e. change due to time. For this, we also observe the experimental units of the control group before the start of the treatment for the test group. We also observe the experimental units of the control group after the end of the treatment for the test group. Control group units themselves do not receive the treatment. Let X_o^c denote an individual observation for the control group before the start of the treatment for the test group. Let X_1^c denote observation for the same unit after the treatment of the test group. Then the difference $DC = X_1^c - X_o^c$ will denote the change in the level of the phenomenon due to change in time. This difference will have no treatment component, as the units belonging to the control group do not receive any treatment. The-average of such oo values will estimate the average change due to time. In that case, the average treatment effect (ATE) can be expressed as.

ATE = average D value-average D^c value,

under the condition that experimental units are randomly assigned to the two groups. This design is more informative because of presence of the control group and can be viewed as an improvement over the experimental design Before and without control. For example, there may be healing over time for some diseases. In case we are having a control group and a test group and observe both the groups before the start of the experiment and say after one month after the end of the treatment on the test group, we may examine the conditions of the patients belonging to both the groups and draw precise conclusion about the effectiveness of the treatment and effect of time alone.

There is another situation where observations are not available at the beginning of the experiment but a test group and a control group are available. In this case, the average level of the phenomenon after the treatment (A) and the average level of the phenomenon without the treatment (W) can be noted down. If experimental units are randomly assigned to test group and control group, then it is reasonable to measure the treatment effect by the difference between A and W. Thus,

Treatment effect = A – W.

The following is a schematic presentation of the three informal experimental designs described so or may be observed that in the schematic presentation we are having period as one dimension and there are two periods. The other dimension is treatment and there are two groups, viz., with treatment and without treatment. The measure under study is the level of a phenomenon. The objective of the study is to· measure the treatment effect based on the available information.

Observation

Design	**Type of group**	**Period I**	**Poriod II**	**Treatment effect**
Before-and After without control	Test group	X_o	X_1	$T = (X_1 - X_o)$
Before-and After with control	Test group Control group	X_o X_o^C	X_1 X_1^C	$T = (X_1 - X_o) - \left(X_1^C - X_o^C\right)$
After-only with control	Test group control group	- -	A w	T = (A - W)

It may be pointed out that Before and After with control is the best informal experimental design. However, depending on the situation or availability of the information we may opt for the other two informal designs. A fourth possibility, i.e., After only without control cannot be used for estimating treatment effect without any additional knowledge on the experimental units.

Formal experimental designs are large in number and hence we will discuss only a few basic experimental designs. The rest will be available in any advanced book on design of experiments. Before we introduce those formal designs let us formally define the term treatment. The term treatment describes different condition or conditions under which experimental units are to be placed. Different promotional schemes may be viewed as different treatments. Different training programmes may also be considered as different treatments. In fact, wherever we can think of different courses of actions which can be applied on different individnals, units or groups with a common objective, we may consider those courses as different treatments. In its simplest form we may consider Completely Randomized Design (CRD) as the basic formal experimental design. It makes use of the two basic principles–replication and randomization.The principle of local control is not used· in the concept of CRD.

CRD lay out : In CRD t treatments are allocated to n experimental units in a random way with r, experimental units receiving the i-th treatment, i =1, 2, ... , t, and where

$$n=\sum_{i=1}^{t} t_{r_i}$$

In practice, experimental units are assigned unit identification numbers from 1 to n and then a random permutation of these units is selected. Next, first treatment is applied to first r, units, second treatment is applied to next r_2, units and so on. If in the random permutation units are numbered as U_1, U_2, ... Un, then the scheme of CRD will be as follows :

Treatment	**Experimental units assigned to it**
1	$U_1, U_2, \ldots\ldots, U_{r_i}$
2	$U_{r_{1+1}}, \ldots\ldots\ldots\ldots\ldots, \underline{U_{r_{1+}r_2}}$
$\vdots$	$\vdots$
i	$\underline{U_{r_{1+\cdots+}r_{i-1}}+1}, \ldots\ldots, \underline{U_{r_{1+}r_2+\cdots+r_i}}$
$\vdots$	$\vdots$
t	$\underline{U_{r_1+\ldots\ldots+r_{t-1}+1,}} \ldots\ldots\ldots , \underline{U_{r_1+r_2+\ldots\ldots+r_t}}$

The simplest formal experimental design that uses all the three principles of experimentation, viz., randomization, replication and local control is Randomized Block design (RBD). To introduce local control, n experimental units are grouped into r blocks each block having a size t. Thus, $n = rt$. To reduce the error term each block is to be made as homogeneous as possible and between block variations as high as possible. Next, in each block t treatments are assigned to t experimental units in a random fashion and in this way in all the r blocks all the experimental units are assigned treatments. It is easy to note that the replication for each treatment is r as each treatment is assigned once and only once in each of the r blocks. Thus, each treatment is randomly assigned to experimental units within a block upholding the randomization principle. Also, each treatment is repeated r times following the replication principle. Further, formation of blocks to reduce within block variation it turn, reduces the total error and adheres to the principle of local control.

Latin Square Design (LSD) is a generalization 9randomized block design where local control is carried out along the two directions, viz; row and column. LSD is of matrix c/ / , form and can be constructed by replicating t times' each of the t treatments. This fixes the number of experimental units to $n=t^2$. Construction of LSD involves a few additional restrictions. In the LSD matrix with t rows and t columns there are t^2 experimental units and t treatments are so assigned that in each· row and in each column each treatment must be present only once. We present below three typical LSD for treatment numbers equal to 2, 3 and 4 respectively. Writing A, B, C, D to denote different types of treatment, standard Latin Square for t = 2, 3 and 4 are

LSD for t=2	LSD for t=3	LSD for t= 4
A B	A B C	A B C D
B A	B C A	B C D A
	C A B	C D A B
		D A B C

From standard Latin squares we may construct other Latin squares by permuting the rows, columns and letters (expressed by alphabets). The total number of distinct and meaningful Latin squares that can be generated from standard LSD of size t × t is given by

$$t!\ (t - 1)!$$

It may be noted that LSD considers all the three basic principles - replication, randomization and local control. It is made of two RBDs. One RBD is with t rows as t blocks and the other RBD is with t columns as t blocks.

CRD is useful for small initial experimentation. It gives good result if the experimental units are homogeneous in nature. There is no restriction on the total number of experimental units. Flexibility is there regarding the extent of replication. Frequency of replication may vary from treatment to treatment. RBD is useful for moderate number of treat-ments. In fact, RBD is the most widely used formal experimental design. It is flexible too except for the fact that the total number of experimental unit must be a positive and integer multiple of the number of treatments. By flexibility we mean that there will be no major change in the statistical design if information for one entire block gets damaged. The number of replication will be reduced by 1 only. The only problem one faces with RBD is the creation of blocks. If within block homogeneity cannot be ensured then RBD is hardly an improvement over CRD. LSD design eliminates row and column effects from the error term and hence provides with greater efficiency. But it has many restrictions. The foremost restriction· is that number of replication should be equal to number of treat-ments. Second problem arises out of analytical difficulties in case of loss of partial data. Problem also arises if the number of treatment becomes large.

We shall end our discussion by referring to another type of design, known as factorial experiments. In place of a single factor one can consider multiple factors with each factor having multiple levels, for a two-factor factorial design if the first factor has s_1 levels and the second factor has s_2 levels then there will be in total $s_1 \times s_2$ factor combinations. For a k-factor factorial design if the i-th factor has s, levels then the total number of factor combination will be equal to $s_1 \times s_2 \times .... s_k$. These factor combinations may be compared using a CRD or RBD or LSD. For example, if we consider 2^2 experiment, i.e., 2 factors each with 2 levels, there will be 4 factor combinations. If we denote these factor combinations as

A : Both the factors at first level

B : First factor at second level and second factor at first level.

C : First factor at first level and second factor at second level

D : Both the factors at second level.

Then the standard LSD for conducting this 2^2-experiment will be given by

$$\begin{matrix} A & B & C & D \\ B & C & D & A \\ C & D & A & B \\ D & A & B & C \end{matrix}$$

For RBD if we consider r blocks then in each block there will be 4 experimental units. Four factor combinations A, B, C, D are to be randomly assigned to units. For CRD it the factor combinations A, B, C and D are repeated r_A, r_B, r_C and r_D times respectively then there will be, in total $n = r_A + r_B + r_C + r_D$ experimentai units.

3.4 Summary

Research design provides with a conceptual framework within which the research activity is to be carried out. Major components of a research design are sampling design, observational design, statistical design and operational design. Sampling design deals with the method of selecting the items to be studied. Observational design describes the conditions under which the selected units are.to be observed. Statistical design gives a plan for analysis, indicating the requirement OJ sample size and the suitability of analytical techniques. Operational design indicates the procedure for making the earlier three designs operational. To complete the research work in right time with right cost and. in a right manner thoughtful development of a research design is a must.

However, this development depends on the nature of research to be carried out. There are basically four types of research study. To become more familiar with a phenomenon one carries out exploratory or formulative research. To accurately portray a situation one takes the course of descriptive research. In diagnostic research one examines the frequency of occurrence of an event in conjunction with another event. In case one is interested to study the relationship among a set of variables the appropriate research will be hypothesis-testing.

Research design for an exploratory or formulative research should preferably be flexible in nature to ensure greater insight and generation of new ideas. Survey of related references, experience survey and case analysis are the three useful methods for this research. For experience survey, purposive or judgment sampling will be more appropriate than probabilistic sampling. Choice of cases should also be done according to research objective. Mostly unstructured instruments are used for. getting observations. In view of the flexible nature of the entire system neither the statistical design nor the operational design can be preplanned.

Research designs for descriptive and diagnostic research activities are similar. Rigid designs are used to minimize bias and maximize reliability for this purpose, keeping in mind the cost and time restrictions. Sampling design is based on probabilistic sampling to provide with valid estimators for the population parameters. Structured instruments are of use for collection of data. Both the statistical design and operational design can be preplanned in these cases.

Research design for hypothesis testing research encourages experimental design a which may be both informal and formal types. The basic principles used, in formal experimental design are replication, randomization and local control. Replication along with randomization provide valid estimate of error and replication along with local control lead to diminution of error. Informal experimental designs are of three types, viz., Before and After without control, Befor and After with control and After-only with control. Before and After with control design provides the most precise measure of treatment effect. In case of formal experimental design there is a plethora of designs out of which CRD, RBD and LSD are three very important and basic designs. For multilevel multifactor situation one may adopt factorial experiment and design the re-search work in both confounded and un-confounded manner.

3.5. Questions

Long answer type questions.

1. What do you mean by a research design ? Indicate the importance of the same.
2. "Flexibility is the key word for research design in exploratory research". Discuss.
3. Compare and contrast research designs for exploratory and descriptive researches.
4. Explain the three principles of an experimental design.
5. Which informal experimental design do you prefer ? Give reasons for your preference.
6. Explain the construction of a formal experimental design based on the principles of randomization replication. Indicates its merits and demerits.
7. ‘‘RBD is the simplest formal experimental design that uses all the three principles of experimentation’’. Discuss RBD, indicate its layout and give your choice in response to the quoted sentence.
8. Can we consider LSD as a generalization of RBD ? If yes, explain the concept of generalization. If no, give arguments in your favor.
9. Explain the concept of experience survey and its use in research design.
10. ‘‘No single research design is suitable for all types of research studies’’. Discuss.

Short answer type questions :

1 . What is case analysis ? When do we use it ?

2. Explain with an example roles of test group and control group.
3. Research design for descriptive research is rigid. Do you appreciate ? Explain.
4. Explain randomization and replication and their joint effects.
5. Describe the following terms.

   Extraneous variable, confounded design, treatment.
6. In a Before and After without control experimental design how do you determine the treatment effect ?
7. Give the schematic presentation of the three informal experimental designs.
8. Starting from the following standard LSD construct two new designs.

   A B C D
   B C D A
   C D A B
   D A B C
9. Explain the formation of blocks in RBD.
10. Why is purposive or judgment sampling suitable in experience survey ?

Objective type questions :

Indicate whether the following statements are true or false.

1. Hypothesis-testing research portrays the characteristics of a population.

   True □ False □

2. To become familiar with a phenomenon one should carry out diagnostic re-search.

   True □ False □

3. Experience survey is the survey of experienced people

   True □ False □

4. Purposive or judgment sampling is a probabilistic sampling.

   True □ False □

5. Descriptive research design aims at minimum bias and maximum reliability

   True □ False □

6. Replication and local control help in reducing error.

   True □ False □

7. Treatment combinations are known as main effects.

   True □ False □

8. Treatment effect can be measured through after-only without control informal design.

   True □ False □

9. Test group is the collection of experimental units upon which treatment is to be applied.

   True □ False □

10. In a 23 factorial experiment there are 6 factor combinations.

    True □ False □

11. In LSD number of treatment equals number of replication.

    True □ False □

12. RBD is useful for small initial experimentation.

    True □ False □

13. In RBD if within block variation is maximum then the efficiency of the design is maximum,

    True □ False □

Unit 4 □ Methods of data collection

Structure

4.1. Data

The basic ingredient of empirical research is ‘data’. The generic source of the word ‘data’ is Greek.‘data’ is plural, the singular of which is ‘Datum’. In normal usage the word data is often taken as singular- a collective singular term which indicates information captured in precise terms. Data could be qualitative as well as quantitative. However, in most cases, the information that comes under 'data' is in quantitative form.

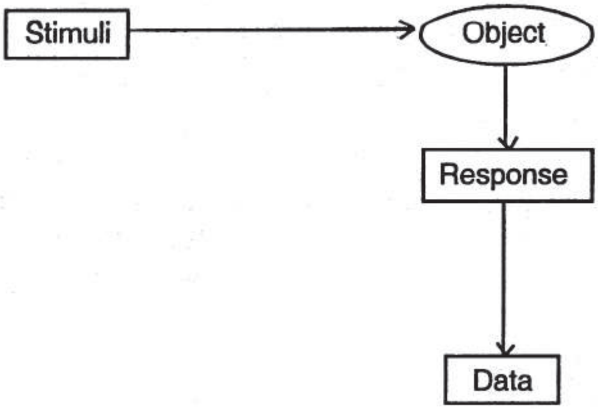
In empirical research, the quality of research depends very much on the quality of data. A researcher should, therefore, be careful while handling: the data, Even if the researcher has a good d6al of exposure to quantitative techniques and even if the researcher has proficiency in handling computer software packages, it may so happen that he fails to achieve a desirable quality in his research output because the data quality was not good. A researcher would get a better insight into the research issue if he is aware of this problem and considers the issue of collecting the data more carefully.

4.2. Data Generation

For this one must know the origin of data. For a research, a researcher uses two kinds of data primary and secondary. Primary data are such data wich have. been generated by fue researcher himself. On the other hand, the secondary data are such data which are collected by the researcher from a source. The researcher himself has not generated these data. While adhering to this broad division the researcher should note that every data is basically, primary data because someone must have generated the data which appears as secondary to the particular researcher.

While using the secondary data and also while generating the- primary data, a researcher would be benefited if he understands the mechanism by which the data set is generated. What we use as data is the outcome of the interaction between what is known as stimulus and the object on which the stimulus is imparted. The mechanism by which stimulus or stimuli generate data by working on the object, may be described in the form of the following flow chart :

MECHANISM OF DATA GENERATION



When a researcher visits a respondent and asks questions basically he is imparting 'stimuli' in the form of questions. The 'object' on which the stimuli are being imparted is the respondent in this case. The outcome is 'response' out of which the researcher gets information. The information is then processed to get qualitative or quantitative data.

The nature and the quality of the data depend, to a large extent, on the nature of stimuli applied to the respondent, which may by classified as systematic and unsystematic stimuli. Under systematic stimuli all units are exposed to the same standardised stimuli and stimuli follows a given pattern. Under unsystematic stimuli there exists respondent specific variation in the stimuli.

What do we mean by systematic stimuli? In social research, we often collect data with respect to a set of respondents (‘object’ in the above flow chart). Under systematic stimuli, we communicate the issue to each respondent in a pre-determined set of questions. The wording with respect to each question is carefully drawn before meeting the respondents and ideally the same wording have to be followed with respect to each respondent. In systematic stimuli, the second necessary requirement is that the stimuli which are imparted follow a given pattern. This implies that the ordering of the questions should remain the same for each respondent. Under unsystematic stimuli, the patterpistic rigidity is not maintained. Again the questions may not be predetermined, neigh? it is necessary that the wording would remain the same for each respondent.

Depending on the nature of stimuli, the researcher would expect two types of responses from the respondents. A respondent might behave systematically while meeting the query of the researcher. There is a possibility that the respondent would behave unsystematically while answering the questions even if the stimuli are imparted in a systematic manner. Logically speaking, there are four alternative possibilities of which three would describe the alternative formats of data collection. These alternative formats could be described as below :

4.3. Formats of Data Collection

The first is the Formal structured setting. The data collection setting is called formal structured when systematic stimuli are coupled with systematic response. If the systematic stimuli are coupled with unsystematic ·response the setting will be called formal unstructured setting. Unsystematic stimuli coupled with unsystematic response gives an informal setting. It is understood that unsystematic stimuli cannot be coupled with systematic response.

The Alternative Formats of Data Collection

STIMULI	Systermatic	RESPONSE	
		Systermatic	Unsystermatic
		Formal Structured Setting	Formal Unstructured Setting
	Unsystermatic	Impossible Setting	Informal Setting

Formal Structured Setting is stated to be the most preferred setting because the stimulus and the response both being systematic, it is likely that the non-sampling error would be relatively less in the data collected according to this format. However, the researcher may often find that the nature of research is such that the data collected through a formal structured setting would contain less information than what one would have found if say the 'informal setting'. In such a situation the researcher may opt for the same. The other point is that, the formal structured setting may not be attainable for the researcher. It should be kept in mind that the formal structured setting is the outcome of systematic behaviour on the part of the respondents as well. It may so happen that in spite of the best of his effort, the selected group of respondents that the researcher faces behaves in unsystematic way while responding to the stimuli. In that case the setting for data collection would be formal unstructured in nature, even though the researcher wanted to have a formal structured setting.

Forms of Communication and Data Collection Methods

Systermatic	**RESPONSE**		
	Non-Verbal	Oral-Verbal	Written-Verbal
Informal Setting	Participant Observation	Conversations, use. of informants	Articles, Letters, Editorials, Biographies
Formal Unstrucutred Settings	Systematic Observation	Interview Unstructured	Questionnaire open-ended
Formal Structured Settings	Experimental Techniques	Interview structured	Questionnaire structured

As we have discussed, there are three alternative settings under which the primary data can be generated. We did not, however, consider the form of response that the researcher gets while collecting the data in any of the three settings. There could be two basic called *non-verbal* and the other is *verbal*. Verbal mode. of communication is, again,subdivided in two forms, *viz., oral verbal and written verbal.* Non-verbal response is response through body language. A person. while seeing an advertisement may spontaneously jump in expression through body language. This is an example of non-verbal communication. The advertisement is the stimulus in this case and the expression through body language is the response. The point to be noted is that the communication here is indirect.

Since there are three modes of communication and there are three different settings under which the stimuli-response interaction takes .place, there are nine alter-native forms in which the primary data can be generated. The possibilities are described in Table 3.3. While reading the table a researcher could note that every individual form of 8a1a generation has its root in the setting in which the data set is generated and the mode of communication used by the respondent. For example, if we say that the data set has been generated by administering structured questionnaire, the researcher would note that the primary data generated by this method is the outcome of a formal struc-tured setting in which the stimuli had been systematic and the responses were also systematic. A structured questionnaire-would also imply that the response was communicated in written verbal form. (An example of structured questionnaire is given in Appendix. Again, if the data set is generated by experimental technique the researcher will unders and that the mode of communication was non-verbal in this case and the data collection setting had been formal structured, that is, systematic stimuli were coupled with systematic response for generating the data set.

4.4 Collection of Pimary Data

4.4.1 Introduction

Quite often we meet with the situation where available information is not sufficient to cater to the information need. It becomes a major hurdle on the path of research progress.To get1id of this hurdle researcher has to collect information and data that are appropriate for the specific study under consideration. Collected information can be directly utilized; collected data may have to be processed further to extract information. In this sense, information may be defined as processed data. Now, dep~ding on the time, cost and resource constraints one has to draw a suitable plan or action for the collection of pertinent data not available with him/her so far. We may classify data under two subheads, viz., primary data and secondary data. Plan of action for collection of primary data will be different from that of secondary data. By primary data mean those data that are collected for the first time directly by the investigators of a,... research project. There are two ways of collecting the primary data,–either through experimentation or through survey work. In case of secondary data, i.e., the data that leave already been collected by someone else, a researcher has to adopt a directed research method. They may be available in a published form or in an unpublished form. [The main task here is to identify the reliable sources of secondary data. If these sources are not reliable, the reliability of the data will be questionable. Any bias on the part of the compiler gets mixed up with such secondary data. If the researcher is not aware of the methods of collection of such data and the extent of bias and accuracy then the researcher, without himself being biased, will arrive at a biased conclusion or an inaccurate result. It is also important to. examine the suitability of secondary data for the current research work. The objective of the earlier study that has generated the secondary data should be matched with the objective of the current study. In case they do not match the suitability of the secondary data becomes questionable. Even if the objectives match, tabulation pattern of the earlier work may not match with the current-requirement and hence the same has to be made suitable, if needed by undertaking re-tabulation of the secondary data. In case objectives and scopes match with current objectives and scopes one has to examine the adequacy of the secondary data in terms of the covered area, the period of study and the level of accuracy. Thus, to eliminate the risk of error the researcher must be satisfied with the reliability, suitability and adequacy of the secondary data. Saving of time and cost cannot be the sole criterion in taking the help of secondary data.

We have already mentioned that both published and unpublished sources of data are to be scanned. Published data can be obtained from the statistical reports of the government of India, state government and local bodies. Central Statistical Organization is engaged in collection of data and their dissemination on a regular basis. Statistical information collected through well-structured survey work is available from the National Sample Survey Organization. Annual Survey of Industries provides with a rich database on the organized sector of industries of our country. There are various international bodies, foreign governments and global organizations to disseminate & information on the global economic activities and sociopolitical environment. Annual reports of companies, newsletters, booklets are also of use for gathering data for specific companies. Besides, there are journals, magazines, books and newspapers to provide with statistical information on a wide range of subjects Technical reports of various associations, stock exchanges, banks and commercial institutions, published reports of universities and academic bodies and various other public records can be of immense help for a researcher provided the reliability, suitability and adequacy of data satisfy the data-need of the researcher.

Unpublished sources are difficult to trace. But on many occasions they provide with the most valuable information. Unpublished theses may contain rich data, which are not available with published documents mainly because objectives of standard publications and reports are general in nature and objectives of research works are very specific in nature. Besides, diaries, letters, autobiographies are useful unpublished sources of information. Many diaries are giving information on consumption and distribution patterns of earlier days and may not only throw light on an individual's attitude towards life and society but also give an idea about the then socio-economic environment.

4.4.2 Methods of Collection of primary data

There are several methods for collection of primary data. But observational method, interview method and questionnaire method are the three most important methods of data collection that are being frequently used. Let us describe these methods primarily keeping in mind the data needs of descriptive research, diagnostic research and exploratory research. In hypothesis testing research, also known as experimental research, observations are generated by the. experiment and the objective of the experiment is to determine the possible relationships among the phenomena under study. In survey method one observes the phenomen, which exist in the population and the existing relationships can be· studied from the collected data. Thus, method-wise there may not be any difference when we talk about observational method for experimentation and the same for survey work. But methodology wise they differ.

4.4.3 Observational method

By observation, under the observational method; we · mean scientific observation made under systematic planning and recording. To be more precise, we use the term scientific observation to differentiate it from the day-to-day observations that we all make in our daily lives. In scientific observation the instrument for making observation must measure what it aims to measure and must measure with consistency. In place of depending on the respondents' views, investigators directly observe the phenomenon. As a result, respondents' bias gets completely eliminated if observation is made with perfection. There is another advantage with this method. Since the investigator makes observations what we record is the current happenings and not what the respondents plan to do in future or did in the past. It has minimum problem of nonresponse that may arise in other methods due to unwillingness on the part of the respondents.

To be systematic in taking observations, an investigator must preplan about what is to be observed, how to observe and where and when to observe. Especially for descriptive research structured observation is preferred to have uniformity in the. condition of observation and in the definition of experimental units. However, unstructured observation may have to be. taken into consideration when preplanning is difficult. This happens in the case of exploratory or formulative research where ideas remain vague initially and become clear gradually as the research work progresses.

Sometimes observations are made by the investigator by directly participating as a member of the group to be observed. The objective of this participant observation is to have the same experience as the members of the group experience. In that case natural behavior of the group members can be closely observed and recorded. Investigator may collect rare information which otherwise cannot be obtained. Only thing that the observer should keep in mind is the importance of objectivity in the recording system. In that case emotional participatfon will not be able to disturb the rational objective of participation. The other alternative to participant observation is non-participant observation, i.e., where the presence of the investigator cannot be felt or may remain unknown to the group members, he/she is observing.

Observational method, in spite of its wide applicability, has a few limitations. It is an expansive method as observations are to be made by skilled persons from the actual field of study. Sometimes, even with repeated visits a unit or a person may not be reachable and this may lead to incompleteness in the collected data because those who are not reachable may possess a common characteristic. There may be situations where subjective interpretation is needed to make an observation. In that case the subjective bias get, mixed up with the collected data. This problem arises when uncontrolled observations are made in a natural setting.

4.4.4 Interview method

By interview one means verbal responses against verbal stimuli. The responding individual is called interviewee and the stimulator is known as interviewer. In case of a face to farce response we refer it as personal· interview, In case of telephonic interview or video interview we call it distant interview.

The method of collection of information through interview may be either *carried out directly* by asking the person of interest or carried out indirectly by asking questions and receiving responses from other persons who can give information on the person of interest. Mostly a structured set of questions is prepared and asked and responses, recorded by using some standardized techniques. In a perfectly structured interview even the sequence of asking questions remains fixed. As a result, such an interview requires a comparatively lesser skill on the part of the interviewer increasing the feasibility of this method to a great extent. However, more in-depth information can be collected through unstructured interview. It provides flexibility in questions to be asked and their order. Interviewer enjoys the freedom of asking additional questions if the situation so demands or may drop a few questions if he/she does not feel their relevance in a particular situation. For unstructured interview time needed is more, comparability is less and requirement of skill on the part of the interviewer is more.

Even then, unstructured interview is mostly favored for. exploratory or formulative research. It gives the scope for identifying new dimensions of the problem to be studied and is extremely useful when the researcher is not very clear about the types of responses.

Unstructured interview may be focused, clinical or non-directive. In an unstructured focused interview the task of the interviewer is to concentrate on a particular issue and see that the discussion does not go out of track. Interviewer has the freedom to ask additional questions of his/her choice without becoming defocused. In an unstructured clinical interview one is mainly interested in individual's experience on the subject of interest. To achieve this aim, interviewer is left free to frame on-the-spO questions and get response therefrom. In an unstructured nondirective interview there is hardly any direct question to be asked. Interviewer is to motivate the respondent to opine on the suggested topic and should give maximum freedom to the respondent.

Interview method of data collection works well if the investigator is friendly, intelligent, conversational and- rational. A friendly and conversational interviewer can over come resistance, if any, obtain personal information with ease and can note down the spontaneous reactions which otherwise may go unrecorded. Intelligent and rational presentation cf questions can help in getting additional in-depth information. Depending on the academic attainment of the interviewee the interviewer may adjust the wordings of the questions and can get a warm response.

There are a few limitations of the interview method. Except for distant interview, the cost of data collection is generally very high. It is also a very time consuming affair. Sometimes very presence of the interviewer may over stimulate the interviewee who in turn, may give imaginary information and disturb the whole process. One special problem, we have already mentioned, is the problem of contacting some respondents. Doctors, businessmen, government officials are mostly busy with their personal or official work and may not be reachable at all.

4.4.5 Questionnaire method

Collection of information through questionnaire (or schedule) is a very popular method. Generally, a questionnaire is prepared and sent to the concerned responding units or individuals with a request to fill up the questionnaire and return to the sender within a stipulated time. A Questionnaire is having at least two blocks of questions. The first block is the identification wherein the respondent gives information on his/ her personal identification as per the questions. Name, address, sex, age etc. are most commonly used questions for this block. This block is meant for verification of information if the responses fail to meet consistency check. It is also needed for sample checking. A part of the information provided under this block is kept confidential. Other parts may be used for cross tabulation purpose. Age apd Sex are the two items that may be of interest from both tabulation and analysispoints of view. The other block is the information or response block where supplied information throw light on· the research subject and if needed may' be quoted as individual information. Response block must be careiully frameJ Questions must be clear, to the point, and logically arranged. No special knowledge on the point of the respondent can be assumed. All the terms should be well defined and self-explanatory. Individual questions may be closed or oper type. In case of a closed type question the respondent will be provided with a list of alternative answers and the respondent has to pick up one of those answers. For this, answers provided in the questionnaire must be both exclusive and exhaustive type. Otherwise there will be confusion in the minds of the respondents. Open type questions are those where exhaustive list of answers are not available or cannot be included to save space. Along with a few specific answers, one common type answer is added to include all possible answers. Respondents are requested to specify their answer if they happen to favor the common type answer. Ordering of the questions should be such that the sequence of questions can be viewed as a funnel. They should, gradually, become pinpointed to extract the specific answer starting from a general setup and general/broad answers.

Structured questionnaires are those questionnaires where majority of the questions are of closed type and questions are arranged in a fixed sequence. If majotity of the questions are of open type we may call it an unstructured questionnaire. Structured questionnaires' are easy to mail and to get response without the help of an investigator. The analytical treatment can also be preplanned and easy to employ. The cost of collection of data is minimum. But the non-response may also be high. In fact, if questions are simple, easily understood, do not put any unnecessary strain on the on memory or intellect of the respondent and do not enquire about personal matters like wealth, character etc. Then the chance of response will be high. Otherwise non-response will be on the higher side. To ensure active cdoperation from the side of the respondent researcher may enclose with the questionnafre a brief write up explaining how to fill up the questionnaire. It is also desirable that before the start of the questionnaire the objective of the res.earch be printed, indicating the scope of the research.

It may be mentioned in this context, that methods for collection of data through questionnaire and through schedules ore similar in nature. Schedule is a proforma containing a set of questions like a questionnaire. While questionnaire method is mostly a mail method,schedule method is a direct enumeration method. Enumerators meet the respondents and fill up the schedules based on the responses received. Enumerators explain the questions to the respondents and are trained to handle questions from the side of the respondents, if any. Quality and skill of the enumerators are very-important in schedule method so that non-response can be reduced to a great extent. In fact questionnaire method is very cheap but non-response is very high. On the contra&, schedule method is very expensive but non-response is very low. While questionnaire method is slow as the researcher should give sufficient time to the respondents for personally filling up the questionnaire and mail it to the sender, the schedule method is very fast as the enumerator notes down answers on one sitting. It is also clear that questionnaire method can be used for educated respondents. There is no such restriction for schedule method because respondents do neither read nor fill up the questionnaire. Incomplete response is also very high in questionnaire method. Incompleteness is rare is'cse of schedule method. Under questionnaire method one can cover a very wide area. This wide coverage is nearly ruled out in schedule method. Thus, there are many advantages and disadvantages of both these methods. Depending on the nature of the research problem, availability of fund and time, level of accuracy required and the requirement of application tools one may select either a questionnaire method or a schedule method.

4.4.6 Other methods for collection of data

There are a few other methods for collection of data. They may not have a universal appeal. Nevertheless, they are interesting and sometimes they are more appropriate and feasible. Consumer panels, warranty cards, store audits, pantry audits, mechanical devices and projective techniques are of use for collection data for business research mainly. Let us, briefly, explain these methods along with their scopes.

Consumer panels are sets of consumers who agree to maintain detailed information of their daily consumptions and share those information with the investigators. These panels are of two types, viz., *transitory consumer panel and continuing consumer panel.* A transitory consumer panel is formed to collect information for a period of transition and is helpful for before and after with out control type of experimental designs, and also for repetitive surveys spreading over a fixed period of time. Initial information is collected from the members of the panel before a phenomenon takes place and final information is collected from the same panel members after the phe-nomenon has happened. In a continuing consumer panel the period of study is indefinite and the information is collected from the panel members on-a periodic basis. One may use direct interview or mail method for collection of data. Transitory consumer panels are useful for examining the effects of advertising, changes in policy decisions, effects of product modification where one-time act is involved. A continuing consumer panel is used for examining the consumer behavior over a long period of time, their brand selection, brand loyalty and brand switching behavior. It is also being used to gauge public opinion, radio or TV listener-ship, channel popularity and so on. For a. transitory consumer· panel drop out problems may not be very significant. For a continuing consumer panel investigator has to take due measure for retaining the consumer panel in tact and device a mechanism for replacing the unwilling members by similar but willing consumers.

Warranty cards are those cards that are provided with consumer durables. Consumers are asked to fill up those cards and either post them or hand over them to dealers for getting the warranty facility. In this process marketers of the consumer durables generate consumer database for future action, if any. Researches may directly obtain this database from the marketer and get a nearly ready sampling frame.

Store audits, conducted by the marketers through their salespersons, can throw light on purchasing pattern and can be used for estimating the market size and market shares. Store audits are conducted periodically and are comparable with consumer panel data.

In the later case consumers report their consumption pattern. But in the former case, retail stores provide the information on purchase pattern. Store audit data are authentic and available on a regular basis without any problem of nonresponse.

Pantry audits, conducted by the investigators, can provide wit consumption figures at the consumer level for the complete basket of goods. It may be supplemented by direct interview to· note down the reasons and plans of the consumers. At the end, investigator gets an idea about which types of consumers prefer which types of brands, what features of the brands make the consumers happy or unhappy, what are the average levels of consumptions, what are the future plans etc.

Mechanical devices have made inroad into the field in the recent past for collection of data by indirect method. For example, eye camera can record the focus of the eyes of the consumer or the respondent. as the case may be. Jf one knows the focal point of interest of an advertisement future designing may be easier and more effective. Audiometers can be of use to know the types of TV programmes preferred by the viewers. One may use psycho-galvanometer to measure the excitement from visual stimuli.

Projective techniques, also known as indirect interviewing techniques, can also be of help for collection of data on motives, urges or intentions, which are usually difficult to record through direct interviews. In case of a direct interview, respondent may try to hide the facts or may be unable to respond. In the indirect interviewing technique the individual's responses to the stimuli are interpreted through psychological conceptualization made before in the literature. Important projective techniques are word association test (to note· down the first word that comes in the mind of the respondent when the investigator reads out award), Sentence completion fest (to ask the respondent to complete an incomplete sentence by using his/her own. thought on that subject), Story completion test (to ask the respondent to complete an incomplete story), Verbal projection test (to know from the respondent why other persons behave in particular way), play technique (to observe how respondents act in a given role assigned by the investigators) and pictorial techniques (to ask the respondents to describe pictures shown to them in an ambiguous way).

4.5 Selection criteria

Given a plethora of methods for collection of data, a researcher has to make a final selection of a method for his/her research activity. This may be a difficult task because consequences of this decision have for reaching effects on the outcomes of the research study. It is therefore, necessary to look into several issues before arriving at a decision on the rhethod for collection of data.

The first thing to be matched is the objective of the research study with the objective of the method for data collection. We have classified research objective under four genera) heads, viz., exploratory or formulative research, descriptive research, diagnostic research and hypothesis-testing research. We have already indicated that if the objective of the research is hypothesis-testing the concerned design must be an experimental one. Observational method may be a pref erred choice for this case. However, one cannot completely rule out the possibility of undertaking other methods. Thus, objective alone may not be sufficient to help us arriving at a conclusion. Problem of choice becomes more acute when one decides to go for survey method. There are different sub-methods of survey method each having its own advantages and limitations. It is therefore necessary to search for other criteria as well.

Precision requirement is the other important issue where the res·earcher should give attention. When precision requirement is high direct observation or direct interview Will be favored over questionnaire method. The aimed at coverage is also an important consideration in this context. If wide coverage is needed in short time, questionnaire method may be favored over direct interview, observational method, and schedule method of enumeration. Problem arises when the precision requirement is high and coverage requirement is also high. In such a case researcher has to strike a balance between precision and coverage and the method of data collection may also depend on the time at disposal of the researcher. High precision and wide coverage demand for more time and more fund. The researcher should take into consideration the fund available for this purpose. If one can employ. more fund one may be able to reduce the time requirement by employing more investigators and more mechanized system of data collection. Fund, in fact, is a major consideration that limits the choice set of methods for collection of data.

4.6 Administration of surveys

Survey method being a major method for collection of data let us indicate the basic steps to be followed for administering a survey work. Since we need to refer to a research activity for presentation of survey administration, it will be sufficient if planning stage, execution stage and analysis and reporting stages are covered in our discussion. The stage of evaluation need not be touched upon.

The **planning stage** consists of the following sub-stages :

1. Defining the purpose of the survey. .

   The purpose of the survey must be clearly spelt out so that investigator can be clear about it and can make use of it whenever some problem is faced at a later period.

2. Taking note of the available resources.

   The investigator must take note of the resources available under his/her disposal. The entire planning, execution and analysis cum-reporting exercise depends heavily on the availability of the resources. Most important resources are fund, manpower and time. Fund indicates the financial limit within which the survey is to be completed. Time indicates the permissible period within which the survey work is to be finished. Manpower indicates the level of skill available and number of hands available for carrying out the entire work as per planning.

3. Defining the aimed at level of accuracy.

   The level of accuracy desired by the researcher will be the other guiding force in planning the survey work. This level of accuracy should be so targeted that the purpose of the survey can be met within the available resources.

4. Defining the population. .

   The population, also know as the universe, must be categorically defined. The results of the survey will be applicable to this entire population and hence the geographic, demographic and other boundaries of the population must be stated without any scope for confusion.

5. Determination of the data to· be collected.

   The purpose of the survey and the objective of the research will all indicate the data requirement. Accordingly, the researcher should specify the characteristics to be observed or measured, opinions to be collected and attitudes to· be noted during the survey work. The questionnaire or the schedule of enquiry can be framed keeping in mind the above decisions. It is desirable that a draft questionnaire be prepared and tested over a small group of individuals. This helps in identifying the gaps and flaws in wording, sequencing and coverage of the draft questions. Based on the ideas so gathered the questionnaire is to be revised and finalized for future administration.

6. Selecting the method for data collection.

   The selection of the method for data collection is very important, as the method is a direct determinant of the level of accuracy, cost of survey and extent of nonresponse. The first choice is between interview method and mail method. Cases where observations are to be made, the method of measurement must be indicated. For example, agricultural production may be measured either exactly or approximately through eye estimation. The types of instruments needed for making measurements should be indicated.

7. Defining the sampling unit.

   The ultimate unit of the population, which is to be studied for the purpose of the survey, is known as sampling unit. For example, whether a family or a member of the family is to be studied must be clearly spelt out. The complete list of the sampling units, known as sampling frame, must be prepared to help the sampling scheme. There are situations where a complete frame is not available. In those situations researcher has to define a hierarchy of sampling units to reduce the task of listing sampling units.

8. Designing the survey.

   Survey design is the most crucial stage of planning. A researcher must decide, during this stage the probabilistic sampling procedure to be adopted, restrictions to be imposed, if any, like with replacement or without replacement and optimum level of flexible variable, if any. It may also be finalized whether a pilot survey is to conducted before the actual execution of the main survey.

9. Preparing the list of sampled units.

   Using the random number table or other devices the researcher must select the sampling units and frame a list of those sampled units. Indications should be given about how to identify those units during the fieldwork.

10. Training of field investigators.

    The investigators should be properly trained for the survey work to be under taken by them. This involves exposure to purpose of the survey, identification bf the sampled units, definitions of the technical terms used in the questionnaire etc.

    The next stage is the execution stage which involves identification of sampled units in the field, filling up of the questionnaire and remedial measures to be taken in case of contingencies.

The **analysis and reporting stage** includes the following stages.

1. Scrutiny of data.

   The collected information through filled up questionnaire should be checked for consistency and completeness. In case of incomplete questionnaire or inconsistent information the questionnaire under consideration should be sent back to the field for fresh collection of information. Scrutiny of data also involves cleaning and dressing of data.

2. Tabulation of data.

   The qualitative variables must be provided with code numbers and the quantitative data field must be standardized and the entire information may be summarized either through hand tabulation or through machine tabulation.

3. Statistical analysis.

   The primary tables generated during tabulation of data are to be put to statistical analysis. This means estimation of population measures, determination of the extent of error and carrying out of statistical inference for testing the working hypothesis. Derived tables can also be prepared from the primary tables to present pinpointed information on the population characteristics.

4. Writing of report.

   The final report should contain detailed description of all- the stages of work and should present all the statistical and other information. Proper interpretation of the statistical inference must be added and final conclusions and recommendations are to be included.

5. Storing of information for future use.

   After the completion of the report writing, adequate care should be taken to store the raw data along with the primary and derived tables. These information should be so stored as to make use of in the future survey works/research studies.

4.7 Summary

The data set is generated by the interaction between 'stimuli' and 'object'. The results of the instruction are captured in the form of data. Information is processed data. Quite often the researcher has to collect data for converting them to useful information. Depending on the time, cost and resource constraints one has to plan for collection of pertinent data. Data are of two types, viz., primary data and secondary data. Primary data are those data which are collected by the researcher for the first time either through experimentation or through survey work. Secondary data are those data which have already been collected by others. The researcher has to adopt a directed search method to avail the secondary data. While collecting the secondary data?-'researcher must be satisfied with reliability, suitability and adequacy of the secondary data. There are both published and unpublished sources of secondary data. Statistical reports of the central and state governments and of the local bodies are the major sources of published data.

To collect primary data one may take course of the observational method, interview method, questionnaire/ schedule method and a few other methods. Under the observational method a researcher directly observes the phenomenon of importance and thereby eliminates the respondents' bias. For descriptive research, one may like to opt for structured observation. For exploratory or formulative research unstructured observation is suitable. Further, the investigator may have the experience of a group while observing the same. This is known as participant observation. Where the presence of the investigator cannot be felt the method of observation is known as nonparticipant observation. Unfortunately, the observational method is very expensive. Sometimes, subjective bias may get mixed up during observation. Sometimes a unit or a person may not be reachable even after repeated attemps.

Interview method records the verbal responses against verbal stimuli. The method of data collection may be either based on face-to-face response or based on distant interview.

Collection of data may be viewed as direct if the respondent is contacted to give responses personally. In other cases, we go by indirect responses. For the ease of tabulation and subsequent analysis, one may consider structured interview based on a preplanned set of questions of fixed sequence. For in-depth information, one may, however, prefer unstructured interview to give more freedom to the inves-tigator for asking the questions. Interview method is very effective if the investigator is friendly, intelligent; conversational and rational. Cost of data collection under interview method is usually on the higher side. It is a time consuming method too.

Questionnaire/Schedule method goes by either the mailing of the questionnaire to the respondents or visiting the respondents with the schedule of enquiry. Questionnaire or schedule must be made of at least two blocks, viz., identification block and information or response block. As suggested in case of interview, questionnaire may be either structured or unstructured types. Coverage of area is very high and also the nonresponse in case of questionnaire method. In case of schedule method nonresponse is nearly absent. Cost of undertaking schedule method is very high while cost of undertaking questionnaire method is very low.

Other methods of data collection include consumer panels, warranty cards, store audits, pantry audits, mechanical devices and projective techniques.

Selection of the suitable method for collection of data depends on the objective of the research, precision requirement and availability of resources like fund, manpower and time. In case the researcher opts for a survey work for collection of data he/she must meticulously plan for survey administration, must carefully execute the pan during VG fieldwork and decide about the analysis of data for final presentation of the report.

4.8 Questions

Long answer type questions :

1. Enlist different methods for collection of data. Which one is the most suitable method, according to you, if the researcher is interested to examine the awareness of a brand in the product field of dental cream?
2. Describe different sources for collection of secondary data.
3. For a descriptive research covering a wide area, which method do you propose to adopt for collection of primary data?
4. If Interview method and observational method are very expensive why do the researchers sometimes opt for these methods for collection of primary data?
5. Explain the differences between questionnaire method and schedule method. Which of these two methods do you prefer for in-depth analysis of commodity baskes of consumers' in a small town?

6. Explain with an example the use of consumer panels in collecting data on consumption pattern.
7. Indicate the basic considerations based on which a researcher- should select the method for data collection.
8. Describe, in brief, the basic steps to be followed in the administration of a survey.

Short answer type questions :

1. Explain th~ terms /primary data and secondary data.
2. When do you prefer structured observation and when do you prefer unstructured observation?
3. Describe the limitations of observational method.
4. What is interview? When do you prefer interview method?
5. Compare structured interview with structured observation.
6. When do you recommend unstructured interview?
7. Explain the terms closed end question, open-end question. Give examples.
8. 'Transitory consumer panels are used to collect information for a period of transition.' Explain.
9. Explain the use of store audits as a method for collection of data.
10. Write a note on projective technique.
11. Enlist the sub-stages of planning stage for administration of surveys.
12. Describe the analysis and reporting stage for survey method.

Objective type questions :

Indicate whether the following statements are true or false.

1. Secondary data are data,/which are collected for the second time by the researcher.

   True □ False □

2. Savings of time and cost are the sole criteria for taking the course. of secondary data.

   True □ False □

3. For descriptive research structured observation is preferred.

   True □ False □

4. In participant observation respondents participate in a meeting.

   True □ False □

5. Non-participant observations are made when respondents refuse to participate.

   True □ False □

6. Personal interview refers- to interview on personal matters.

True □ False □

7. In a structured interview questions and their sequences remain fixed.

True □ False □

8. Unstructured interview may be focused, clinical or nondirective.

True □ False □

9. Presence of the interviewer is to over stimulate the interviewee.

True □ False □

10. Information or response block provides us with the· identification of the respondent.'

True □ False □

11. When question become gradually pinpointed we term the sequence as a funnel.

True □ False □

12. In a continuing consumer panel, panel members provide information before and after the occurrence of a phenomenon.

True □ False □

13. Warranty cards can be used to generate consumer database

True □ False □

14. Audiometer is used under interview method.

True □ False □

15. Projective technique is based. on projection of eye camera.

True □ False □

16. Sentence completion test is a very popular tool of questionnaire method.

True □ False □

17. Population is also known as universe.

True □ False □

18. Sampling frame is the frame where sampled units are kept.

True □ False □

19. Scrutiny of data involves cleaning and dressing of data

True □ False □

20. Derived tables are prepared from the primary tables.

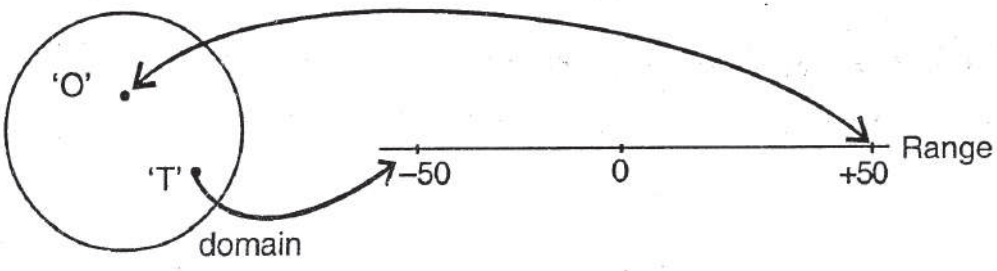
True □ False □

Unit 5 □ Techniques of Measurement and Scaling

Structure

5.1 Introduction

Measurement is a process of mapping. It maps different aspects of the members of a domain on to some other aspects of a range set where range is. mostly the real line or a segment of the real line. The process of mapping works on the principle of correspondence. For example, environmental opportunity and environmental threat, denoted by 'O' and 'T', are the members of the domain set. We may assign a score +50 to 'O' and -50 to 'T'. This rule of assignment of score may be viewed as a rule of correspondence. The range set contains -50 and 50 as the two scores. A neutral situation may correspond to score zero. If there is one to one correspondence, as in the above case, then the score +50 will imply opportunity. in the domain set, the score -50 will imply threat in the domain set and the score o will imply neutral state in the domain set.



Development of a measuring instrument is a complex and difficult task, especially when we need to measure the attitude of say, customers. For example, it is easy to measure some properties like height and weight but it is not that easy to measure the traits like risk-oriented, risk-aversive or risk-planner. It is also absolutely necessary to know the measurement scales to decide upon the appropriate statistical analysis. All types of scales do not permit the use of all types of statistical analysis.

5.2. Measurement Scales

Measurement scales can be grouped under four broad heads: nominal scale, ordinal scale, interval scale and ratio scale. Under nominal scale we transform categorical data, i.e. qualitative or descriptive data into numerical data through assignment of numerical codes for various, categories. Consider the case of job categories: Government service, non-governmental service, business, self-employed, professional and others. These desrlptions of the job-status, which an enumerator may observe during a survey work, are not as such quantifiable. However, we may assign codes 1, 2, 3, 4 and 5 respectively to these five categories. These codes are numerical codes as these are expressed in terms of numerical digits. But no numerical operation can be carried out with these codes. The difference between non-governmental service code and government service code is same as the difference between self-employed pro-fessional code and business code. But status-wise or otherwise these differences are not same. Basically values on the nominal scale are numerical in names. Since these numerical codes of the nominal scale are playing the roles of labels only, we cannot add or subtract these numerical codes. We can count the frequency of occurrence of a particular code, i.e., the number of objects that fall into a particular codecategory. Some statistical analyses are allowed to be undertaken. One such analysis is the chisquare test for association of attributes.

Under ordinal scale one transforms the performance data into numerical data by expressing the order of preference. For example, consider the subject of brand aware- ness. Let there be three brands, viz., Lux, Dove and Cinthol and the objective of a survey is to examine the extent of brand awareness. According to the order of preference we may assign ranks to these brands. Suppose rank is 1 for maximum awareness, rank is 2 for next best awareness and rank is 3 for minimum awareness. Then, if Lux is of rank 1, Dove is of rank 2 and Cinthol is of rank 3, we may say that the brand awareness values. of these brands on the ordinal scale are 1, 2 and 3 respectively. It is easy to note that objects measured on an ordinal scale can be ordered. But we cannot claim that Dove is as far below Lux in brand awareness as Cinthol is below Dove. This is because rank 2 minus rank 1 is not necessarily equal to rank 3 minus rank 2. Thus, numerical operations such as addition, subtraction, multiplication and division cannot be applied on numerical values generated by an ordinal scale. In case of measurements from an ordinal scale, one may calculate medians, percentages and rank correlation coefficients for describing the population in terms of statistical measures. Nonparametric tests which are constructed through ranks, can be used without problem.

Under interval scale, we have a constant unit of measurement. But this unit of measurement is arbitrary. Interval scale corresponds to what one can think of as a measurement. Only the zero point and the unit of measurement are arbitrarily fixed. The most common examples of interval scale are Fahrenheit scale and centigrade scale. We measure temperature on both these scales. These scales have different zero points and also different units of measurement. What is o" centigrade on the centigrade scale is 32 Fahrenheit on the Fahrenheit scale. Further, 100 units on the centigrade scale is equivalent to 180 units on the Fahrenheit scale. So if for the same temperature of an object measurement on the centigrade scale is C and that on the Fahrenheit scale is F then

$$\frac{C-0}{100}=\frac{F-32}{180}\text{ or, }\frac{G}{5}=\frac{F-32}{9}$$

Almost all the commonly used statistical measures and operations can be employed on data measured through an interval scale. Arithmetic mean; standard deviation and correlation co-efficients can be calculated without any problem. Techniques such as regression analysis, factor analysis can also be applied on interval data.

Ratio scale is an improvement over interval scale in the sense that it possesses a unique zero point. The unit of measurement is constant but need not be fixed. When measure the height of individuals either in feet or in meters we employ ratio scale. For both feet and meter scales absolute zero point is same. However, units of measurement are constants but different. One inch is equal to 2.54 centimeter. For ratio scale, if an object is twice longer than another object then the ratio of the corresponding measures on any ratio scale will be 2 : 1. Ratio. scales of measurement are very common in physical science and are rarely used in behavioral science. All the statistical measures can be calculated on a ratio scale and all the statistical tests can be applied on data obtained from a ratio scale.

5.3. Development of measurement tools

Quite often in social science researcher has to develop own measuremont tools to measure attitudes and behaviors of individuals, as there is no standard reasurement tool. The developent of a measurement tool has to pass through four stages. First stage is, the concept development. The second stage involves specification of the dimensions of the concept the third stage calls for selection of indicators for different dimensions and the fourth stage deals with formation of an overall index as a measurement tool. Let us describe these four stages in their above-mentioned sequence.

Concept development means deciding about the major concepts that are directly related with the study and making a clear understanding about those concepts. For example, if we are interested to examine the efficiency of an organization or a product field we have to develop the idea about efficiency. Major concepts related to efficiency may be profit, average cost of production, competitive position and extent of innovation. There may be other concepts also. However, if we would like to restrict our study to these four measures of efficiency then these four will be the underlying concepts. We need to develop clear understanding about these four concepts. For example, for examining profit we may consider economists' definition of profit, minimum profit concept, profit after tax concept or actual profit concept. We need to be clear about which profit concept we would like to pursue.

Once the. concepts are clearly understood the researcher has to identify the dimensions of the concept and specify them in clear-cut terms. Sometimes intuition plays an important role in deciding about the different dimensions of the concept. Sometimes, one may work back through logical deduction and arrive at the important dimensions. For example, if the concept under consideration is brand identity, we may work back to deduce the major dimensions of brand identity as physique, personality, culture, relationship, reflection and self-image.

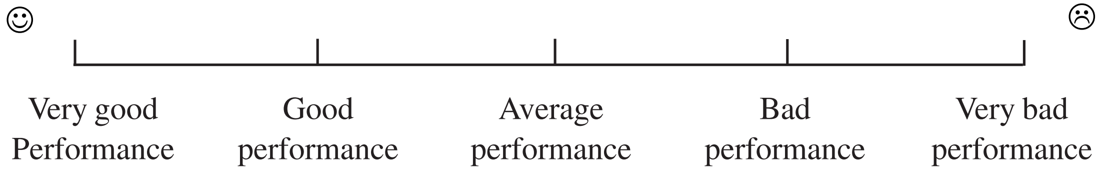
After the specification of the different dimensions of the concept, a researcher must develop measuring indicators for each dimension. These indicators may be scales of measurement, sets of questions or the devices as may be developed by the re- d searcher. For example, if we consider competitive position as a concept and one of its dimensions of measurement as market share then the measuring instrument will be sales in volume and values for the company under consideration and for the total industry.

The end step aims at combining the individual indicators into a workable whole. This combination of indicators is also known as an idex number. Index number, being an overall measure, can help the researchers in arriving at the final conclusion. For example, if we know the relative importance of different concepts one may consider an weighted average of the indicators of the concept values and obtain the overall measure of importance.

5.4. Scaling Techniques

Scaling techniques are many in number and need to be judiciously selected. Results of the research and the level of accuracy highly depend on the scaling technique adopted for measurement. For problems like measurement of attitudes and opinions the availability of .a valid measurement is virtually under cloud. It is the duty of the researcher to find out a valid measurement. Validity and reliability of the measuring instruments are to be ascertained first before these are put to use. The actual use is to assign numbers to different states of attitudes or opinions of individuals. An individual may be directly placed on a scale or indirectly placed on a scale based on the responses given against a set of questions. Let us describe a few important scaling techniques which are broadly classified under two general heads, viz., rating scaling and ranking scaling.

Graphic rating scale is an important rating scale where various judgmental points are placed along a line to form a complete continuum. For example, a judgemental point may be of the type ‘‘average performance’’ when we seek the opinion or a group of customers on the brand achievement of a specific brand. Five judgental points as very good performance, good performance, average performance, boad performance; and very bad performance' may represent the full continuum There is no restriction on the number of judgmental points to be used for measurement. It may start from two points and end up to any number of points. ‘‘Good’’ and ‘‘bad’’ or ‘‘right’’ and ‘‘wrong’’ are the examples of two point graphic rating scale. During the application of this scale the evaluator puts a tick mark against the right judgemental point, as demonstrated below.



One may use a box in place of a line in a similar way.

Itemized rating scale provides with a sequence of statements, the sequence being an indicator of progressive or. regressive adherence to a property. The respondent has to select one of the statements that is closest with the respondent's personal view. For example, consider the case of a brand 'A' and the enquiry about its leadership in the market. Consider the following sequence of statements, the first statement confirming the leadership of brand 'A' in the market and the last one disowning the same. Inbetween statements are regressive in nature :

- Brand 'A' is always the market leader.
- Brand 'A' is often the market leader.
- Brand 'A' is sometimes the market leader.
- Brand 'A' is infrequently the market leader.
- Brand 'A' has never been the market leader.

If the respondent's evaluation is closest with statement number four then he/she will circle the fourth option.

In this itemized rating scale the listed statements may not describe the actual view of the respondent. But the tabulation work and the subsequent analysis become easier. It respondents are allowed to write their views the'tabulation work may become extremely difficult.

These rating scales are easy to use, require less time and enjoy wide applicability. For itemized rating scale, development of the statements is a difficult task as one has to foresee the possible alternative views which are exhaustive but not many in number. There are some problems assocrated with· rating scale. The respondents are sometimes unwilling to make extreme comments/ratings. This attitude results in high heaping near the central point. This type of error is, therefore, described as the error of central tendency. Another type of error known as the 'error of hallo effect' arises when the respondents are asked to rate different items without having clear idea about them. In those cases rational judgment takes the back seat and the very purpose of rating scale gets defeated.
Ranking scale is an alternative to rating scale. It works based on the principle of comparison among the similar objects and the· respondent is asked to make his/her choice. There are two types of ranking scales one based on the method of pair-wise comparison and the other based on rank.
Pair-wise comparison method records the views of the respondent regarding the preferred choice between two objects. In case there are n such objects then the respondent has to examine all the ${}^{n}C_{2}$ (= n(n-1)/2) combinations of objects and make ${}^{n}c_{2}$ preferred choices. Thus, ordinal data are generated which can be used to develop scores on an interval scale. Thurstone has proposed the law of comparative judgment and using that law such a transformation from ordinal data to interval data can be undertaken. Let us describe thts procedure.
Let
f_{ij}= the number of respondents perferring the i th object over the j th object, i = 1, 2, ..., n, j=1, 2, ...n (≠ i)
f_{ii} = 0, i=1, 2...., n,
Then

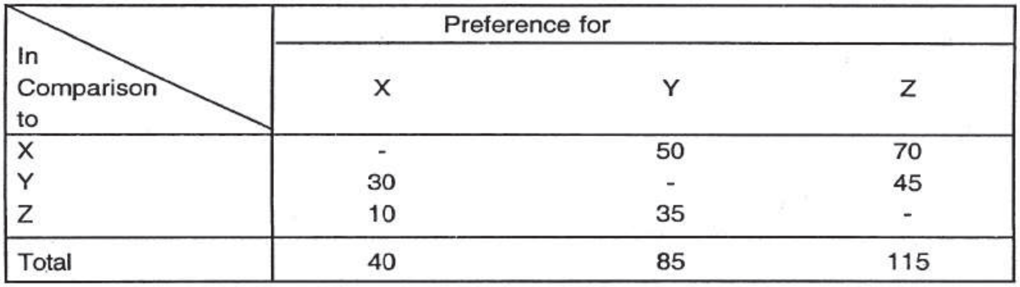
$C_i = \sum_{i=1}^{n} f_{ij}$, the total number of choices for a given object j, j=2., 2,....., n.
We can calculate, M_j, the j-th mean proportion as
M, =[C_1+ N/2]/(nN), j=1,2, ..., n, where N is the number of respondents.
Let

τ_j = standard normal value corresponding to the mean proportion, M_j, j = 1,2,...., n,
Then

$$\left\{\tau_j - \min_{1\le j\le b} \tau_j\right\}$$

is the value of tth object on an interval scale measuring the characteristic based on which the comparisons have been made.

Let us consider an example tor a clear understanding. Consider three brands X, Y and Z of a product field. 80 respondents are asked to compare them pair-wise. There are three possible pairs (X,Y), (X, Z) and (Y, Z). Following table summarizes the responses of the respondents.



In Comparison to	Preference for		
	X	Y	Z
X	-	50	70
Y	30	-	45
Z	10	35	-
Total	40	85	115

This table indicates 50 out of 80 respondents preferred Y to X, 70 out of 80 respondents preferred Z to X and 45 out of 80 respondents preferred Z to Y. Then the number of choices for brand X over other brands is given by (30+10)=40 as given under column totals. Similarly, number of choices for brand Y over other brands is (50+35)=85 and number of choices for brand Z over other brand is (70+45)=115. Thus, 40, 85 and 115 are the Ci values for brands j as X, Y, Z respectively. Corresponding mean proportions are calculated as follows :

For brand X,

$$= M_x = \frac{C_x + N/2}{nN}$$

$$= \frac{40 + 80/2}{3(80)}$$

$$= \frac{80}{240}$$

$$= 0.333$$

For brand Y,

$$M_y = \frac{C_y + N/2}{nN}$$

$$= \frac{85 + 80/2}{3(80)}$$

$$= \frac{125}{240}$$

$$= 0.521$$

For brand Z,

$$M_z = \frac{C_z + N/2}{nN}$$

$$= \frac{115 + 80/2}{3(80)}$$

$$= \frac{155}{240}$$

$$= 0.646$$

Then the corresponding τ_j values can be obtained from the standard normal table. For brand X, the mean proportion is less than 0.5. Hence the τ_j value is negative and the tabulated value comes out as -0.432. Thus, τ_x = -0.432. Similarly τ_j value for brands Y and Z are respectively τ_y = 0.053, τ_z = 0.375. Since the Min τ_j = -0.432 we have the final values of brands X, Y and Z as

{ -0.432 - (-0.432)} = 0,
{ 0.053 - (-0.432)} = 0.485, and
{0.375 - (-0.432)} = 0.807.

respectively. Thus with respect to the characteristic based on which comparisons are made the values of brands X, Y and Z on an interval scale are respectively 0, 0.485 and 0.807.

Rank order method is the other method of ranking scales. Under this method, the respondents assign ranks for each of the objects under consideration. The best object will receive rank 1 and the worst object will receive rank n where there are in total n objects. Then the rank values for each object are added to arrive at the rank sum for each object. Based on the rank sum one may finally rank the objects based on the views of all the respondents.

This is an easier and faster method when compared with pair-wise comparison method. As such the values we obtain are on an ordinal scale and not on an interval scale as in the previous case. But transformation of ordinal data to interval data is possible.

5.5 Construction of scales

For measuring the attitude or opinion of the respondents, there is no standard scale available for measurement. The researcher has to develop his/her own scale. This makes the construction of scale an important task in doing research in the social science. Let us describe the differential scale of Thurstone type, Summated scale of Likert type and semantic differential scale for use in such cases.

Differential Scale using consensus amongst a panel of judges is based on a set of items relevant to the topic of research. Given a large set of items in the form of statements, judges are asked to group them in eleven groups. First group of statements is the one which a judge thinks is antithetical to the issue. Next most unfavorable statements come under the second group. This process continues with eleventh group containing the most favorable statements as decided by each judge. This method of grouping provides with a composite position for each statement. Major disagreement among the judges will lead to dropping of that statement which belongs to polar apposite groups. For the statements, which are retained, researcher has to assign a· score between 1 to 11 depending on its group number suggested by each judge. lf there are 3 judges then there will be 3 scores for each selected item. Median score will be- final score of that item. ln the final set of statements, those statements will get births, which are widely dispersed over the full range 1 to 11. Given this final choice, researcher will apply the same on the respondents to record the statement, which they individually agree with. Opinion of a respondent is assigned a score based on the median value of the statement, the respondent agrees upon.

Development of a differential scale takes ,[ot of time and money. At the end, subjective decision process on the part of judges gets intermingled with the scoring pattern. Thus, objectivity in the scaling technique gets severely curtailed. However differential scale is a reliable instrument for measuring a single attitude of the respondents.

Summated Scales of Likert type are constructed through item analysis approach. Respondents are reguested to express their agreement or disagreement, on a 5-point scale, with some preplanned set of statements. In a conventional Likert's scale these r~f 5 points are respecfively f Strongly disagree, Disagree, Neither agree nor disagree, (/l Agree and Strongly agree. Response categories often carry scores and scores are assigned to a respondent. The totalea score measures the attitude of the respondent. Alternatively, each statement of opinion may be studied through aggregation of individual opinions on that statement and study of the response pattern.

Initially, the researcher may consider a large number of statements, which express favorableness or un-favorableness in a definite manner. Then a trial test must be undertaken on a small group of respondents with a five-point scale of measure for favorableness or un-favorableness. The score 5 may be assigned to most favorable attitude and the score 1 may be assigned to most unfavorable attitude. Total scores are obtained to examine the discriminating power of the considered statements. Those statements which are having high discriminating power, should be selected for inclusion in the final measuring instrument. Other statements may be discarded.

The following is an example where five statements are used to assess the,perception about the use of a personal computer (PC). Each of these statements are to be examined on a five point scale as strongly agree, agree, neither agree nor disagree, disagree and strongly disagree.

Statements are the following :

1. A PC gives a better work environment.
2. A PC meets certain specialized needs of high volume computation.
3. A PC can increase efficiency through various functional supports.
4. A PC is a symbol of status.
5. A PC is more expensive compared to benefits it offers.

The Likert type scale is easy to construct and takes lesser time when compared with construction of Thrustone type scale. It is also more reliable as because a respondent has to answer against each statement. However, it has some limitations too. There is no reason to believe that the five positions on the scale of measurement are equispaced and hence the scale is. more of an ordinal scale than of an interval scale. Sometimes respondents may give answer according to what they think fit without having a direct exposure on the topic of interest. A person without having any work experience on a PC may give answer to all the five statements cited earlier.

Semantic differential scale is a type of factor scale, which is helpful in uncovering d/ the latent aspects of an attitude. Respondents are asked to indicate their choices among a set of bi-polar phrases for describing their feelings towards an issue/object. A statement set is framed along the different dimensions of the issue. The semantic scale presumes that the position in-between the two extreme poles describes the neutral position.

Consider for example, the following statement set that describes different dimensions along'which the brand image of a nasal drop can be studied. Against the statements bipolar responses are given for the responsive choices of the respondents.

Statement	**Set Bipolar responses.**	
Clearance of blocked nose	Effective	Ineffective.
After effect	Long lasting	Short lived
Color	Pleasant	Unpleasant.
Strength of flavor	Strong	Mild
Cooling Effect	Strong	Weak
Wrapper	Good	Bad.

To be effective the researcher first selects the concepts to be studied. The underlying proposition is an object that can have many dimensions of connotative meaning. These meanings are located in multidimensional property space, which is also known as semantic space. Next step is to select the scales. Usually,bipolar rating scales are of seven points.

For example, in case of clearance of blocked nose the one pole is effective for which a score +3 is. assigned and the other pole is ineffective for which a score - 3 is assigned. In between these two polar points +3 and -3 there are 5 other points in the diminishing degree of effectiveness. These are respectively +2, +1, 0, 1 and -2. It is easy to note that this scale is an interval scale and produces interval data. Thus, attitudes can be measured along different direction in terms of different intensity. The complete set of responses presents a comprehensive picture of meaning of an object along with the measure as suggested by the panel of judges.

5.6. Multidimensional scaling

The key objective of multidimensional scaling is to. provide with a low-dimensional picture. This reduction in dimension retaining nearly the original order of similarity or distance among the objects can help the researcher in having a visual inspection and arriving at a clear interpretation. When the multidimensional observations are numerical in nature and Euclidean distances in p-dimensions can be computed, the researcher can seek a q (< P), dimensional representation of observations. In a p-dimensional setup if $\underset{\sim}{x}_{px1}$ and $\underset{\sim}{y}_{px1}$ are two multivariate observations of two items respectively the Euclidean distance between these two items is given by

$$d^{(p)} = \left[\left(\underset{\sim}{x} - \underset{\sim}{y} \right)' \left(\underset{\sim}{x} - \underset{\sim}{y} \right) \right]^{\frac{1}{2}}$$

If the researcher has n items there will be m = n(n-1)/2 combinations of two items each. The distance between the i-th item and j-in item· can be represented by $d_{ij}^{(p)}$in the p-dimensional setup. In case vfe consider lower dimension q(<p) then the distance between i-th and j-th items can be represented by $d_{ij}^{(q)}$. The choice of q should be such that the following measure, E, will be. minimized :

$$E = \frac{\sum\sum_{i<j} \left[d_{ij}^{(p)} - d_{ij}^{(q)} \right]^2 / d_{ij}^{(p)}}{\sum\sum_{i<j} d_{ij}^{(p)}}$$

In this approach the Euclidean distances in the p-dimensional and a-dimensional setups are directly compared to ensure closeness amongst them. This closeness in-directly retains, as far as possible, the original ordering of the distances of pair-wise objects. If E value is zero the choice of q dimension is perfect, if E value is 2.5% it is excellent, if E value is 5% it is good, if E value is 10% it is fair. Rest is a poor fit.

It is possible to arrange n items in a low-dimensional coordinate system using only the rank orders of the m = n(n-1)/2 original similarities or distances and not their actual values. When the ordinal data is used to have a lower dimensional geometric representation the approach is called non-metric multidimensional scaling. When the original similarities or distances are used to,have a lower dimensional geometric representation, the approach is called metric multi-dimensional scaling.

5.7 Summary

Measurement is a process of mapping. It maps different aspects of the members of a domain on to some other aspects of a range set. Mostly the mapping is on a real line or an interval of the real line. Measurement scales are of four types viz., nominal scale, ordinal scale, interval scale and ratio scale. In nominal scale numerical values assigned to objects are numerical names. This is mainly used. for transforming categorical data. No numerical operation can be carried out on nominal data. However, one can count the frequency of occurrence of a particular code or nominal value. In ordinal scale one transforms the performance into numerical data by indicating the order of preference. Numerical operations such as addition, subtraction, multiplication and division cannot be carried out on ordinal data. But some statistical measure like rank correlation coefficient, median value and some statistical tests based on ranks can be carried out. Under interval scale, there is a constant unit of measurement but the zero point of measurement is arbitrary. All the commonly used statistical measures and tests can be used for analyzing interval data. Ratio scale is where unit of measurement is constant and the zero point is fixed. All the statistical measures can be computed and all the numerical operations can be undertaken .

For research in the field of social science, quite often the researcher has to develop own measurement tool. This is due to the fact that the frequent problem faced in social science is to measure the attitude and behavior of individuals. Stages of development of a measurement tool are concept development, dimension identification, selection of indicators of each dimension and formation of an overall index.

Scaling techniques are many in number. These scales can be grouped under two general heads, viz., rating scale and ranking scale. Under rating scale graphic rating scale and itemized rating scale are popular. Graphic rating scale presents a continuum on which various judgmental points are placed. This continuum may be of the form of a line or a box. The respondent puts a tick mark on the suitable position of the continuum as per his/her opinion. In itemized rating scale a sequence of statements are given in order of adherence to a property. The respondent selects the statement which is closest to his/her personal view. Rating scales are easy to use and enjoy wide I applicability. But they suffer from error of central tendency and error of hallo effect. Ranking scale is an alternative to rating scale and provides with ordinal data. In case pair-wise comparison is made, the ordinal data can be converted into interval data. In the rank order method direct ranks are given to objects and is easy to use. Data generated is ordinal data but can be converted into interval data also.

Popular methods for construction of measuring instruments are differential scale, summated scale of Likert type and· semantic differential scale. In case one wants to go for low-dimensional data for geometric representation one may use multidimensional scaling. Lower dimensional data are helpful in making interpretation and presentingpictorial outlay.

5.8. Questions

Long answer type questions.

1. What do you mean by measurement? In what way do the following scales differnominal scale, ordinal scale, interval scale and ratio scale?
2. Give an example with explanation for each of the following data : nominal data, ordinal data, interval data and ratio data.
3. Discuss the usefulness of rating scale with special reference to itemized rating scale.
4. ‘‘Development of a measurement tool has to pass through four stages.’’ What are those four stages? Explain each of these stages by referring to one research problem of social science.
5. Explain the usefulness of pair-wise comparison method for rankingn objects. How do you convert ordinal data into interval data under this method?
6. The following table presents pair-wise comparison data for 3 Cold drinks : CocaCola, Pepsi and Thums up.

	Preference for		
Brand	**Coca Cola**	**Pepsi**	**Thums up**
In comparison to			
Coca Cola	-	130	160
Pepsi	170	-	180
Thums up	140	120	-

(a) How many respondents were covered?

(b) How do the brands rank in overall performance?

(c) Obtain the score for these three brands on an interval scale.

7. Describe the method of scale construction for summated scale of Likert type.
8. What differences are there between differential scale and semantic differential scale? Which one do you prefer and why?
9. Explain the need for multidimensional scaling. Explain one metric multidimensional scaling technique.
10. Compare and contrast summated scale and differential scale.

Short answer type question :

1. Explain why c2 test for association of attributes can be applied but no numerical operation can be carried out on nominal data.
2. Can we measure the degree of association between two attributes measured on ordinal scale? Explain.
3. What is the basic difference between interval and ratio scales? Explain.
4. Explain graphic rating scale.
5. Describe the problems associated with rating scale.
6. For large number of items, would you prefer pair-wise comparison method or rank order method? Give reasons.
7. How do you construct a differntial scale?
8. Explain the construction of a summated scale.
9. Explain with an example semantic differential scale.
10. What is the objective of multidimensional scaling? Explain goodness of fit with respect to E-value.

Objective type Questions

Indicate whether the following statements are true or false.

1. Measurement is a process of mapping.

   True False

2. The range set is identical with the real line.

   True □ False □

3. No measure of central tendency can be calculated from ordinal data.

   True □ False □

4. Addition and subtraction are permitted for interval data.

   True □ False □

5. Division is allowed for ratio data.

True □ False □

6. Index is generally a weighted average of indicators.

True □ False □

7. In graphic rating scale only two judgmental points can be used.

True □ False □

8. Graphic rating scale is an ordinal scale.

True □ False □

9. Itemized rating scale is an interval scale.

True □ False □

10. High heaping near the central point is known as hallo effect.

True □ False □

11. Total number of comparison made for (n+1) objects is equal to n(n+1)/2.

True □ False □

12. Ordinal data of pair-wise comparison method can be converted into interval data.

True □ False □

13. Rank order method is a nominal method.

True □ False □

14. Construction of differential scale involves formation of 11 groups.

True □ False □

15. Statements,/ which have high discriminating powers, are included in the final set of a summated scale.

True □ False □

16. Multidimensional scaling reduces the dimension of observation always through metric method.

True □ False □

Unit 6 □ Value and Cost of Information

Structure

6.1 Introduction

Though traditional economist in their profit maximization models did not include cost of information yet cost of information is a very important factor that can have significant impact on the optimum decision rule. Specially, in today's business environment this cost is not only important but also becoming very high. The demand for information is on the rise; so also the cost of information. Even if we are willing to include the cost of information in a decision-making process there is a tendency to consider the cost of data collection as the most important cost item. There is no doubt that collection of data needs substantial resources but that requirement is comparatively less so for as the total research outlay is concerned. The cost of data collection is at the most onethird of the total research budget Project planning, and analysis, interpretation and report writing are the two other major heads each requiring nearly one third of the research outlay.

Budget formulation is one of the most important tasks the researcher has to under-take. Those researchers who apply for fund from different organizations should be able to identify the resources, determine the cost of planning, cost of data gathering and cost of analytical studies and reporting. There are three ways for formulating budget for a research work. The first one is the rule of thumb budgeting, second one is the functional area budgeting and the third one is the task budgeting. Under the rule-ofthumb budgeting one considers, as a convention, a fixed percentage of some other criterion and fixes the same as the budget outlay. For example, marketing research budget may be fixed at 10% of the sales revenue generated in the earlier year. Here, the other criterion is the last year's sales revenue; the fixed percentage is 10%. Under the functional area budgeting, a fixed portion of the total expenditure is considered as the budget outlay of the research activities. This is a very common practice followed by both government_ and nongovernmental organizations, and profit making and nonprofit making organization. This being a fixed portion of the total expenditure, lot of autonomy can be given to functional units for clearing their respective research projects. Under the task budgeting each project is examined on a case-by-case method and techniques such as cost-benefit analysis are employed to ascertain the suitability of the scheme. For schemes, which clear the technical analysis, one finds out the budget requirements and adds them to arrive at the total research outlay.

6.2 Value of information from research

The usefulness of information collected from a research study determines the benefit of the research. A research study may increase revenue, eliminate the risk of losses, strengthen the promotional campaigns, help in fixing the right price for the company's offer and so on. There are cases where a research study can be simply wastage of funds and time. It is therefore extremely important not to look at the cost of a research project only but to simultaneously examine the benefits that will inflow into the system that sponsors the research project. Only when the net benefit is more than the net cost, a project can be cleared for execution. This difference between the benefit and cost is the value of information from research.

It is not very difficult to ascertain the value of applied studies. Applied studies are objective oriented where the objective is specific in terms of revenue or profit or cost or any such measure of the organizational efficiency. In that case one may judge the benefit of a research study in terms of added revenues, or additional profit, or reduction in cost or increase in efficiency as the case may be. Sometimes value of research information can be expressed as the difference between the results of the decisions taken with the research information and the results of the decisions taken without the research information. This later approach is appealing in the context of decision making under risk.

Let us examine, in detail some of the evaluation methods, which are of use and can be of much help in determining the benefits of a research study. These commonly used methods are Prior evaluation, Posterior. evaluation and Option analysis. The decision theory approach with special reference to Bayesian analysis will be separately covered in the next section.

Prior evaluation is mostly carried out based on the information need of an organization and the probable contribution of the proposed research studies. We say probable because neither the costs nor the benefits are predictable or estimable with definiteness in many of the research studies. For example, in case of a proposal for under-taking detailed management audit of the basic operations in a company where the management is in a fix over the problems it is facing in the business environment. The need for taking up such a management audit is beyond question. But the risk involved in terms of costs and benefits are not easily estimable. For cases where prior need is understood but value of information is not clear the research study may be subdivided into stages so as to minimize the overall risk and facilitate the evaluation part. Prior evaluation of the benefits and cost may be difficult for the entire study. But stage-wise evaluation becomes easier as one has to evaluate for only a sub-part and the posterior evaluations will be available for the earlier stages. At any sfge ot the' work, the sponsor may withhold the authorization and discontinue the rest of the work if sufficient benefit does not accrue to the sponsor.

Posterior evaluation measures the value of research only after the research is over. One may consider an objective estimate of the contribution of the research towards aimed at benefit. For example, this benefit, also referred as net benefit, may be corporate profitability when one conducts research in the field marketing. Similarly, this benefit may be expressed in terms of reduction in cost when one conducts research in the field of production. Posterior evaluation is best possible when information is quantifiable in terms of measures of efficiency of the organization under study. It is also desirable that observations of the type before and. after are available to estimate the change in the state of the system organization.

Option analysis examines the alternative courses of actions available with the decision makers. These alternative courses must be well defined and estimated costs and benefits associated with each option should be available. While costs are easy to approximate, benefits are difficult to estimate. Mostly these estimate are crude. One may refer to decision theoretic approach as an example of option analysis.

6.3 Summary

Cost of information is very high in today's business environment. There· is a tendency to consider the cost of data collection as the most important cost item. But, in reality, cost of data collection is only one-third of the total research budget. Project planning and analysis, interpretation and report writing are the two other important cost items.

Since information is costly, budget formulation assumes priority. There are three ways by which budget can be formulated. These are known as rule-of-thumb budgeting, functional area budgeting and task budgeting. Under rule-of-thumb budgeting one considers a fixed percentage of some other criterion and fixes the same as the budget outlay. Under the functional area budgeting a fixed portion of the total. expenditure is considered as the budget outlay of the research activities. Under task budgeting each project undergoes cost-benefit analysis. Those projects which are cleared by the technical analysis, are studied for fund requirement. Total fund requirement of all the cleared projects is the final budget outlay of research.

Benefits of the research are determined by the usefulness at the information generated by research activity. A research study may increase revenue, eliminate the risk of losses, strengthen the promotional activities and so on. It is extremely important to look at both costs and benefits of a project. To determine the· benefits one may undertake prior evaluation, or posterior evaluation or option analysis. Prior evaluation is mostly carried out based on information need of an organization and the likely contribution of research. If the value of information to be generated by the research is not clear at the beginning the research work may be subdivided into stages. This minimizes the overall risk as stage-wise evaluation is easier and the work may be continued or discontinued depending on the net benefit of the earlier outcomes. Posterior evaluation measures the value of research only after the research is over. It is best possible where information is quantifiable in. terms of a measure of efficiency. Option analysis examines alternative courses of action in terms of expected pay off or some related measures. Generally, option analysis is supported by decision theoretic approach. Under decision theory under risk we may evaluate the benefit in many ways. Two_most prominent approaches are perfect information along with its costs and ben efits and sample information along with its costs and benefits. The expected value Of perfect information is the difference between the expected pay off under perfect information and the expected pay off under prior analysis. The expected value of sample information is the difference between the expected pay off under Bayesian posterior analysis and the expected pay off under prior analysis.

6.4 Questions

Long answer type questions

1. 'Increased information need has increased the cost of information.' Critically examine the above statement with special reference to business activities.
2. Explain different methods of budgeting.
3. How do you propose to determine the value of information?
4. Explain the concept of perfect Information and the expected value of perfect information.
5. Using the Bayesian approach, determine the expected value of sample information.

Short answer type questions :

1. Explain three major cost components of research.
2. 'Functional area budgeting gives more autonomy.' Explain.
3. 'Prior evaluation is difficult but necessary'. Why?
4. Which information do you prefer perfect information or simple information? Why?
5. Indicate the major use of Bayer' 'theorem in option analysis.
6. Can the value of sample information be more than the value of perfect information? Give reasons.

Objective type questions.

Indicate whether the following statements are true or false.

1. Cost of data collection accounts for 50% of the total cost of research.

   True □ False □

2. Cost of project planning is an important cost item.

   True □ False □

3. Rule-of-thumb budgeting is based on bargaining power mainly.

   True □ False □

4. Posterior evaluation is best possible when information is quantifiable.

   True □ False □

5. Under prior analysis best course of action is the one that maximizes the expected pay off

   True □ False □

6. Perfect information cannot change the expected pay off in the positive direction

   True □ False □

7. Indicators or reports are made of perfect information.

   True □ False □

8. Posterior analysis is based on Bayesian approach.

   True □ False □

9. The Expected pay off under sample Information cannot be more than the Expected play off under perfect Information.

   True □ False □

10. Posterior analysis is expected to give better pay off than prior analysis.

    True □ False □

Unit 7 ◘ Sampling Design

Structure

7.1. Introduction

The researcher may opt for sample survey when complete enumeration is beyond the scope of the study. If the number of units to be studied is small, the method of enumeration is not destructive and the requisite skill for enumeration is available to cover all the units one may resort to complete enumeration in place of sample survey. However, mostly the number of units to be studied is not small and the cost of research is to be controlled. In those cases one may go for sample survey. It has a greater scope in the sense that requirement of trained personnel, specialized instruments, and requirement of time will be much less in sample survey than in complete enumeration. There may be greater coverage of information and more in-depth study. A sample survey, is also found to be more error-free than complete enumeration because there is a better control of non-sampling error.

A Sampling design is a plan for selecting units from the set of all units, commonly known as universe or population. Sampling design should, therefore, define the universe in clear terms. Sampling design may also indicate the size of the sample. Thus, a sampling design addresses two vital questions: The first question is: How the units are to be drawn from the population? The second question is: How many units are to be drawn? Depending on the nature of the research problem and precision requirement the researcher may decide about how many units are to be drawn and the procedure for drawing those units.

The ultimate units to be studied during the survey work are known as sampling units. A sampling unit for a household survey may be a family. A sampling unit for an opinion survey may be an individual. The list of all the sampling units is known as the sampling frame. It contains the identification of sampling units so that a selected unit from the sampling frame can be uniquely identified. If the sampling frame is not available but can be prepared, the researcher may develop it for his research purpose and also for the purpose of future use. If the sampling frame is not reliable the sample cannot be a representative one. In cases where the sampling frame is difficult to obtain one may divide the selection procedure into multiple stages. In the first stage, sampling frame for only the first stage-sampling units is needed. For the second stage, the sampling frame of the second stage units for each of the selected first stage units is needed. For example, for selecting households on an all India basis the sampling frame of all the households may not be available. But ideally one may select a few states, from the selected states a few districts may be selected. In this way the construction of sampling frame of households gets restricted to only the selected districts. This approach of multistage sampling works well with large-scale sample survey where the sampling frame of ultimate sampling units is not always readily available. Problem of lack of sampling frame arises in cases of moving population like fishes in a pond, birds of migrating type, tigers in a forest and so on. Generally, catch and recatch method is applied for these cases to avoid the problem of constructing complete sampling frame.

The objective of the study, the characteristics of interest and the related measures of the population are all important in shaping the scheme. of sampling procedure. For example, if information is required for different age groups it is desirable to prefix the sampling design in such a way that all the age groups are well represented in the sample. Similarly if sampling units are of unequal importance, like companies of different levels of transaction, one has to assign unequal chances for selection of those units from the corresponding overall population.

Keeping in mind the above concepts and considerations, we first address the problem of selection of sampling units and then the problem of determination of sample size and the estimation of population measures. It is easy to note that if we know how estimate the population mean we know how to estimate the population proportion and the population total. In view of this we shall consider the case of estimating the population mean mainly.

7.2. Sampling Design

Under sampling design, the researcher must propose the concrete plan of drawing the sampling units from the sampling frame. The basis for selection may be either probability value or non-proobability value. In non-probability sampling there may be restriction on the inclusion, of units in the sample or there may not be any restriction at all. For unrestricted non-probability sampling, sampling units may be selected in a haphazard manner without any systematic procedure. There is a tendency of projecting haphazard sampling as a random sampling procedure. But, in reality, these two procedures are completely different. Random sampling, to be described later, is a systematic and probabilistic procedure. Haphazard sampling, on the other hand, is an unsystematic and non-probabilistic procedure. In case of random sampling, one can draw statistically valid conclusion about the population. In case of haphazard sam-pling, no statistically valid conclusion can be drawn. Selected sampling units can at best be described through different measures. One cannot make use of those mea-sures for inductive inference.

For restricted non-probability sampling one may adopt purposive or judgment sampling in the most general sense, purposive sampling means selection of sampling units based on a prefixed purpose. It is expected that the purposive method may result in a representative sample. Unfortunately, this expectation can hardly be sci-entifically established. The extent of bias cannot be estimated and the extent of error remains unclear and unknown. Moreover, for a multipurpose survey, selection of units based on one purpose may not be suitable for throwing light on other aspects of the survey. For example, if one purpose fully selects average performers from a group of sales representatives one may closely estimate the average performance but the population variation in respect of performance can hardly be estimated. A significant underestimation of variation will be reported under such a choice.

For probability sampling there may be both restricted type and unrestricted type sampling procedures. Example of an unrestricted probability sampling is simple ran-dom sampling. Example of a restricted probability sampling is stratified random sam-pling. In this context it may be made clear that probability sampling and random sampling are not equivalent. In probability sampling, probabilities are assigned to sampling units for selection. Random sampling is a special kind of probability sam-pling where assigned probability values are equal. Hence, a probability sampling is not necessarily a random sampling procedure.

Simple random sampling is of fundamental importance in the field of Statistics. The concept of sampling distribution is having the basic assumption of random sampling. By random sampling we mean selection of the sampling units through assignment of equal chance of selection to each unit at every stage of selection. Thus, each unit of the population has the same chance of inclusion in the sample. Simple random sampling is of two types: simple random sampling with replacement (SRSWR) and simple random sampling without replacement (SRSWOR). Let us explain these concepts in detail. In SRSWR, at every stage of selection there are as many units in the sampling frame as they are in the original population. This means, if the universe or the population is made of N units, U_1, U_2, ...U_N then at every stage of selection there will be (U_1, U_2, ..., U_N) units to be sampled and irrespective of the earlier selection each unit will be assigned a probability value 1/N for selection in the current stage. This process of selection is repeated n times where n is the sample size.

For simple random sampling without replacement (SRSWOR), in the selection of first unit there are N units in the sampling frame with each sampling unit having a probability of selection as 1/N. In the second round there will be (N- 1) units in the sampling frame with each unit having a chance of selection as 1/(N- 1). The unit selected in the first round will not feature in the sampling frame of the second round. This process continues till n sampling units are selected. Typically, in the k-th round, k varying from 1 to n, there will be (**N**-k+ 1) units in the sampling frame. Each of these units will be assigned a probability of selection as 1/(N- k +1). All the units selected up to the (k- 1)-th round are to be deleted from the original list of sampling units and the new sampling frame is to be constructed for selection of the k-th unit. Thus, for SRSWOR the chance of selection of a sample is

$[{}^{N}C_{n}]$=n!/[N(N-1)....(N-n+1)].

As there are ${}^{N}C_{n}$, possible choices of n units from a population of N units and all the sample possibilities have equal chances of selection.

In this sampling procedure the precision level can be increased by increasing n. It is inexpensive and simple and provides valid estimator. But a complete frame is needed to make it operational.

Let us consider, for the purpose of demonstration, a small population of size N = 6. Let the population be denoted by U = (U_1,U_2,U_3, U_4,U_5,U_6). Let us draw a sample of size n= 3 from this population both with replacement and without replacement. For making the choice of a unit we take help of random number table. A random number l\ table is a row and column-wise arrangement of digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9/such that each position in that rectangular arrangement is filled up by selecting one of these ten digits with equal chance of selection and independently of selection for other positions. The main benefit arises out of the fact that randomization of the digits can be done once for all and hence the population units are not to be randomly arranged. Once the population is given with each unit having an ordinal number (here U_5 is the 5th unit, say) we have to make a correspondence between the population and the random number series. Let us consider the following part of random number table constructed by L.H.C. Tippett, No. xv pp 12-13.

4652 3819 8431 2150 2352 2472 0043.

Since our population is of size 6 let us decide to accept one digited random number and Jet us make a correspondence between the sampling units and random numbers as :

Unit	Random number	Unit	Random number
U_1	1	U_2	2
U_3	3	U_4	4
U_5	5	U_6	6

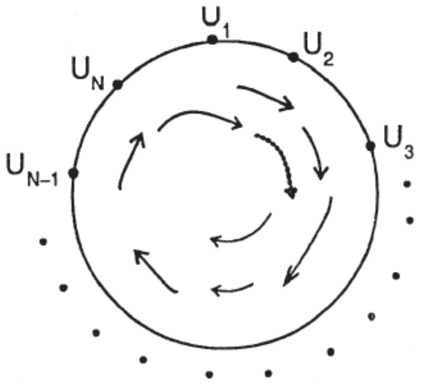
and the rest of the random numbers, i.e., 0, 7, 8 and 9 be rejected for, the purpose of selection for the present problem. Suppose we decide to start from the position of the quoted random number series. Our starting digit is 3. Once the starting position is selected it will remain fixed and subsequent numbers are to be selected either row- wise or column-wise without any gap. Our first selected random number is 3. Hence the first unit to be selected is U_3. The next number-is 8 but 8 has no correspondence with the population units and hence we reject the number 8. The next number is 1. It corresponds to U_1. Hence the next unit selected is U_1. The next number is 9 but 9 has no correspondence with the population units. So we reject the number 9. The next number is 8 which has also no correspondence with the population units. Hence we reject the number 8 too. The next number is 4/ which corresponds to U. Thus, the third unit selected is U_1. This completes the selection of all the 3 units from the given population. The selected sample of size 3 is given by s = (U_3, U_1, U_4). This selected sample meets requirements of both SRSWOR and SRSWR as there is no repetition of units. In SRSWOR we have tc reject the selection of a unit ff that unit has already been selected in the sample and go for a fresh selection.

Let us consider a population of size 987. This means N = 987. The population units be represented by U_1,U_2,U_3,...,U_{986}, U_{987}. Thus, the population is U=(U_1,U_2, ..., U_{987}). To draw a sample of size 4 from this population we make use of the same random number series. This time we have to consider 3 digited random numbers because N is a 3 digited number. Let the starting point be the first position of the quoted random number series. The first random number is 465 and we select the unit U_{465}. The next random number is 198 and we select the unit U_{238}. The next random number is 198 and we select the unit U_{198}. The last random number is 431 and the selected unit is U_{431}. Thus, the selected sample by both SRSWR and SRSWOR (as there is no repetition of units in the selected sample) is

s = (U_{465}, U_{238}, U_{198}, U_{431}).

Let us consider another example where the population size is N = 110. In such a case we have to consider 3 digited random numbers. But many of the 3 digited random numbers will be more than 110 and hence there will be many rejections if we do not accept random numbers above 110. Here we adopt a different procedure. Since the highest integer multiple of 110 which is also a 3 digited number is 990 (= 9×110) we consider random numbers acceptable for selection to vary between 001 to 990. We reject random numbers 000 and 991 to 999. Thus only 10 random numbers we are rejecting. Rest will have correspondence with population unit. Cor-respondence will be made through remainder method. Given any 3 digited random number we shall divide it by 110 and note down the remainder. If the remainder is zero we consider U,,, For the rest of the cases, i.e., from 1 to 109 we consider units from U, to U,. Once this correspondence between 3 digited random numbers and population units is established let us draw a sample of size 3 by SRSWOR method. Let the starting point of the random number series be the 9th position. Then the first 3 digited random number is 843. This is less than 990 and more than 000. We accept this number. Dividing 843 by 110 we get the remainder as 73. So the selected sampling unit is U. The next number is 121. The corresponding unit selected is U_{11} because the remainder is 11. The last number is 502 with remainder as 62. The corresponding unit is U_{62}, Thus, the sample selected is s = (U_{73}, U_{11}, U_{62},)-

Systematic Sampling is a special type of sampling design where the process of selection of sampling units is systematic and simple but not fully random. In the ideal case, the population size N is an integer multiple of sample size n. Thus, in the ideal case N = kn where k is an integer. Then, one may divide the population into k blocks each with n units. The first block is made of units U_1, $U_{k+1,}U_{k+1}$, $U_{2k+1, ...,}$ $U_{(n-1)k+1,}$ the second block is made of units U_2, U_{k+2}, U_{2k+2}, ..., $U_{(n-1)k+2}$ and so on, the last block being made of U_k, U_{2k}, ..., U_{nk}. In place of selecting each unit randomly, in systematic sampling one selects one of these k blocks by the random procedure. If r is the selected random number between 1 and k, then the selected sample is U_1, U_{k+1}, ..., $U_{(n-1)k+r}$ Thus, units are selected periodically with period as k. Once the first unit selected is U_r the subsequent selected units are periodically drawn with period k, i.e., (k + r)-th unit, (2k + r)-th unit and so on. In case the ideal situation does not hold true i.e., N is not equal to kn wed find the value of k as N = kn +k' where 1 < k' < n. In such a case the selection of unit follows a periodic pattern with the period as k and the starting point r, randomly selected between 1 to N. Units are arranged along a circular path so that starting from U, we move up to UN and after UN there comes the first unit U,, then U, and so on. Systematic sampling is a very simple. and convenient method. It takes less time to select the complete sample. Results obtained from the systematic sampling are mostly satisfactory. However, problem arises in case there is a periodic arrangement of units in the population. For example, if every 15th house is a rented one there is chance that all the selected houses will be of rented type.



To explain this method of systematic sampling further, let us consider a population of size 110 from which a sample of size 10 is to be selected. Here N = 110 and n = 10. Hence k = N/n = 110/10 = 11. We have to first select a random number between 1to11. We need to select a 2 digited random number and let us follow the remainder method. Thus, admissible random numbers will range from 01 to 99 and rest of the numbers i.e., 00 will be inadmissible. If we start from the 1st position of the quoted random number series the two digited random number comes out to be 46. Using remainder method we get the r value as 02. Thus, the first unit selected is U_2. Rest of the units will be selected with a periodicity of k = 11. Hence the selected sample is s=U_2,U_{13},U_{24}, U_{35}, U_{46}, U_{57}, U_{68}, U_{79}, U_{90}, U_{110}).

Stratified random sampling is a restricted version of the simple random sampling. In simple random sampling, sampling units are drawn from the entire universe. But, in a stratified random sampling, the universe is divided into different strata and from each stratum sampling units ar drawn based on simple random sampling. This restricted approach aims at meeting dual objectives. The first objective is to provide with information for each stratum separately and for the entire population jointly. The second objective is to reduce. the extent of error and this can be achieved by constructing strata in such a way that within stratum variation is minimum and between strata variation is maximum. In other words, this means that the units are as homogeneous as possible within a stratum and are as heterogeneous as possible between the strata. The strata must be exclusive and exhaustive representation of the population with the criterion of stratification, i.e., the stratifying factor dictated by the objective of the study. Stratifying factor may be sex if we want to present information separately for males and females and make a comparative study. Stratification may also be based on age, educational level, and geographic region if the objective of the study so demands. Within each stratum, as we propose to drawn samples based on simple random sampling, the sampling procedure will be exactly same as described under. simple random sampling, with or without replacement. However, one major issue needs to be addressed. This is about the sample size, n_1, to be drawn from the i-th stratum. Let N_1. be the size of the i-th stratum and let there be, in total, k strata. Then N =N_1, +N_2,+...+ N_k, and the total sample size n=n_1, +n_2,+...+ n_k, Under Bowley's **proportional allocation** n_1 is proportional to N_1. Thus, n_1, = C N_1 for some constant C independent of i, i=1,2,...,k. This independent of i, i=1,2,...., k. This means, $\sum_{i=1}^{k} n_i = C\sum_{i=1}^{k} n_i$ or n = CN or C=n/N. Hence n_i=n N_i / N, i=1,2, ...k.

Proportional allocation is a simple and convenient method which can be implemented without any additional information. Neyman's formula for optimum allocation takes into consideration both stratum size and stratum variability. If X.. is the value of the character under study for the j-th unit of the ith stratum, j = 1, 2, ... , N_i and i = 1,2, ... k, then the population mean, for the i-th strata, X_i and population variance for the i-th strata, S_i^2, are given by

$$X_i=(X_{i1} +...+X_i N_i) / N_i = \frac{1}{N_1}\sum_{j=1}^{N_i} X_{ij}$$

$$S_i^2= [(X_{i1}-M_i)^2+...+X_i N_i - M_i)^2]/ (N_i-1).$$

Then, the **Neyman's optimum allocation** is given by

$\mathbf{N_i=nN_iS_i/[N_iS_i +N_2S_2+...+N_kS_k]}$

For example,- for a population with size **N** = 1,000 with 4 strata, the respective stratum size **N.** and standard deviation S. values are given as :

$N_i = 400, S_i = 15; N_2= 300, S_2 = 20; N_3 = 200, S_3 = 10 ; N_4 = 100, S_4= 5.$

If it is decided to draw samples with the overall sample size n = 100 then under the proportional allocation sample sizes will be

$n_i = nN_i/N = 100(400)/1000 = 40,$

$n_2= nN_2/N= 100 (300)/1000 = 30,$

$n_3= nN_3/N = 100 (200)/1000 = 20$ and

$n_4= nN_4/N = 100 (100)/1000 = 10$

Under the Neyman's optimum allocation we have

$n_i = nN_iS_i/N_iS_i + N_2S_2 + N_3S_3 + N_4S_4)$

$= 100 (400) (15)/[(400)15 + (300)20 + (200)10 + (500)5]$

$= 100 (400)(15)/[6000 + 6000 + 2000 + 500]$

$= 100 (6000)/14500$

= 41 (nearest integer)

$n_2=n N_2S_2/N_iS_i + N_2S_2 + N_3S_3 + N_4S_4)$

= 100 (300) (20) *I* 14500

=41 (nearest integer)

$n_3=nN_3S_3 /(N_iS_i + N_2S_2 + N_3S_3 + N_4S_4)$

$= 100 (200) (10)/14500$

=14(nearest integer)

$n_4=nN_4A_4 /(N_1S_1 + N_2S_2 + N_3S_3 + N_4S_4)$

$= 100 (100) (5)/14500$

=4 (adjusted for a total sample size as 100)

It may be noted that stratified random sampling is more representative than simple random sampling. It provides with greater precision because of the stratification of the population into homogeneous strata with between strata variation as high as possible and within stratum variation as low as possible. It has administrative convenience due to division of the population into different strata. There may be reduction in respect of both time and cost of data collection. However, the division of population into homogeneous strata requires prior knowledge on the population. Similarly, optimum allocation assumes knowledge of population standard deviation. But this information may not be available with the researcher.

In **cluster sampling** the population is divided into some recognizable subgroups which are smaller in size. These subgroups are called clusters. Thereafter a sample of clusters is selected by simple random sampling from the population of clusters. For the selected clusters all the sampling units belonging to these clusters are to be studied for the survey work. For example, if we have to conduct an opinion poll in the city of Kolkata we may divide Kolkata into different localities according to post office. Each locality will .be treated as a cluster. If there are 117 such clusters. We may select 12 clusters from these 117 clusters and cover all the households for those selected clusters. This method of cluster sampling is popular when data is to be collected for some common characteristics of the human population. If the clusters are large in number and small in size one may expect a reasonable level of precision. It is also desirable that the clusters be prefer-ably of nearly equal size.

Multistage sampling is carried out in stages where the units to be sampled are regarded as made of a number of first stage units, each first stage unit is made of a number of second stage units and each second stage unit is made of a number of third stage units and so on. Thus, ultimate sampling units are reached via multiple stages through higher stage units. One may adopt simple random sampling or some other method for selecting units in each stage of sampling. We have already indicated that for implementing multistage sampling one does not· require the complete sam-pling frame. Sampling frame of first stage units is needed for sampling in the first stage. Sampling frames of second stage units of the selected first stage units are needed for sampling in the second stage and so on. These ensure administrative ease and hence multistage sampling is a popular design for conducting large-scale sample survey. Though this flexibility in the sampling procedure may reduce the cost and time for conducting the. survey work, it may reduce the level of precision also. In fact, it is generally less efficient than some suitably designed single stage sampling procedure. Variability of the estimates under multistage sampling depends on the composition of the first stage units, second stage units and so on.

Convenience sampling is one in which a sample is selected based on the con-venience of locating and contacting the sampled units for the purpose of,collection of information. Neither the probabilistic method nor the judgmental method are taken into consideration. Results obtained from convenience sampling are not satisfactory as they hardly depict the behavior of the complete population.

Quota sampling is a type of stratified sampling where the population is divided into multiple strata. The investigator is informed about the number of units to be covered by him from a particular stratum assigned to him. How these units will be selected will be left entirely to the discretion of the investigator. The investigator may or may not resort to simple random sampling method to select the units. He/she may go by his/her judgment or convenience to get the result as early as possible. For example, in a survey on job opportunity an investigator may be asked to cover 200 individuals belonging to general category, 50 individuals belonging to scheduled caste category, 29 individuals belonging to Scheduled Tribe category and 50 individuals belonging to OBC category. The investigator enjoys complete flexibility in selecting such individuals while the number of individuals to be surveyed from each category remains fixed. Thus, quota sampling is a· mixture of stratified and convenience sampling. In case the investigators are highly skilled and if they are closely supervised this method may give desirable results in less time and cost. However, it suffers from all the limitations of nonprobabilify sampling. The extent of error in estimation cannot be worked out. As a result, the error term cannot be controlled.

7.3. Choice of sampling technique

As it is difficult to suggest a sampling technique which is uniformly better than the rest of the techniques, choice of a sampling technique depends on the objective of the study and the field/ research situation. In different situations different techniques would be useful. In general, the size of the population, availability of sampling frame, availability of resources and time, precision requirement in terms of sampling error and administrative ease, all play decisive roles, collectively or individually, in making a choice of the sampling method.

It is also hot clear Whether a probability sampling provides a better solution than a non-probability sampling. For example, if it is a choice between simple random sampling and purposive sampling one cannot decide in favor of one over the other. A rearcher should weigh the relative strengths of thesg two methods and arrive at ile best choice. If the sample size is small compared to the population size, the purposive sampling technique may be giving a better result than the simple random sampling. The simple random sampling, on the other hand, may gain edge over purposive sampling as the ratio of sample size to population size increases.

Quite often stratified random sampling provides with a higher level of precision than the simple random sampling. This happens especially when the population can be suitably subdivided. into homogeneous strata. Sometimes, multistage sampling may give better result than the single stage sampling methods.

A good sampling design is the one which ensures a proper representative sample and describes the aimed at behavior of the population in a better way. The design that can reduce the sampling error in a significant way is a good design and should be kept in mind while making a final choice of the sampling technique. The systematic bias that arises out of errors in the sampling procedure and cannot be reduced by a mere increase in the sample size, rs to be kept under control.

7.4. Summary

Since cost of research, time of study and error are less in case of sample survey **c//·** than complete enumeration one may adopt sample survey method for collection of data. Scope is also more for sample survey. To conduct a sample survey one must plan for selecting units from the set of all units. The set of all units is known as universe or population. The plan of selecting units is known as sampling plan. UItimate units to be selected are sampling units and the list of sampling units, from which samples are drawn, is known as sampling frame.

Selection of sampling units may be based on probability sampling and non-brobability sampling. Each one of them may be either restricted type or unre-stricted type. For unrestricted probability sampling, sampling units may be selected in a haphazard way. This is unsystematic, and non-probabilistic and hence no valid con-clusion can be drawn. For restricted non1probability sampling one may adopt purposive or judgment sampling. Purposive means based on a prefixed objective.

Example of an unrestricted probability sampling is simple random sampling. It is of fundamental importance in the field of statistics. Simple random sampling without replacement means selection of units from· the population with equal probability with-out returning the selected units to the population. In case selected units are put back to the population for fresh selection, we term it as simple random sampling with replacement. Random number tables are of help in selecting sampling units.

Systematic sampling is a quasi-random procedure where there are k sets of sampling units, each, set containing equal number of units, equal to the sample size n. Then a set is selected at random. Stratified random sampling is an example of restricted probability sampling where the entire population is divided into homoge-neous strata with within stratum variation as minimum as possible and between strata variation as maximum as possible. From each stratum samples are drawn based on simple random sampling principle. Number of units to be drawn for different strata can be determined by the principle of proportional allocation or optimum allocation.

Other important variants of sampling procedure are cluster sampling, multistage sampling and quota sampling. In quota sampling procedure is similar to stratified sampling except the fact that the investigator based on his/her judgment selects units from different strata. In cluster sampling population is divided into recognizable groups of small sizes. Then, randomly selected groups constitute the total sample. In multi-stage sampling, sampling units are selected in stagest: first stage, second stage and so on, the last stage giving the ultimate units to be selected. Thus, first stage units are made of second stage units, second stage units are made of third units and so on. The last stage units are the ultimate units.

Since so many techniques of sampling are available, one has to make a choice of sampling technique. This choice should be based on size of the population, availability of the sampling frame, availability of resources and time, precision requirement and administrative ease.

7.5. Questions

Long answer type questions.

1. What do you mean by a sampling plan? Indicate why one may opt for sample survey for collection of data.
2. Explain the concept of non-probability sampling, both restricted type and unre stricted type.
4. Describe two variants of simple random sampling explaining, in details, the sampling procedure.
4. From a population of 1500 units draw 5 units using simple random sampling without replacement method, by explaining the sampling procedure in details. Make use of the following random numbers :

   4652 3819 8431 2150
   2352 2472 0043 3488
5. Is systematic sampling a fully random sampling? If yes, give justification by explaining the sampling procedure. If no, give reasons for your comments by making reference to the procedure of selection of sampling units under this scheme.
6. What are objectives of stratified random sampling? Explain how those objectives are met in stratified random sampling.
7. Describe Neyman's optimum allocation. If population variances for all the strata are equal what will be the form of optimum allocation?
8. Indicate the advantages of multistage sampling.
9. Is there any sampling design which is uniformly better than all other designs?
   How do you make the choice of a sampling technique?

Short answer type questions :

1. Explain the terms sampling unit, universe and sampling frame by considering a problem of your choice:
2. How is a random number table constructed?
3. Explain the remainder method of drawing samples. When are random number tables used?

4. Is quota sampling a stratified random sampling? Give reasons.
5. Cluster sampling is a type of simple random sampling where sampling units are clusters of ultimate units. Do you agree? Justify your answer.
6. Give an example where judgment/purposive sampling can be useful.
7. What is the major disadvantage ot noh5rob ability sampling?

Objective type questions :

Indicate whether the following statements are true or false.

1. Complete enumeration method results in higher non-sampling error.

   True □ False □

2. Sampling frame is the list of selected units in the sample.

   True □ False □

3. Haphazard sampling is an unrestricted probability sampling

   True □ False □

4. Probability sampling is always unrestricted type.

   True □ False □

5. Effective sample size (number of no~dentical sampling units in a sample) in SRSWOR is always greater than or equal to the effective sample size in SRSWR.

   True □ False □

6. A random number table is an arrangement of digits O to 9 in a haphazard manner.

   True □ False □

7. Stratified random sampling is an unrestricted probability sampling because units are selected randomly from each stratum.

   True □ False □

8. Systematic sampling gives better result if the population units are. arranged in a periodic fashion in respect of the character under study.

True □ False □

9. Proportional allocation is applicable only when population proportion is to be studied.

True □ False □

10. Convenience sampling is a special case ·of clus.ter sampling.

True □ False □

Unit - 8 Tabulation and Analysis of Data

Structure

8.1. Introduction

When data starts pouring in, tabulation and analysis of data assume importance.

There are some preprocessing needs for undertaking first,tabulation of data and then analysis of data. To make the tabulation work easier the collected data must pass through the stages of **data preparation, data examination** and **data mining.** By data preparation we mean editing of data, coding of information for ease of storage and dissemination/ and data entry. The first objective of data preparation is to have a look *al/* at the raw data to detect errors, omissions and repetitions and to correct these errors, omissions and repetitions to the extent possible. In this process one can ensure a minimum quality level in the collected raw data and maintain the precision standard aimed at during the stage of analysis. During the stage of analysis it becomes ex-tremely difficult, time consul}g @ad fund consuming to rectify the errors, omissions and repetition. During edition, the editor checks for accuracy, consistency, uniformity and completeness of the raw data'so that raw data can be arranged to simplify the coding and tabulation works. The other objective of data preparation is to codify the re-sponses of sampling units or experimental units or respondents so that raw data can be grouped into a limited number of classes for ease of interpretation and conclusive action. Data entry into a computing system follows such a detailed codification work. It converts gathered information to a suitable media so that viewing and manipulation can become easier. Data entry through keyboard is a manual method where a key board operator keys in the dressed raw data and save the same under a file name. This data file can then be used by the researcher for his/her own purpose, or by the future researchers. Sometimes, manual methods of data entry can be replaced by mechanical methods of data entry. Optical scanning instruments can sense answers given by darkening one of the alternative small circles, each circle representing one answer. This scanning method reduces the time for data preparation and the extent of error in this processing. Voice recognition system is also capable of recording interviews and can be an alternative to telephone interview.

The objective of data examination is to explore, examine and display the data on hand to have greater insight into the problem to guide the analysis if needed. There may be revision in the data analysis approach earlier fixed during the stage of re-search design. Thus, data examination emphasis 'hat research should be problem based and facts-based, and not merely technique -based.

The objective of data mining is to discover knowledge from database. The extrac-tion of knowledge depends on identification of unique and useful pattern of information in the data. In this sense, data mining is comparable with traditional mining activities where miners search for ore from the underground. Like mining where one has to sift a large amount of material to locate an extractable patch of ore, in data mining one has to study bulk of data to find out a novel pattern of interest. Since we are talking about a pattern we are having in mind a multidimensional study against the usual marginal analysis of statistical variates.

8.2. More about data preparation, examination and mining

8.2.1. Data preparation

The first step in data preparation is editing of data. The purpose of editing, as pointed out earlier, is to ensure that data are accurate, consistent with related infor mation, uniformly entered in the documents/ and complete in all aspects of the study .

There are two types of editing of data. One is field editing and the other is central editing. Field editing is to be carried out soon after the collection of raw data from the field. If gaps are found, a call back should,necessarily,be made to bridge the gaps so observed. Guesswork is undesirabie as the investigator's bias may creep in, add-ing to the error in the subsequent analysis and conclusion. If inconsistencies are found, the same should be checked from the repeat visit in place of arbitrarily accept-ing one of the inconsistent reports. Field editing also involves a routine checking of collected data by the supervisor through actual verification of a prefixed percentage of responses. This re-interview of the respondents, at least on some questions, is aimed at validating the field data.

The next stage of editing is known as central editing to have a thorough check for consistency and accuracy. Generally, central editing means editing by a single editor for ensuring maximum consistency. However, for large scale studies one editor may deal with one section only. In that case inconsistencies in answers among different sections cannot be verified. In case such checks are very much needed, the. re-searcher. may identify the points of possible inconsistency and may entrust one editor to look into such points of inconsistency. In case of inconsistency, one may either decide about the proper answer based on information provided in other sections or contact the respondent for correct information, provided budget of research permits to visit the respondents second time.

Once the editing work is over,coding work starts. By coding we mean assigning numeric or alphanumeric codes to symbolically present the response against a data item. There is a chance that a presentation that categorizes responses into limited number of categories may lead to loss of some information. But this loss of informa-tion leads to efficient analysis of data.

There are four rules that may govern the formation of categories. Formation of categories should be such that these categories will be appropriate for the research problem and the research purpose. By *appropriate* we mean suitability of the codifi-cation system for testing the null hypothesis and for showing the relationships among some variables and/ or among some classes. After appropriateness comes the ques-tion of *exhaustiverless* of the codification system. If the codification system is not exhaustive there is a chance that full range of information cannot be captured. Only a smaller part may get reported due to smaller number of alternative classification codes. *Exclusiveness* is the next requirement. Exclusiveness means clear categoriza-tion of information. For two exclusive codes the respondent can clearly choose his/ her responding code. In this way, for multiple exclusive codes, it becomes easier to identify the correct code of presentation and present both correct and complete informa-tion. Lastly, there is always a need to have a *common principle* for codification of all the items.

It is desirable that a codebook be prepared and kept ready for reference purpose. Pre-coding is having added advantages and it is desirable that, to the extent possible, questionnaire should be pre-coded. Data entry will be an easier task and the time taken for data entry will be much less if the questionnaires are pre-coded. However, pre-coding is possible for cases where possible alternative answers can be antici-pated. In case of open questions, categorization of answers and hence codification can be undertaken after going through the responses received from the field enquiry.

8.2.2. Data examination

It ls desirable to discover the important aspects of data and then carryout analysis C [

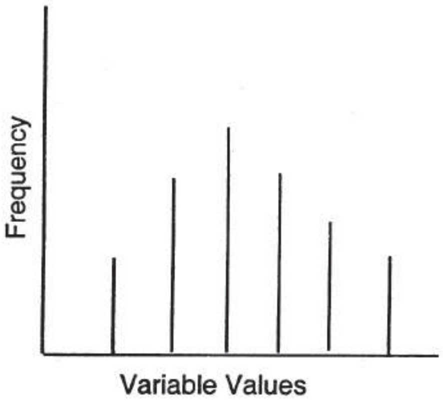
for final confirmation. This means problems should get higher importance than tools and techniques. There are approaches where data guide the subsequent analysis. These are known as exploratory data, analyses. Under such approaches one searches for important evidences, significant clues and hidden directions. There are visual presentations and graphical displays to help the exploratory works. These are more sensitive tools than summary statistics, which mainly present the average behavior in respect of location, dispersion or shape of the underlying distribution.

Frequency table is a simple but basic instrument for displaying data. It arranges data from lowest to highest value, either individually or in groups. It describes the frequency of occurrence of those individual values or groups of values so as to indicate their relative importance. Following is an example of frequency table for individual values. It describes accidents in mines per month covering 200 mines. It may be noted that for 70 mines there is no occurrence of accident; for 50 mines there is one accident per month and so on.

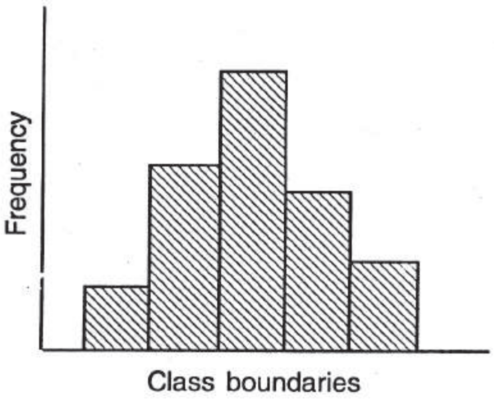
No. of accidents in a mine per month	**Frequency**
0	70
1	50
2	40
3	25
4	5
5	7
6	3
7	0
Total	**200**

To represent such a distribution through graphs one may plot along the horizontal axis the variable values (here, the no. of accidents) and along the vertical axis the frequency values. Such a diagram is known as a **column diagram** or a **frequency bar diagram.** If top of the columns are joined by line segment to get a closed polygon the resultant diagram is known as frequency polygon. In case of a continuous random variable (like height, weight etc.) it will be meaningful to develop frequency table against different non-overlapping but exhaustive intervals of values. Thse intervals are known as class intervals. Generally, class width obtained as the difference be-tween the upper class boundary and the lower class boundary is kept constant across class-intervals. In such a case one may calculate frequency density for each class interval where frequency density = frequency / class width:

In place of column diagram one may construct **histogram** for such continuous cases where along the horizontal axis one marks the class boundaries and along the vertical axis one marks the frequency densities. Then the rectangles are constructed for each class-interval by considering the class width as *one* side of the rectangle and frequency density as the other side of the rectangle. The area of the rectangle drawn on each class interval will represent the frequency of that class interval. Typical looks of a column diagram and of a histogram are given below.



Column diagram



Histogram

A popular exploratory data analysis technique which is very similar to histogram approach is known as **stem-and—leaf diagram.** Stem-and-leaf diagram has several unique advantages. For example, there is no loss of information as this diagram presents in column/ row form the actual values of the variable. It presents {he distri-bution of actual values of the variable within the selected exhaustive but non overlapping intervals. The range of values is clear from a look at the diagram and the exact range value can be realized. If we present row-wise data then each row will represent a stem and each individual datum will represent a leaf. A typical example of a stem-and-leaf diagram is given below based on 20 observations on heights in inches.

Stem 58 58 59 59

Stem 60 61 61 63 65 66 66 66 69

Stem 71 71 73

Stem 80 83

In case we present the stems along the horizontal axis the resultant look will be similar to column diagram or histogram.

Another exploratory, data analysis technique is **boxplot technique.** It is an exten-sion of five-number summary of data in terms of median, upper quartile, lower quaftile, the largest observation and the smallest observation. Unlike mean and stan-dard deviation, which are non-resistant statistical measures, the above mentioned five measures are resistant to change. In the box plot there is a rectangular plot that takes care of 50% of the data values. It has a central line that represents the position of the median and is parallel to the width of the plot. The edges of the plot are known as **hinges.** From the right and left hinges one has lines drawn up to the maximum and minimum values respectively. These lines are known as **whiskers.** For example, for the height data presented for the demonstration of stem-and-left diagram, the-median value is 66. The lower quartile is 60.5 and the upper quartile is 72. The maximum value is 83 and the minimum value is 58. Then the box plot can be drawn as follows :

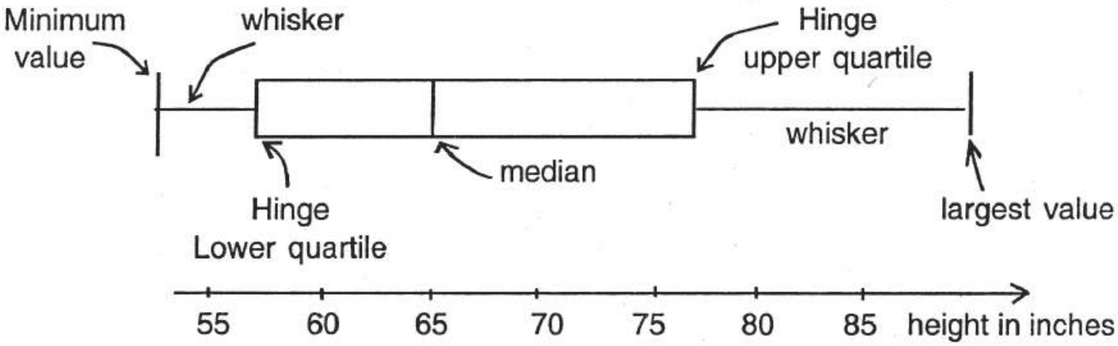


Diagram: Boxplot

To determine the outliers one has to draw the· inner fence and outer fence along the two sides of the plot. The inner fence is 1.5 times the interquartile range (IQR = upper quartile - lower quartile) added to the upper hinge to get the upper value of the inner fence. To get the lower value of inner fence we have to subtract 1.5 times the IQR value from the lower hinge. The upper value of the outer fence is obtained by adding 3 times the IQR value to the upper hinge and the lower value of the outer fence is obtained by subtracting 3 times the IQR value from the lower hinge. An observed value is said to be outlier if it lies between the inner fence and outer fence· in either direction. Observations that lie outside the outer fences are known as extreme or far outside values.

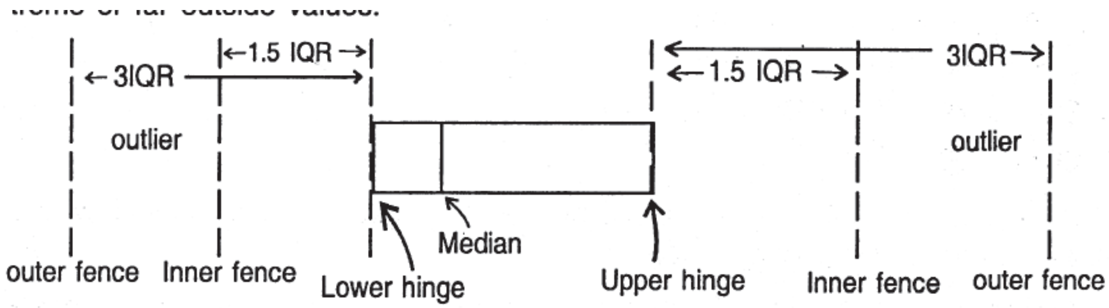


Diagram : Boxplot analysis for outliers and for outside values

In case one can separate out outliers and for outside values it may be helpful is reducing the error for drawing conclusion about average behavior in the population. However, outliers are unusual observations and are important sources of information. These units may be separately studied for the in-depth analysis and drawing specific conclusions.

The box plot can also be used for other purposes. It is easy to conclude from the box plot whether the underlying distribution is nearly symmetric or positively skewed or negatively skewed. It is also a general practice to construct multiple boxplots for multiple groups of observations to make a comparative study *among* the groups. Equality of the group medians can be studied from the multiple box plots. In this sense boxplots may play a role close to the hypothesis testing.

Sometimes there is a need to re-express the data set on a new scale using a single mathematical transformation for each set point. This need arises out of com-patibility with other data set or improved interpretation. Sometimes transformations are undertaken to stabilize the variance or to ensure symmetry in the data. Standard-ization with respect to location and scale is a very common linear transformation used for comparing different sets of scores, determination of normal probability values and making the original data free of the unit of measurement. Non-linear transformations are also being used to stabilize variance of the test statistic. For example, for sample proportion, p, the corresponding transformation for stabilization of variance is Sin-^{1}(p)$^{1/2}$, The stabilized variance is 1/(4n) where n is the sample size. For a Poisson variate, X, the corresponding transformation is (X)^{12}, the stabilized variance being (1/4). Similarly, transformations for stabilization of variance of sample variance, s^2, is log s^2 and sample correlation coefficient, r, is z = tanh^{-1}r. A popular non-linear transformation is power transformation. One may use **spread-and-level** plots to arrive at a reasonable choice of power transformation. In the spread-and-level plot the log of the median is plotted against the log of the interquartile range to find the slope of the plot. The power transformation is x^P where the original variable is x and where p = 1- slope.

8.2.3. Data mining

The idea of data mining is similar to mining activities where ores are extracted from the mother earth. In data mining one extracts knowledge from the database to find **cut** the novel and useful patterns of information that may lie hidden in the data set. ihe evolutionary steps of data mining are **Data collection** for retrospective static data delivery, **Data access** for retrospective dynamic data delivery at record level **Data navigation** for.retrospective dynamic data delivery at multiple levels and **Data mining** for prospective, proactive information delivery. Since data mining discovers or at-tempts to discover rules and patterns in the data set, the decision-makers get largely benefited from data mining in their process of supporting, reviewing and examining the current business decisions. · ‘

Data mining is frequently used in management research for pattern discovery and for prediction of trends and behaviors. Under pattern discovery, one may find out retail purchase patterns, extent of ‘fluctuation in sales volume, frauds committed, if any, and error committed, if any, during data entry. Regarding prediction of trends and behav-iors, one may employ data mining packages to identify the schemes that are more effective in attracting customers, to determine the population segments in respect of behavioral similarity, to develop risk models for different segmented markets, to dis-cover the significant buying trends and many more specific predictions.

There are many data mining techniques available at present. Clustering, data visualization, artificial neural networks, free models, classification are some important techniques widely used for generation of business knowledge. In complex data mining problems, one may make use if genetic algorithms, fuzzy logic and fractal transforms. Genetic algorithms are useful for optimization through directed search and identification of meaningful interdependencies. Fuzzy logic is an extension of traditional Bool-ean logic. In the Boolean logic there are only two possibilities: true or false. The fuzzy logic has enlarged the possibility-set by accommodating the concept of partial truth that lies in between the extreme states of truth and false. Fractal transforms are extremely useful in identifying a small group of units of similar characteristic from a very large group of units. It works based on the principle of data compression.

Data mining process mostly passes through five steps. The first step is to decide whether the entire data set will be used or a sample part of the data set will be used. For high volume data, or for limited processing power, one may opt for sample data set in place of complete data set. In case it is needed to study the behavioral pattern for every individual record, there is no other way out but to handle the entire data set. Since we decide about sampling of the data set or· not, this step of data mining process is known as **Sample** step.

The next step is **Explore** where sampled data or complete data are visually or numerically studied to identify the group behavior, if any, or behavioral trend of the data. Also the outliers are identified and studied at this stage for a possible cleaning of the data or a possible enlargement of the data. The third step is **Modify** which means undertaking of data modification based on the ideas generated during the stage of exploration. Data modification may involve data reduc-tion in respect of dimension of the study. Principal Component Analysis is a useful statistical tool for reduction of dimension. Clustering and or fractal transform may be employed for formation of groups or compression of the data. There may be addition or transformation of the data for gaining better in sight into the problem. The fourth step is **Model** where construction of a model is the subject of interest for both describ-ing the data and getting described by the data. The last step of data mining process is **Assess** wherein the performance of the model is being studied. **A** general practice is to apply the model on that part of the data, which are not used during the stages of exploration, modification and modeling. In case the entire data are used for the development of the model,! the model can be applied on some known data. If the model predicts the actual behayior, which is known in this case, we may claim the model to be a good one.

8.3 Data Tabulation

Data tabulation is aimed at presentation of data in a tabular form. In fact, there are various types of presentation. Depending on the way the researcher would like to view the data, the mode of presentation may be decided upon. In case the researcher is interested to know how a particular situation changes over time, he/she is having in mind **time-series data or historical data.** In case the researcher is interested in region-wise information his/ her way of looking will result in **spatial-series data.** In both time-series data and spatial-series data, the identity of individual data is important information and should always be kept in record. During the analytical studies, this information should be made use of to provide a critical insight into the data. Since individual data are important, we may club these two types of data under a common name· of **non-frequency data.**

There are situations where identify of individual data is not important. The re-searcher, in these situations, is more interested in the behavior of the group to which those individuals or units, belong. This type of data is known as **frequency data** because in such a case researcher is more interested to know how frequently a particular phenomenon occurs rather than which unit is having that characteristic.

The presentation of data may be textual or tabular. Tabular presentation is more popular because it gives a very compact presentation. There are different parts of a fable and these parts, taken together, lead to a complete presentation. A table must . have a title to briefly describe the contents of that table. Title should have an identifying number for future reference. Within the structure of the table there should be stub at the extreme left for describing the row characteristics/ and there should be caption in the upper part for describing the various columns of the table. The title, stub and caption jointly constitute the box head of the table. The body of the table is the main part where figures are given as per row-column description. At the end of the table footnotes are included to indicate the data source, the scope and coverage and reliability of particular data-elements.

It is desirable that the table should be balanced in number of rows and columns. Too long or too wide tables are not desirable. Arrangements of items of information should be logical and sequential in nature.

Though tables are made of numerical information numerical data may not be necessarily available at the start. For some cases we start with numerical data because the character under study is quantitative in nature. For example, heights, weights, incomes of individuals are having quantitative character. Quantitative characters are known as variables. There are characters that,a/e not quantifiable. These are known as qualitative characters, commonly referred as attributes. For example, awareness about a brand is an attribute. There may be two possibilities. Respondent may be aware of it or may not be aware of it. If n respondents are studied we may initially collect n qualitative information. But. when we start counting the number of respondents who are aware of the brand we start generating quantitative information. If number of respondents are aware of the brand, then (n- f) number of respondents will be unaware of the brand. The entire information can be summarized as follows:

Table

Result of survey on brand awareness

State of knowledge	**Number of respondent**
Aware	**f**
Unaware	**n - f**
Total	**n**

In case n = 500, f = 375, i.e., out of 500 respondents 375 are aware of one brand, then remaining 125 respondents are unaware of the brand. Thus, in the population estimated proportion of awareness is p=375/500=0.75. This is also known as relative frequency.

Relative frequency of awareness being 0.75, relative frequency of un-awareness is (1-0.75) = 0.25.

It is not a must that for an attribute there will be necessarily two classes. Number of classes may be two or more. If there are k classes with ti as the frequency of the i-th class, i = 1,2.....k and n as the total number of respondents, then

n=f, +f,+.........+ f,

Further, relative frequency of the i-th class will he given by (f/n) i= 1, 2,......, k.

Tabulation for a quantitative character (i.e., a variable) depends on the nature of the variable values. If the variable takes isolated values we term it as discrete variable. If one variable takes any value within a range of variation it is called a continuous variable. For a discrete variable, if x_i, x_2...., x_i, ... are the variable values then tabulation scheme presents the frequency f_i, with which x_i, appears in the sample of n observations. For a finite number of variable values, say x_i, i = 1, 2, ..., k/the frequency table will look like the following :

Value of the variable	**Frequency**	**Relative frequency**
x_1	x_1	x_1
-	-	-
-	-	-
-	-	-
x_i	x_i	f/n
-	-	-
-	-	-
x_k	f_k	f_k/n
Total	**$n=f_1+f_2+...+ f_k$**	**1.00**

The cumulative frequency of less than type for the i-th class, as obtained from the above table, is given by

$$F_i = f_1 + f_2 + ... + f_1 \ i = 1, 2, ..., k.$$

The cumulative frequency of greater than type is given by

$$S_1 = f_i + f_{i+1} ... + f, \ i = 1, 2, ..., k.$$

For infinite number of variable values we may allow a reasonable number of classes to exclusively and exhaustively cover all the units and all the variable values, the last class being an open ended one. For example, for the case of number of accidents there cannot be any upper bound on the number. To make a finite presentation, one may consider classes as 0 accident, 1 accident, 2 accidents, 3 accidents and 4 and more accidents. Mostly, when we deal with an attribute or a discrete variable the nature of the data may try to indicate the desirable nature of the class formation. A researcher has to listen to the variation in the data set to arrive at a suitable classification system.

In case of a continuous variable, the classification is to a g·reat extent artificial in nature. A researcher has to decide about the classification system within a general framework that classes must be exhaustive and exclusive in nature, the number of classes should neither be large nor small and the class widths should preferably be equal. For a k-class classification if l_i; is the lower class limit and u, is the upper class limit, i =1,2....., k, then choice of u, and l_{i+1} should be such that even theoretically no observation can lie between u_i and l_{i+1}, i= 1, 2, ... , k - 1. The class frequency f_i, indicates that there are f_i, observations that lie between l_i, and u_i, inclusive of l_i, and u_i, i=1,2, ..., k. However, to retain the continuous character of the data set we introduce the concept of class boundaries in the sense that the lower and upper class boundaries of the i-th class will be $(u_{i+1}+l,)/2$ and $(u_i+l_{i+l})/2$ respectively, i = 1,2, ..., K, where u_0 and l_{k+i} are conceptual possible values of two extreme classes. The class mark of the i-th class is given by $(u_i + 1_1)/2$,i= 1,2, ..., k. Further, it is desirable that class widths should be equal, i.e., $\frac{(u_i + l_{i+1})}{2} - \left(\frac{u_{i-1} - l_i}{2}\right) = C$ a constant for all i =1, 2, ..., K. Let us explain these concepts using the following table :

.

Table

Frequency table on daily per capita expenditure (in Rs.)

Classes	**Frequency**
0- 19	14
20 - 39	**23**
40- 59	**27**
60 - 79.	**21**
80 - 99	15
Total	**100**

Consider the class number 3 for which the class limits are 40 and 59. Thus 1_3 = 40 and u_3 = 59. The class boungares are :

lower class boundary of the 3rd Class = $(u_2+ 1_3)/2=(39+40)/2=39.5$,

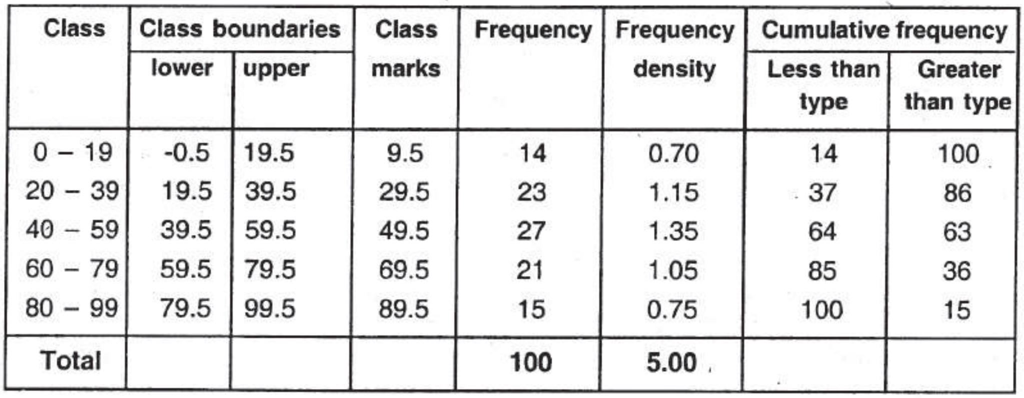
upper class boundary of the 3rd class =$(u_3+ 1_4)/2=(59+60)/2=59.5$.

Here class width is a constant and is equal to (59.5 - 39.5) = 20. Further, class mark of the 3-rd class is (40+59)/2=49.5.

The following table presents class boundaries, class marks, frequency, frequency density and cumulative frequency of less than and greater than type for all the 5 classes for the above problem.

Table

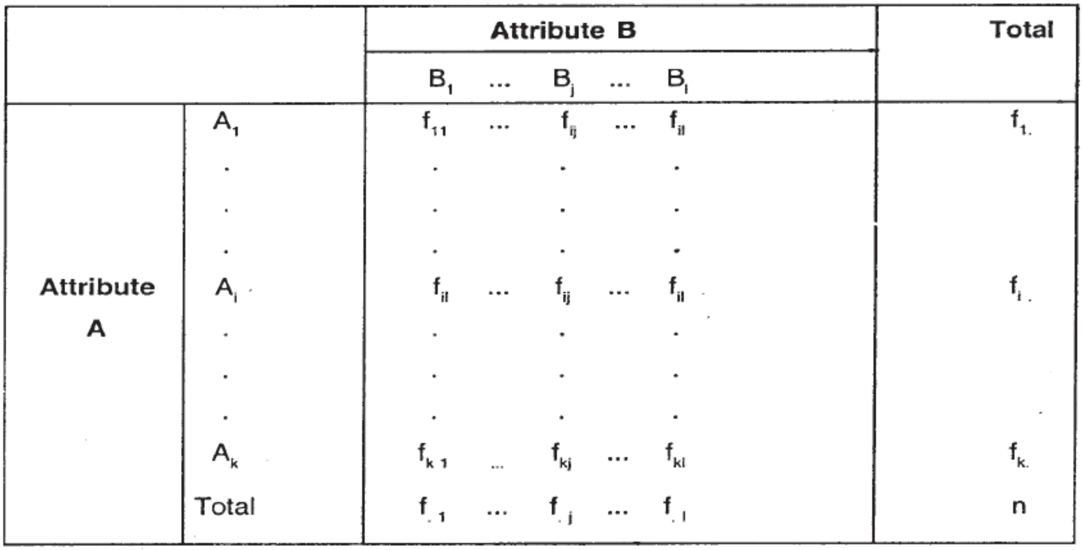
Detailed frequency table on daily per capita expenditure (in. Rs.)



Class	Class boundaries		Class marks	Frequency	Frequency density	Cumulative frequency	
	lower	upper				Less than type	Greater than type
0 – 19	-0.5	19.5	9.5	14	0.70	14	100
20 – 39	19.5	39.5	29.5	23	1.15	37	86
40 – 59	39.5	59.5	49.5	27	1.35	64	63
60 – 79	59.5	79.5	69.5	21	1.05	85	36
80 – 99	79.5	99.5	89.5	15	0.75	100	15
Total				**100**	**5.00**		

Quite often, one observes more than one character at the same time for each unit/ individual. The corresponding tabular presentation giyes rise to multi-character fre-quency table. In case of two characters there will bea two-way classification. For two· attributes A and B with A occurring ink forms A_1, A_2, 2...., A_k, and B occurring in I forms B_1, B_2........, B_i, the number of occurrences of A_i and B_i in n observations be f_{ij} i = 1,2, ..., k, j = 1,2....., ...1. These give rise to the following bivariate table :

Table



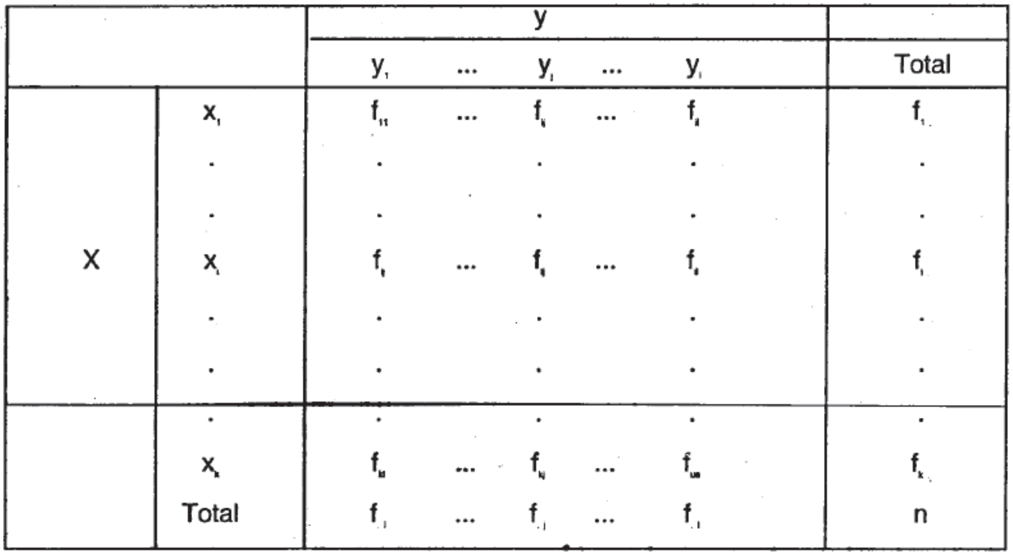
		Attribute B					Total
		B_1	...	B_j	...	B_l	
Attribute A	A_1	f_{11}	...	f_{ij}	...	f_{il}	$f_{1.}$
	.	.		.		.	
	.	.		.		.	
	.	.		.		.	
	A_i	f_{il}	...	f_{ij}	...	f_{il}	$f_{i.}$
	.	.		.		.	
	.	.		.		.	
	.	.		.		.	
	A_k	f_{k1}	...	f_{kj}	...	f_{kl}	$f_{k.}$
	Total	$f_{.1}$	...	$f_{.j}$	...	$f_{.l}$	n

If we ignore the attribute B then the last column presents the frequency distribu-tion of attribute A. Similarly, if we ignore the attribute A the last row presents the frequency distribution of attribute B. Here

$$f_{.j} = \sum_{j-1}^{k} f_{ij}, \ \ j = 1, 2, \ \ ......, 1,$$

$$f_{i.} = \sum_{i-1}^{1} f_{ij}, \ \ i = 1, 2, \ \ ......, k,$$

For two discrete variables X and Y taking values x_1, x_2, ..., x_k and y_1, y_2, ..., y_1, respectively the corresponding bivariate frequency table will take the following look :



		y					
		y_1	...	y_j	...	y_l	Total
	x_1	f_{11}	...	f_{1j}	...	f_{1l}	$f_{1.}$
	.	.		.		.	.
	.	.		.		.	.
X	x_i	f_{i1}	...	f_{ij}	...	f_{il}	$f_{i.}$
	.	.		.		.	.
	.	.		.		.	.
	.	.		.		.	.
	x_k	f_{k1}	...	f_{kj}	...	f_{kl}	$f_{k.}$
	Total	$f_{.1}$	...	$f_{.j}$	...	$f_{.l}$	n

In the above discrete bivariate frequency table f represents the joint frequency of occurrence of x, and y, in n observations. As usual

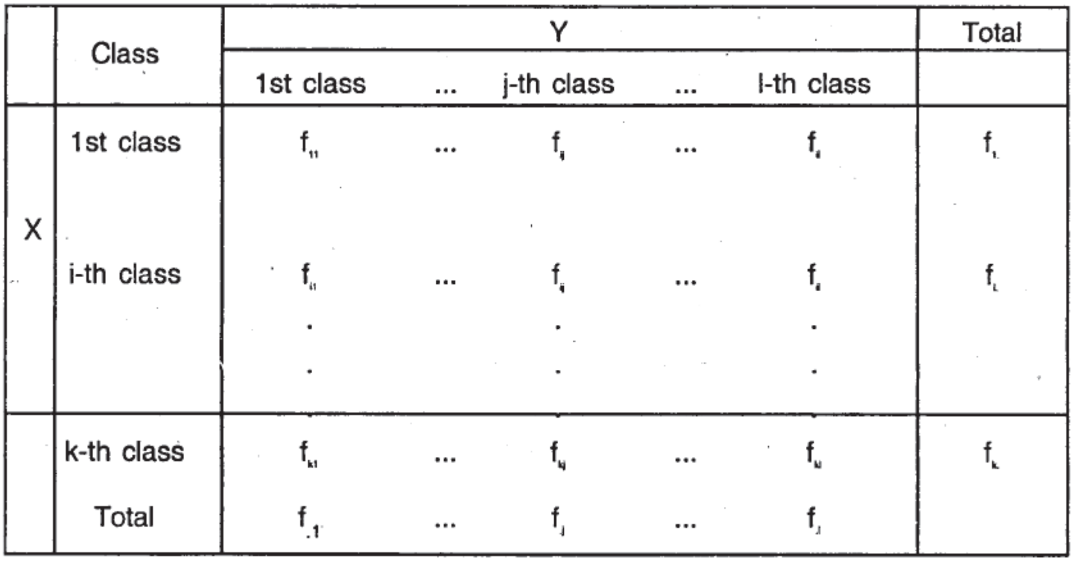
$$f_i = \sum_j f_{ij} \text{ and } f_{.j} = \sum_i f_{ij}$$

$$f_{i.} = \sum_{j=1}^{1} f_{ij} \text{ , } i = 1, 2........, \ k,$$

for i = 1, 2, ..., k. j = ... 1 and n is the sum of all the frequencies.

$$\text{i.e.n} = \sum_i \sum_j f_{ij}$$

For bivariate continuous variables the i-th class with respect to the variable X be denoted by (xl, xu) and the j-th class with respect to the variable Y be denoted by yl, yu). If f_{ij} is the frequency of occurrence of variables belonging to both (xl, xu,) and (yl, yu) classes for X and Y respectively then such a classification will give rise to a bivariate frequency table.



	Class	Y					Total
		1st class	...	j-th class	...	l-th class	
X	1st class	f_{11}	...	f_{1j}	...	f_{1l}	$f_{1.}$
	i-th class	f_{i1}	...	f_{ij}	...	f_{il}	$f_{i.}$
		·		·		·	
		·		·		·	
	k-th class	f_{k1}	...	f_{kj}	...	f_{kl}	$f_{k.}$
	Total	$f_{.1}$	...	$f_{.j}$	...	$f_{.l}$	

The last column gives rise to the frequency distribution of X, also known as marginal distribution of X. The last row gives rise to the frequency distribution *of Y,* also known as the marginal distribution of Y.

8.4. Data analysis

Initial analysis of data is to focus on some basic features of the data. These fea-tures describe the data in a general way summarizing the nature of variation in the data set. It has been found that observations vary in such a way that there is a tendency to cluster around a central value. This is known as the central tendency of the data. Mostly, we **measure Pntral tendency** in terms of an average. The most commonly used average is the arithmetic mean. Other widely used measures are median and mode. **Arithmetic mean** is obtained by dividing the sum of values of all the observations by the number of observations. If there are k possible values, i-th value being x_i with frequency of occurrence as f_1 then the arithmetic mean (AM) is defined as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} f_i x_i \text{ where, } n = \sum_{i=1}^{k} f_i$$

For a frequency table of a continuous variable we may approximate x_i by the class mark of the i-th class. The resultant value will be a close approximate of the sample mean $(\overline{x})$.

Median is the middle most value of the observations when arranged either in ascending or in deescending order of magnitude. If the total number of observation, fn is odd, then there is a unique medium and the value of the median is the {(n+1)/2}-th observation of the ordered arrangement. In case n is even, median is the arithmetic mean of the (n/2)-th and (n/2 +1)-th observations of the ordered arrangement.

Mode is the value of the variable having the maximum frequency of occurrence.

The extent of variation in a data set is measured in terms of the tendency of deviation from the central value. There are three widely used **measures of deviation/ dispersion.** These are range, mean deviation and stand and deviation. **Range** is the simplest measure of dispersion, measured in terms of the difference between the highest and lowest observed values. **Mean deviation** is the arithmetic mean of the absolute deviations of individual observations from any measure of central tendency. If C is any measure of central tendency, the mean deviation (MD) about C is given by

$$MD_c = \frac{1}{n}\sum_{i-1}^{k} f_i\,(x_i - C)$$

For a continuous frequency table, x_i is approximated by the class mark of the i-th class.

Standard deviation (SD) is the position root mean square deviation about. arithmetic mean and is given by the formula ;

$$SD = \left\{\frac{1}{n}\sum_{i-1}^{k} f_i\,(x_i - \overline{x})^2\right\}^{\frac{1}{2}}$$

The relative measure of dispersion is the ratio of the measure of dispersion and the measure of central tendency. One widely used relative measure of dispersion is the **coefficient of variation** which is the ratio of the standard deviation and the arithmetic mean. That is, $CV = \frac{SD}{x}$

$$= = \sqrt{\frac{1}{n}\sum_{i=1}^{k} \frac{(x_i - \overline{x})^2}{\overline{x}^2} f_i}$$

This is the standard deviation of the normalized observations (x/AM) i = 1,2, ..., k, where AM stands for arithmetic mean.

The extent of departure of a frequency distribution from its central value is known as

skewness of the frequency distribution. For a unimodal distribution the symmetry ensures equality of mean, median and mode values. For a positively skewed distribution

mean > median > mode

and for a negatively skewed distribution

mean < median < mode

Hence, a measure of skewness (SK) is given by

$$\text{SK} = \frac{\bar{x} - \text{Mode})}{\text{SD}}$$

Alternatively, SK=$\text{SK} = \frac{3(\bar{x} - \text{Median})}{\text{SD}}$

Skewness can also be measured in terms of central moment of order 3, m_3, where the central moment of order r is given by m_3

$$m_3 = \frac{1}{n}\sum_{i=1}^{k} f_i (x_i - \bar{x})^3$$

and the measure of skewness, g_1, is

$$g_1 = \frac{m_3}{(\text{SD})^3}$$

The other feature of a frequency distribution is the degree of peakedness, also known as **kurtosis.** A mesokurtic distribution is the one that exhibits moderate peakedness. A leptokurtic distribution is one that exhibits high peakedness. Flatness is the key feature of a platykurtic distribution. Writing m, as the central moment of order 4 where

$$m_4 = \frac{1}{n} = \sum_{i=1}^{k} (x_1 - \bar{x})^4 f_i$$

we may suggest a measure of kurtosis as g_2

$g_2 = m_4/(\text{SD})^4$

For a bivariate situation, measures of relationship are of importance and use. Product moment coefficient of correlation measures the degree of correlation between two variables via covariance. If there are n pairs of observations (x_i, y_i), i= 1,2, ... ,n, on two variables X and Y, then the covariance between X and Y, denoted by Cov(X, Y) can be measured as follows:

$$\text{Cov (X, Y)} = \frac{1}{n} = \sum_{i=1}^{k} (x_1 - \overline{x})(y_i - \overline{y})$$

The coefficient of correlation between x and y, r_{xy}, is given by

f_{xy} =Cov (X,Y)/(SD$_x$SD$_y$)

where SD$_X$ and SO$_Y$ are the respective standard deviation of X and Y.

·The value of r_{xy} lies between 1 and -1. If correlation coefficient is either 1 or -1 variables X and Y are linearly related with positive slope in case of 1 and negative slope in case of -1.If correlation coefficient is close to either 1 or -1 we have reasons to believe that variables X and Y are almost linearly related with positive slope in case of 1 and negative slope in case of -1. Zero value for r_{xy} means lack of correlation but not necessarily independence.

In case x and y are measured on ordinal scale one may use ranks in place of ordinal values. The resultant measure of correlation coefficient is known as rank correlation coefficient. The formula simplifies to

$$r = 1 - 6\sum d_1^2 / \{n(s^2 - 1)\}$$

where d. is the difference between the ranks of the i-th individual in respect of the two variables/attribute, and n is the number of pairs of observations.

In case the correlation coefficient has a high absolute value one may make use of the linear relationship to predict one variable based on the knowledge on the other variable. For example, variable y may be predicted by Y based on the knowledge on x, prediction equation being

Y =a+ bx.

To estimate a and b one may take the help n pairs of observations (x, y,), i = 1, 2,... , n, on x and y and employ the least square method to choose a and b in such a way that

$$S = \sum_{i-1}^{n} (y_i - a - bx_i)^2$$

is minimized. Differentiation of S with respect to a and b gives rise to least squares normal equations for the determination of a and b. These equations are given below:

$$\sum_{i-1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$$

Solving these equations we get solutions $\hat{a}$ and $\hat{b}$ for a and b as

$$\hat{a} = \overline{y} - \hat{b}\overline{x}$$

In case of attributes A and B we examine the. extent of association in terms of Pearson's coefficient of contingency, C_{AB}, where

$$C_{AB} = \{\chi^2_{AB} / (n + \chi^2_{AB})\}$$

with

$$\chi^2_{AB} = n[\sum f^2_{ij} / (f_{i.} f_{.j}) - 1]$$

Since the upper bound of C_{AB} is less than 1, one may consider Tschuprow's mea-sures of association given by T_{AB} where

$$T_{AB} = [\chi^2_{AB} / n\{(k-1)(1-1)\}^{\frac{1}{2}}]^{\frac{1}{2}}$$

where k is the number of forms of attribute A and I is the same for attribute B.

8.5. Summary

Tabulation and analysis works follow the stage of data collection. To make the tabulation work easier the collected data must pass through the stages of data prepa-ration, data examination and data mining. Data preparation involves editing of data to ensure accuracy, consistency and completeness. Data editing may be carried out in the field and is known as field editing. After field editing one undertakes central editing to have a thorough Sheck preferably by a single editor. After the editing work, starts· the work of codification to categorize the data for efficient analysis. Codification should be done in such a way that appropriateness, exhaustiveness, exclusiveness and uniformity in principle are adhered to.

Data examination aims at discovering important aspects of data. It identifies the problem and attaches higher importance to problem than tools and techniques to be used for analysis. To examine the data one may take the help of visual presentation and graphical displays. Frequency table is a simple but 'basic instrument for displaying data, common types of diagrams are column or frequency bar diagram and histo-gram. A popular exploratory analytical technique is stem-and-leaf diagram where each row of the row-wise presented data represents a stem and each individual data represents a leaf. Another important exploratory technique is boxplot technique. It can identify the outliers and extreme observations/for outside values.

Sometimes there is a need to re-express the data set in terms of some mathematic transformations. Transformation may be of use to ensure compatibility with other data sets to improve interpretation and to stabilize variance. A popular non-linear transformation is power transformation. One may take the help of spread-and-level plot to identify the power of the transformation.

Data mining is to extract knowledge from the database. The evolutionary steps of data mining are data collection, data access and data navigation. It is widely used for pattern discovery and for prediction of trends and behaviors. Sample, explore, modify, model and assess are the five stages through which the data mining process passes through.

Data tabulation is aimed at presentation of data ‘in tabular form. **Mode** of presentation depends on the way researcher would like to view the data. There are time series data and spatial series data, frequency data and no-frequency data. In tabular presentation, i.e. presentation in the form of tables, tables must be of balanced number of rows and columns with clear description of the subject of interest, rows and columns. Tabulation method depends on the nature of character under study. A character may be an attribute or a variable, A variable may be discrete or continuous. Formation of class depends on the nature of the character.

Data analysis at the initial stage involves calculation of measures to summarize the data set. There are features of interest like central tendency, dispersion, skewness and kurtosis. For each feature alternative measures are available. A researcher has to make a choice from among different alternatives. In case of a multi-character study a measure of relationship assumes importance. For a bivariate set up, for interval data or ratio data one may calculate correlation coefficient and for attributes one may use Pearson’s co-efficient of contingency or Tschuprow’s measure of association. Regres-sion analysis is also of use for predicting one variable, from the knowledge of the other variable in case the correlation coefficient is close to + 1 or -1.

8.6. Questions

Long answer type

1. What do you mean by data preparation? Explain the concept of editing of data indicating the need for the same.
2. Describe the four rules of data categorization.
3. Indicate how boxplot technique can be used for data examination.
4. Describe a few graphical techniques for examining data.
5. What is a spread-and-level plot ? Suggest a few transformations for stabilization of variance.
6. Enlist the evolutionary steps of data mining, briefly explaining each one of them.
7. Indicate the need for data mining and describe some techniques available for this purpose.

8. Describe the five stages of data mining process.

9. What do you mean by tabular presentation of data ? Describe the main features of a table.

10. Enlist the basic features of the data, a researcher is commonly interested to measure. Suggest a suitable measure for each of the enlisted features.

Short answer type questions.

1. Explain stem-and-leaf diagram.

2. What is interquartile range *?* How is it used in boxplot technique ?

3. Give one example each for, time series data and spatial series data explaining their importance.

4. Explain the classification system for a continuous data set.

5. Calculate arithmetic mean, standard deviation, g, and g, values *for* the data set, given in the text, on daily per capita expenditure (8.4}.

6. Refer to accident data in section 8.2.2. Find out the mean, median and mode values. Also calculate range and mean deviation about median.

7. Using the following data set on sales and expenditure on advertisements calcu-late the correlation coefficient (figures given in Rs. 00

Sales	**Expenditure on Ad.**	**Sales**	**Expenditure on Ad.**
340	45	482	48
460	50	450	46
480	50	500	49
520	62	495	47
470	47	500	48

8. Using the Sales and expenditure on advertisement data as given in problem 7 determine the regression equation for predicting Sales (y) from expenditure on adver-tisement (x).

9. Rank the data set of problem 7 and calculate rank correlation coefficient. 10. Compute a measure of association for the following data.

	Male	**Female**	**Total**
TB of respiratory system	354	176	530
other forms of TB	140	130	270
Total	**494**	**306**	**800**

Objective type questions

Indicate whether the following statements are true or false.

1. By data preparation we mean editing of data, coding of information and data entry.

   True □ False □

2. Field editing is also known as central editing.

   True □ False □

3. Categorization of data into listed number of categories may lead *to* loss of some information.

   True □ False □

4. If codification system is not exhaustive full range of information cannot be captured.

   True □ False □

5. Code book is kept ready for reference purpose.

   True □ False □

6. Column diagrams an improvement over frequency bar diagram. n

   True □ False □

7. Stem-and-leaf diagram is a technique for detection of outliers

   True □ False □

8. In boxplot, edges of the plot are knows as whiskers. h]

   True □ False □

9. Inner fence is 1.5 1QR away from upper hinge.

True □ False □

10. Far outside values are known as outliers.

True □ False □

11. Spread-and-level plot is used for variance stabilization.

True □ False □

12. Data mining mostly passes through five steps.

True □ False □

14. Range is a measure of central tendency

True □ False □

15. For a negatively skewed distribution mean < median < mode.

True □ False □

Unit - 9 □ DI Estimation and testing. of hypothesis

Structure

9.1. Introduction

The basic objective of data collection and subsequent data processing is to estimate or test for the population characteristics of research importance. For example, we may ·tqtbe interested text know the per person per day average consumption of milk in the state of West Benga1. For this, we may draw a random sample of 100 persons to enquire *4.*aPout their per day consumption of milk covering the state of West Bengal. Average per day consumption of milk can then be calculated and the resultant value may be proposed as an estimate of the per person per day average consumption of milk in the state of West Bengal.

The need for model building is also experienced by the practitioners to describe the future behaviour of individuals/ units/markets. lf a model is described by a mathemati-cal form it involves parameters which are situation-dependent. These parameters must be determined to make use of the model. Such determinations based on sample information are known as estimation of parameters.

Thus, two types of estimation we need to address. The first type of estimation is for *finite population* characterstics using sample observations gathered through probabilistic procedure. The second type of estimation is for *infinite population* de-scribed through probability models and for which random sample observations are available.

In case there is a conjecture or hypothesis available with the investigator that hypothesis may be tested in the, light of the given observations. lf a probability model is given along with random sample observations the hypothesis to be tested is ex-pressed in terms of the parameters involved with the model. This type of testing of hypothesis is known as *parametric testing*. Parametric tests may *be* carried out based on *small sample exact tests* or *large sample approximate tests.* In case the underlying probability model is not specified, one adopts *non-parametric test.*

9.2. Estimation for finite population

9.2.1. Estimation of population mean under SRSWOR

Let the population be denoted by **U** = (U_1, U_2 .... , U_N) with value of the characteristic X as X_i for the i-th unit U_1. In other words, X_1= X (U_1), the value of X for U_1. Then the population mean is given by

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Let the population variance be denoted by S^2 where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_1 - \bar{X})^2$

Consider s = (U_{i1}, $U_{i\,2}$, ... , Uin), a simple random sample drawn without replacement from U. The sample size is n and (i_i,i_2..., l_n) $\subseteq$ (1,2, ... ,N) with $i_i \neq i_k$, for j $\neq$ k. Then the sample mean $\bar{x}$ is given by

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_{i_j}$$

To estimate $\bar{X}$ we may use $\bar{X}$ and it is easy to note that $\bar{X}$ is an unbiased estimator of $\bar{X}$ in the sense expectation of $\bar{x}$ is $\bar{X}$, i.e.

E($\bar{x}$) = $\bar{X}$.

The corresponding standard error (SE) is the positive root of the variance of $\bar{x}$.

Thus,

$$\mathrm{SE}(\bar{x}) = \bar{X}\sqrt{\mathrm{var}(\bar{x})}$$

can be snow that $\mathrm{SE}(\bar{x}) = \sqrt{\frac{S^2}{n}\left(1-\frac{n}{N}\right)}$

Population proportion may be considered as a special case of population mean where the character X is so .selected that it takes the value 1 when a particular atttibute is present with the population unit and value O when that particular attribute is not present with the population unit. In that case $\bar{X}$ can be viewed as P, the population proportion, $\bar{x}$ can be viewed as p, the sample proportion. Since $\bar{x}$ is an unbiased estimator of $\bar{X}$, in particular, p is an unbiased estimtor of P. The corresponding stand-ard error can be simplified to

$$\mathrm{SE}(p)=\sqrt{\frac{P(1-P)}{n}\left(1-\frac{n}{N}\right)}$$

To estimate the standard error we consider the sample variance s^2 where,

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_{i_j}-\bar{x})^2$$

It is known that s^2 is an unbiased estimator of S^2 in the sense expectation of s^2 is S^2. Then the unbiased estimator of Var $(\bar{x})$ will be

$$\frac{s^2}{n}\frac{N-n}{N}$$

with an estimator of SE $(\bar{x})$ as

$$\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$

9.2.2. Estimation of population mean under SRSWR

In case of simple random sampling with replacement, the population behaviour will be similar to an infinite population. The sample mean, $(\bar{x})$, will remain an unbiased

estimator of the population mean, $\bar{X}$, with

$$\mathrm{Var}\,(\bar{x}) = \frac{\sigma^2}{n}$$

where

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

9.2.3. Estimation of population mean under systematic random sampling

Let the population size N be an integer, multiple of the sample size n. Thus, let N = nk. Then the population can be subdivided into k possible samples each of size n. Let the r-th block be denoted by U_r,U_{k+r}, ... , $U_{(n-1)k+r}$ with corresponding mean value as

$$\bar{X}_r = \frac{1}{2}\sum_{\infty=0}^{n-1} X_{r+\alpha k}$$

In that case $\bar{X}_2$, $\bar{X}_{1.........}$ $\bar{X}_k$ are the k possible estimates of the population mean $\bar{X}$ and in case the r-th block is selected the estimator of $\bar{X}$ is : $\bar{X}_r$. The variance of such an estimator is given by

$$\mathrm{Var(sys)} = \frac{1}{k}\sum_{r=1}^{k}(\bar{X}_r - \bar{X})^2$$

with SE given by

$$\mathrm{SE}{=}\sqrt{\mathrm{var\,(sys)}{=}}\sqrt{\frac{1}{k}\sum_{r=1}^{k}\bar{x}_{r.} - \bar{x})^2}$$

9.2.4. Estimation of population mean under stratified random sampling

Let the population be made of k strata with xi as the i-th in stratum mean and S_1^2 as the i-th stratum variance. Then, the overall population mean $\bar{\bar{X}}$ is given by $\bar{\bar{X}} = \sum_{i=1}^{k} N_i\bar{X}_i \,/\, N$, where N_i is the size of the i-th stratum and N is the total population size. Then the unbiased estimator of the i-th stratum mean $\bar{X}_i$ will be the corresponding sample mean, $\bar{x}_i$, based on a simple random sample of size n_i drawn without replacement from the i-th strata. Combining these sample means with weights equal to (N_1/N) for the i-th strata we get the unbiased estimator of $\bar{\bar{X}}$ as

$$\bar{\bar{X}} = \sum\nolimits_{i=1}^{k} \frac{N_i}{N} \bar{x}_i$$

The variance of $\bar{\bar{X}}$ is

$$(\bar{\bar{x}}) = \sum_{i=1}^{k} \left(\frac{N_i}{N} \right)^2 \text{Var}(\bar{x}_i)$$

$$= \sum_{i=1}^{k} \frac{N_i^2}{N^2} \left\{ \frac{S_i^2}{n_i} \; \frac{N_i - n_i}{N_i} \right\}$$

With S_i^2 as the sample variance of the i-th the stratum· which is an unbiased estimator of the population variance S_i^2 of the i-th statum; we have an unbiased estimator of var $\bar{\bar{X}}$ given by

$$\widehat{(b_1^2 + b_2^2 + \dots\dots + b_n^2)}$$

9.2.5. Estimation of population mean under multistage sampling *l*c.

Consider, for simplicity, two stage sampling with **M** first stage units and N, second stage units for the i-th first stage unit, i = 1, 2, .. , M. Writing $N = \sum\nolimits_{i=1}^{M} N_i$ and $\bar{X}$ as the mean of the i-th second stage units, the grand mean $\bar{\bar{X}}$ will be.

$$\bar{\bar{X}} = \frac{1}{N} \sum_{i=1}^{M} N_i \bar{X}_i$$

Writing $\bar{x}_i$ as the sample mean of the second stage units based on ni second stage units drawn out of Ni second stage units, i =1,2, ... , m where m is the number of first stage units selected out of M first stage units, we have

$$\bar{\bar{X}} = \frac{1}{N} \frac{M}{m} \sum_{i=1}^{M} N_i \bar{X}_i$$

as an unbiased estimator of the grand mean $\bar{\bar{X}}$.

9.3. Estimation for infinite population

9.3.1. Linear estimator

Let us describe the infinite population in terms of a random variable X. Let the expectation of X be μ and the variance of X be σ^2. The problem of interest is to estimate μ based on random sample observations $X_1, X_2..., X_n$ drawn from X. This means $X_1, X_2..., X_n$ are random replicas of X. Let us restrict ourselves to linear estimators based on X,, ..., Let a typical linear estimator T for μ be denoted by

$T=a+b_1X_1 +b_2X_2+...+ b_nx_n$

For T to be unbiased for μ we need to ensure E (T) = μ. But

$E(T) = a + b_1 \mu+ b_2+.......+b_n\mu$

$=a+(b_1+b_2+.... +bn)\mu$

Hence a must be zero and $(b_1 + b_2 + ... + b_n)$ must be equal to 1. Thus,

$T=b_1x_1+b_2x_2 +......+ b_nx_n$

such that $1 = b_1 +b_2+ .... + b_n$.

For such an estimator T of μ the variance of T will be

$$\text{Var (T)} = b_2^1\sigma^2 + b_2^2\sigma^2 + ... + b_n^2\sigma^2$$

$$= \sigma^2 + (b_1^2 + b_2^2 + ... + b_n^2)$$

Since Var (T) measures the error in estimation we need, to minimize var (T). In other words, we need to minimize $(b_1^2 + b_2^2 + ... + b_n^2)$ subject to the Condition $b_1+b_2+...+ b_n =$

1. Using Lagrange multiplier method we construct a futction

L where $L(b_1, b_2....b_a) = b_1^2 + b_2^2 + ...b_n^2 - \lambda(b_1 + b_2 + ... + b_n - 1)$.

Differentiating L (.) with respect to $(b_1, b_2....b_n)$ and equationg each to zero we get the choice of $b_1, b_2, ..., b_n$, But

$$\frac{\partial L(b_1, b_2, ..., b_n)}{\partial b_1} = 0, \quad i = 1, 2, ..., n$$

$$\Rightarrow b_1 = \lambda/2, \ i = 1, 2, ..., n.$$

$$\Rightarrow b_1 = b_2 = ... = b_n$$

But their sum is 1. Hence each b_1 is of value 1/n. Putting tghis Optimum choice of b_i s in the expression of T we get the Best Linear Unbiased Estimator (Blue)

$$\text{as } T = \frac{1}{n}X_1 + \frac{1}{n}X2 + + ... + \frac{1}{n}Xn$$

$$= \bar{X}.$$

9.3.2. Minimum variance unbiased estimator

Let us examine the suitability of the sample mean as an estimator of the population mean when some additional information is given. This additional information will be in terms of the population distribution described through *probability mass function* (pmf) for a discrete case or through *probability density function* (pdf) for a continuous case.

X follows Poisson distribution. Let the pmf of X be denoted by

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}, x = 0, 1, ..., \lambda > 0$$

The parameter λ is the mean of the Poisson distribution. It may be proved that sample means $\bar{X}$ is not only an unbiased estimator of λ, it is the minimum variance unbiased estimator (MVUE) of λ in the sense that among all unbiased estimators of λ has the minimum variance. This is a stronger property than BLUE as there is no restriction of linearity on the proposed estimator.

X follows Normal distribution. Let the pdf of X be denoted by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(x-\mu)^2}, -\propto$$

$$\sigma > 0, -0 < \mu < \propto$$

The parameters of this distribution are μ and σ, μ denotes the population mean and σ denotes the variance. Here again, sample mean $\bar{X}$ is not only an unbiased estimator of μ but also the MVUE of μ for all choices of μ.

X follows exponential distribution. Let the pdf of X be denoted by

$$f(x) = \frac{1}{\mu}e^{\frac{x}{\mu}}, x > 0, \mu > 0$$

The parameter μ is the population mean. The sample mean $\overline{X}$ is the MVUE of μ for all choices of μ.

9.3.3. Methods of estimation

Given the random sample observations X_1, X_2......, X_n from a population described by the probability function (pdf or pmf) f,(x) where θ is the parameter of the dispihgution. One can estimate θ by various methods. The most popular method is the maximum likelihood method where the likelihood function

$$L(\theta) = \prod_{i=1}^{n} f_\theta(X_i)$$

is maximized with respect to θ. Let thiss optimum choice of θ be denoted by $\hat{\theta}$. Then $\hat{\theta}$ will be referred as the maximum likelihood estimator (mle) of θ. For a function g(θ), the mle of g(θ) is g($\hat{\theta}$). Let us consider the case of normal distribution. $\hat{\theta}$ will be a vector in this case, made of μ and σ^2, with

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(x-\mu)^2}$$

The likelihood function will be

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x_i-\mu)^2}{2\sigma^2}}$$

$$= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^2} e^{\frac{1}{2\sigma^2}\sum(x_i-\mu)^2}$$

$$= \frac{1}{\left(\sigma\sqrt{2\pi}\right)^n} e^{\left[\frac{n}{2\sigma^2}(\overline{x}-\mu)^2 - \frac{1}{2\sigma^2}\sum(x_i-\overline{X})^2\right]}$$

Taking logarithm, a monotonic transformation,

log L (μ,cr²) = -nlog $\sigma - \frac{n}{2}\log(2x) - \frac{n}{2\sigma^2}(\overline{x}-\mu)^2 - \frac{\sum(x_i-\overline{x})^2}{2\sigma^2}$

In place of maximizing L (.) we may maximize log L (.).Diferentiating with respect to μ we get

$$\frac{\sigma \log L(\mu,\sigma^2)}{\sigma\mu} = \frac{n}{\sigma^2}(\bar{x}-\mu)$$

The above expression when equated to zero gives

$$\bar{\mu} = \bar{x}$$

Differengiating log L(.) with respect to σ^2 we get

$$\frac{\partial \log L(\mu,\sigma^2)}{\partial\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{\sum(x_i-\bar{x})^2}{2\sigma^4}$$

The above expression when equated to zero gives rise to

$$\frac{\sum(x_i-\bar{x})^2}{\sigma^4} = \frac{n}{\sigma^2}$$

$$\sigma^2 = \frac{\sum(x_i-\bar{x})^2}{n}$$

Thus, the mle of μ is $\bar{X}$ and the mle of σ^2 $\frac{1}{n}\sum(X_i-\bar{X})^2$. Mies are consistent estimators which are not necessarily unbiased. Here, $\bar{X}$ is an unbiased estimator of μ but $\hat{\sigma}^2$ is a biased estimator of σ^2· These estimators are asymptotically efficient and normally distributed.

Method of moments is an important method of estimation. It equates sample mo-ments with the population moments to create a set of estimating equations. For exam-pie, in the normal case we consider the first two population moments as μ and $\sigma^{2.}$ We equate these moments with that of sample moments, i.e., sample mean and sample variance. These result in

$$\hat{\mu} = \bar{X}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2$$

Estimators obtained by method of moments are consistent estimators.

9.4. Testing of hypothesis of parametrice type

A hypothesis is a statement of belief to be tested in the light of the given data. The hypothesis to be tested is called null hypothesis. The purpose of testing is to see its tenability. This means there must be an alternative hypothesis too. The statistical testing procedure in its general form can be presented as a procedure of dividing the sample space into two non-overlapping and exhaustive parts, say ω and W. The sample space is the collection of' all possible observations of size n, say. This means the observed sample $\underset{\sim}{W} = (X_1, X_2, ..., X_n)$ is a point of the n-dimensional real space, denoted by R^n. Thus $\underset{\sim}{W} \in R^n$. Our objective is to divide R^n into two spaces ω and W so that $\omega \cap W = \phi$ and $\omega \cup W = R^n$. If $\underset{\sim}{X}$ falls in the region o we reject the null hypothesis (H_o) and if $\underset{\sim}{X}$ falls in the region W, we accept the null hypothesis. When we reject H_o we favour the alternative hypothesis (Ha). We refer as the *critical region* or rejection region and W as the *acceptance region.* While drawing inference on tenability of H_o in the above mentioned fashion, we commit two types of error. Sometimes we may reject the H_o when it is actually true and we take a wrong decision. This error is known as *type I error.* On other occasion, we may accept the Ho when it is not true. This error is known as *type II error.* In a labular form we may present the four possible situations that may arise out of the true state of the H_o and decisions taken on its tenability.

	Decision	
Actual State	Accept Ho ($\underline{X} \in W$)	Reject Ho ($\underline{X} \in \omega$)
Ho is true	correct decision	Type I error
Ho is false	Type II error	correct decision

Since errors are involved in this statistical testing procedure and are unavoidable because of sample information in place of complete information, the testing procedure fixes the maximum probability of type I error at some 1000. % level. Choice of a , as a matter of convention, is 0.05 or 0.10,or 0.01. The a is known as the level of signifi-cance of the test. It is also referred to the size of the critical region.

The power of the test is determined by

{1- Pr. (Type II error)}

and the objective of the decision maker is to maximize the power subject to size restriction or a few other restrictions.

9.4.1. Small sample tests under normal setup

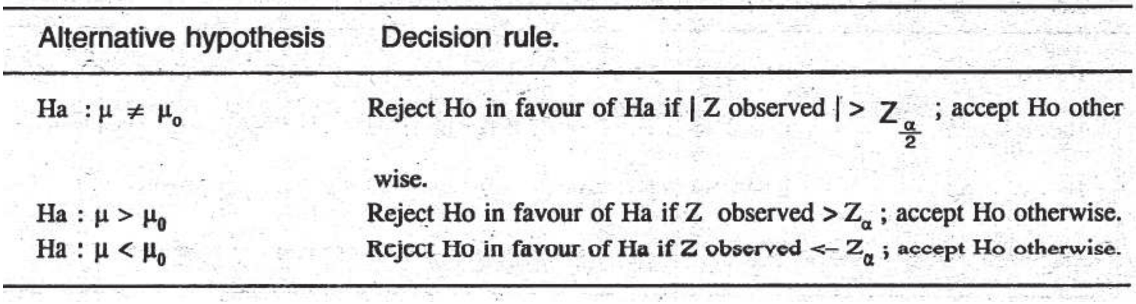
Let us consider $X_1, X_2......, X_n$ as n random sampie observations from a normal population denoted by N (σ, μ^2) with pdf

$$f_{\mu,\sigma^2}^{(x)} = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{1}{2\sigma^2}(x-\mu)^2}$$

Our objective will be to test for location μ and for dispersion σ^2 under various parametric situations.

Test for location when dispersion is known. In this situation we need to test Ho; $\mu = \mu_{0.}$ against different alternative hypotheses. The test statistic is $Z=\frac{\bar{X}-\mu_o}{\sigma/\sqrt{n}}$ where $\bar{X}$ is the sample mean. Under Ho, Zα follows the standard normal CI distribution N(o, 1).

Decision rules for different alternative hypotheses are described below at 100α level of significance with *Zα,* denoting the upper 100α% point of the normal deviate.



Alternative hypothesis	Decision rule.
Ha : $\mu \neq \mu_o$	Reject Ho in favour of Ha if | Z observed | > $Z_{\frac{\alpha}{2}}$; accept Ho other wise.
Ha : $\mu > \mu_0$	Reject Ho in favour of Ha if Z observed > Z_α ; accept Ho otherwise.
Ha : $\mu < \mu_0$	Rcject Ho in favour of Ha if Z observed <– Z_α ; accept Ho otherwise.

For $\alpha = 0.05$, $Z/\alpha_2 = 1.96$, $Z_\alpha = 1645$

Test for location when dispersion is unknown. Here Ho : $\mu=\mu_o$. The test statistic is given by

$$t = \frac{\bar{X}-\mu_o}{s/\sqrt{n}}$$

where s is an estimator of the unknown o value with the expression for s given by

$$\overline{X}_2 = \frac{\overline{X}_i - \overline{X}2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Under Ho, t follows student's t-distribution with degrees of freedom (n-1).

Decision rules for different alternative hypotheses are described below at 100α% level of significanc with $t_{\alpha\, n-1}$ denoting the upper 100α% point of thedistribution with degrees of freedom (n-1).

Alternative hypothesis	Decision rule.
Ha : $\mu \neq \mu_2$	Reject Ho in favour of Ha if | Z observed | > $Z_{\alpha\backslash 2}$; accept Ho otherwise.
Ha : $\mu > \mu_2$	Reject Ho in favour of Ha if Z observed > Z_α ; accept Ho otherwise.
Ha : $\mu < \mu_2$	Reject Ho in favour of Ha if Z observed <$-Z_\alpha$; accept Ho otherwise.

Test for dispersion : Here Ho : $\sigma = \sigma_0$, where σ_0, is a specified value of σ. The test statistic is

$$\chi^2 = \sum_{i=1}^{n}\left(X_1 - \bar{X}\right)^2 / \sigma_o^2 = (n-1)s^2 \ / \sigma_o^2$$

The distribution of this statistic under Ho is a chisquare distribution with degrees of freedom (n-1).

The decision rules under dfferent alternative hypotheses are given below with 100α% level of significance.

Alternative hypothesis	Decision rule.
Ha : $\sigma > \sigma_0$	Reject Ho in favour of Ha if χ^2 observed> $\chi^2_{\alpha;\, n-1}$; accept Ho otherwise.
Ha : $\sigma < \sigma_0$	Reject Ho in favour of Ha if χ^2 observed < $\chi^2_{1-\alpha,\, n-1}$: accept Ho otherwise.
Ha : $\sigma \neq \sigma_0$	Reject Ho in favour of Ha if χ^2 observed <$\chi^2_{\alpha/2}, _{n-1}$ or X^2 observed > $\chi^2_{\frac{\alpha}{2}, n-1}$; accept Ho otherwise.

Here $\chi^2_{\alpha n-1}$ represents the upper 100α% point of the chisquare distribution with (n-1) degrees of freedom. In case $\mu = \mu_o$ is known, the test statistic wil be

$$\chi^2 = \sum \frac{(x_i - \mu_0)^2}{\sigma_0^2}.$$

The decision rule will be same except for a change in degrees freedom as n in place of (n-1).

9.4.2. Two-population small sample tests under normal set up

Let us consider two normal populations X_1 and X_2, X_1 following N (μ_1, σ_1^2) and X_2, following $N(\mu_2, \sigma_2^2)$. Based on n_1, random sample observations X_{11}, X_{12}....., X_{1n_1} let us calculate sample mean $\bar{X}_1$ and sample variance s_1^2, where

$$\bar{X}_1 = \frac{1}{n_1}\sum_{\alpha=1}^{n_1} X_{1\varepsilon}, s_1^2 = \frac{1}{n_1 - 1}\sum_{\alpha=1}^{n_1}\left(X_{1\alpha} - \bar{X}_1\right)^2$$

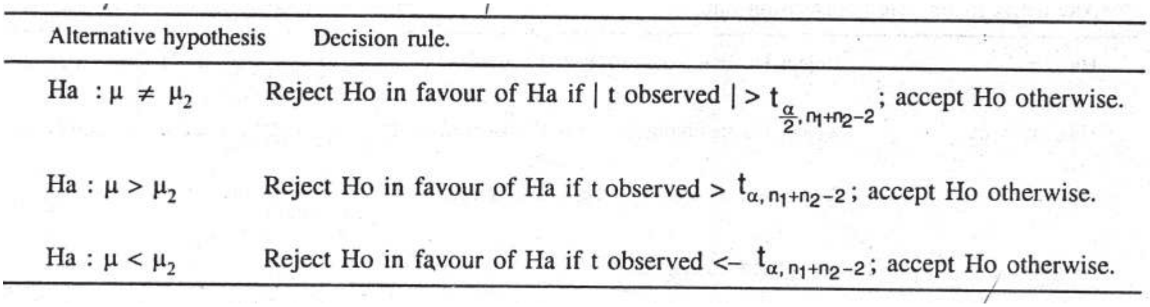
Similarly, based on n_2, random sample observations X_{21}, X_{22},....X_{2n_2} let us calculate sample mean $\bar{X}_2$ and sample variance s_2^2, where

$$\bar{X}_2 = \frac{1}{n_2}\sum_{\alpha=1}^{n_2} X_{2\alpha}, s_2^2 = \frac{1}{n_1 - 1}\sum_{\alpha=1}^{n_2}\left(X_{2\alpha} - \bar{X}_2\right)^2$$

Test for equality of two locations when dispestions are unknown. Let the null hypoth-esis to be tested be Ho : $\mu_1 = \mu_2$ where σ_1 and σ_2 are known.

The test statistic is $$Z = \frac{\bar{X}_1 - \bar{X}2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n2}}}$$

Under Ho, Z follows the standard normal distribution N(o, 1). The decision rules under different alternative hypotheses are described below for 100α% level of significanc;e.

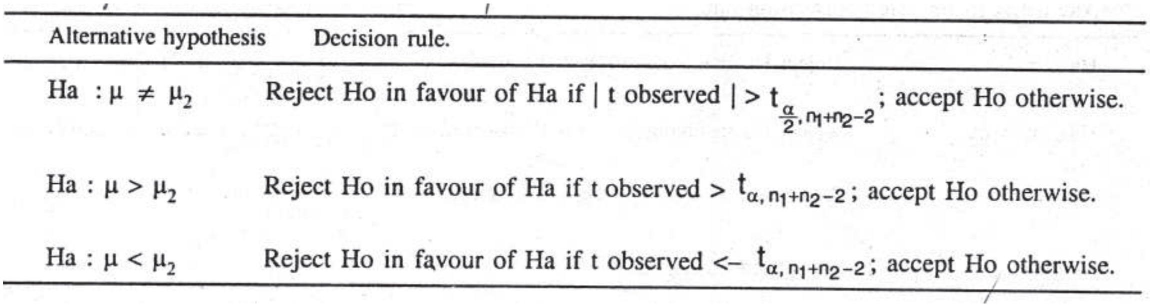


Alternative hypothesis	Decision rule.
Ha : $\mu \neq \mu_2$	Reject Ho in favour of Ha if | t observed | > $t_{\frac{\alpha}{2}, n_1+n_2-2}$; accept Ho otherwise.
Ha : $\mu > \mu_2$	Reject Ho in favour of Ha if t observed > t_{α, n_1+n_2-2}; accept Ho otherwise.
Ha : $\mu < \mu_2$	Reject Ho in favour of Ha if t observed <– t_{α, n_1+n_2-2}; accept Ho otherwise.

Test for equality of two locations when dispersions are unknown but equal. As in the previous case Ho to be tested is Ho H,=,. The statistic is

$$t = \frac{\bar{X}_1 - \bar{X}2}{\sqrt{\frac{1}{n_1} + \frac{1}{n2}}\sqrt{\frac{(n_1 - 1)\, s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}}}$$

This test statistic follows students' t-distribution with degrees of freedom (n_1+ n_2-2) under Ho. The decision rules are given below for different alternative hypotheses with 100α% level of significance.



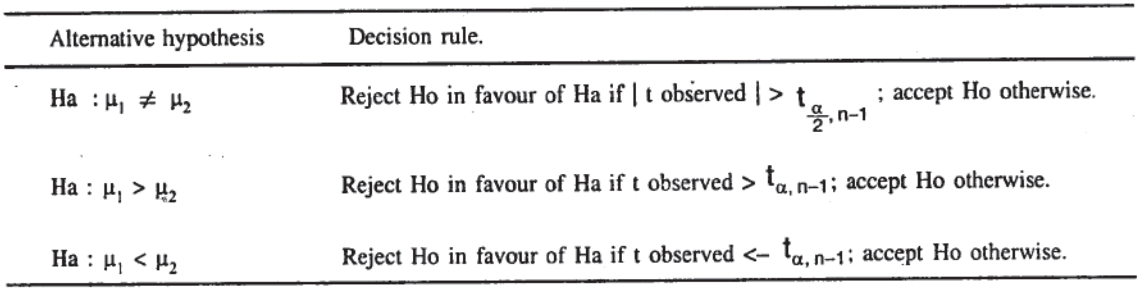
Alternative hypothesis	Decision rule.
Ha : $\mu \neq \mu_2$	Reject Ho in favour of Ha if | t observed | > $t_{\frac{\alpha}{2}, n_1+n_2-2}$; accept Ho otherwise.
Ha : $\mu > \mu_2$	Reject Ho in favour of Ha if t observed > t_{α, n_1+n_2-2} ; accept Ho otherwise.
Ha : $\mu < \mu_2$	Reject Ho in favour of Ha if t observed <– t_{α, n_1+n_2-2} ; accept Ho otherwise.

In case dispersions σ_1^2 and σ_2^2 are unknown but unequal, we may truncate two-sets of observations to minimum of n_1 and n_2. Consider the difference of observatios in a given order and carry out test similar to single population test for location.

Paired t test. In case in the problem of testing for equality of locations, the populations are interdependent we have pair-wise observations $(X_{1\sigma}, X_{2\sigma})$, = σ=1,2,........., n.

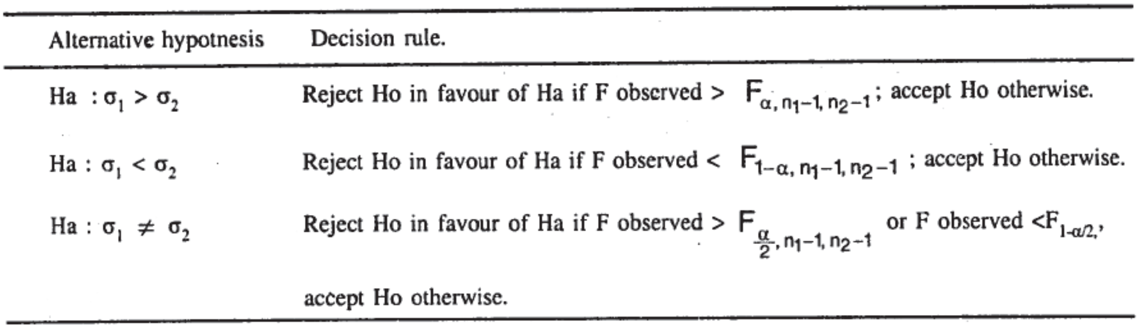
In such a case, writing $d_\alpha = X_{1\alpha} - X_{2\alpha}$, α=1, 2,...n, we construct a new r-andom variable d following normal distribution N (u,- ,, var (d)). The corresponding n random sample observations will be given by $d_1, d_2, ..., d_n$. The problem of testing Ho : $\mu_1 = \mu_2$ can be expressed in terms of expectation of d as Ho : E(d) =0. The test statistic is $t = \frac{\bar{d}}{s_d / \sqrt{n}}$.

which will follow the distribution with degrees of freedom (n-1) under Ho. The decision ,. rules under different alternative hypotheses are given below at 100α% level of significance.



Alternative hypothesis	Decision rule.
Ha : $\mu_1 \neq \mu_2$	Reject Ho in favour of Ha if | t observed | > $t_{\frac{\alpha}{2}, n-1}$; accept Ho otherwise.
Ha : $\mu_1 > \mu_2$	Reject Ho in favour of Ha if t observed > $t_{\alpha, n-1}$; accept Ho otherwise.
Ha : $\mu_1 < \mu_2$	Reject Ho in favour of Ha if t observed <– $t_{\alpha, n-1}$; accept Ho otherwise.

Test for equality of dispersion. We need to test for $\sigma_1 = \alpha_2$. The test statistic is F = s_2^1 / s_2^2. which follows the F distribution under Ho with depress of freedom (n_1-1) and (n_2-1) respectively. The decision rules for the alternative hypotheses of interest are described below at 1000% level of significance.



Alternative hypotnesis	Decision rule.
Ha : $\sigma_1 > \sigma_2$	Reject Ho in favour of Ha if F observed > F_{α, n_1-1, n_2-1}; accept Ho otherwise.
Ha : $\sigma_1 < \sigma_2$	Reject Ho in favour of Ha if F observed < $F_{1-\alpha, n_1-1, n_2-1}$; accept Ho otherwise.
Ha : $\sigma_1 \neq \sigma_2$	Reject Ho in favour of Ha if F observed > $F_{\frac{\alpha}{2}, n_1-1, n_2-1}$ or F observed <$F_{1-\alpha/2,}$, accept Ho otherwise.

Here $F_{\alpha, n_1\text{-}1, n_2\text{-}1}$ is the upper 100α% point of F distribution with d.f. (n_1-1) and (n_2-1).

Test for independence. In the bivariate normal set up lack of correlation is equivalent to independence and hence a test for independence may be carried out by testing the null hypotnesis Ho : $\rho = 0$ where ρ is the population correlation coefficient. Random sample observations ($X_{1\alpha}$, X=$_{2\alpha}$) α=1,2, ...n from the bivariate normal population can be utilized to calculate the sample correlation coefficient, r, where

$$r = \frac{\sum_{\alpha=1}^{n}\left(X_{1\alpha} - \bar{X}_1\right)\left(X_{2\alpha} - \bar{X}_2\right)}{\sqrt{\sum_{\alpha=1}^{n}\left(X_{1\alpha} - \bar{X}_1\right)\sum_{\alpha=1}^{n}\left(X_{2\alpha} - \bar{X}2\right)^2}}$$

The test statistic is

$$t = \left(r\sqrt{n-2}\right) / \sqrt{1-r^2}.$$

which follows the t distribution under Ho with degrees of freedom (n-2).

Since we are tnterested in testing for independence, the alternative hypotnesis will deal with presense of dependence. Thus, Ha : $\rho \neq o$. The decision rule will be to reject Ho in favour of Ha if l t observed l $> t_{\frac{\alpha}{2}, n-2}$,

9.4.3. Some large sample tests

The small sample tests covered so far are based on normal set up. Quite often we come across with situtions where the set up is non-normal. Most of these problems can be handled through large sample normal approximation. It has been noted under the celebrated central-limit-theorem that if $X_1, X_2, ..., X_n$ are n random sample of observations from any population. with mean μ and finite variance α^2 then follows asymptotically normal distribution in the sense that the cumulative distribution function of T can be approximated by the cumulative distribution function of a standard normal distribution when → ∞. We write ‘

T – AN (0, 1).

We shall make use of this result for suggesting.Jarge sample test for testing hypothesis related to non-normal populations mainly.

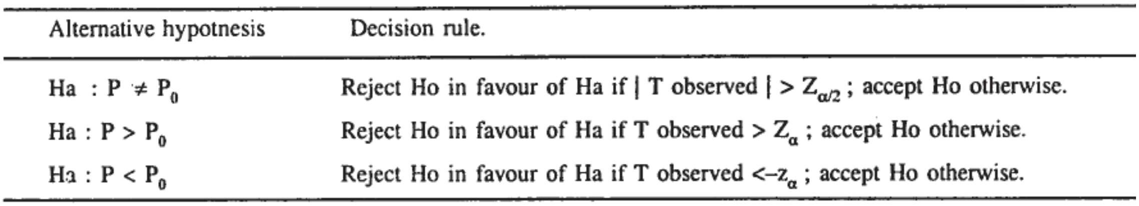
Large sample Test for proportion: Let the random variable X follow binomid distribution with pmf.

$$P(x) = \binom{n}{x} p_x (1-p)^{n-x}$$

and we need to test Ho : P = P_o. Viewing P as the mean of (X/n) we consider the sample proportion $\hat{p}$ = X/n as the sample mean. For large n,

$$T = \frac{\hat{p} - p_o}{\sqrt{p_o(1-p_o)/n}}$$

follows asymptotically normal distribution N(0, 1), under Ho. Hence the large sample decision rules for different alternative hypotheses can be developed along the following lines for 100α% level of significance. ·



Alternative hypotnesis	Decision rule.
Ha : P ≠ P_0	Reject Ho in favour of Ha if | T observed | > $Z_{\alpha/2}$; accept Ho otherwise.
Ha : P > P_0	Reject Ho in favour of Ha if T observed > Z_α ; accept Ho otherwise.
Ha : P < P_0	Reject Ho in favour of Ha if T observed <–z_α ; accept Ho otherwise.

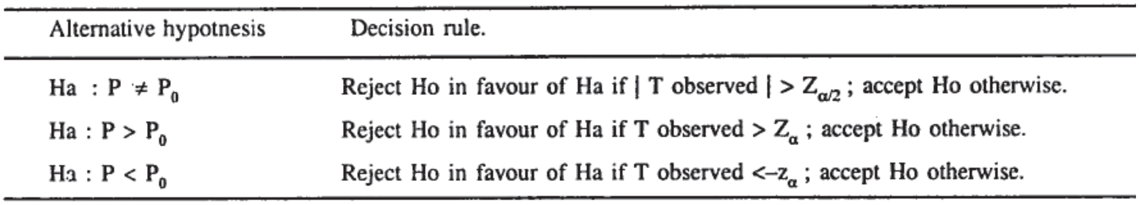
Large Sample Test for equality of two proportions: For two independent binominial populations denoted by pmfs

$$p_1(x_1) = \binom{n_1}{x_1} p_1^{x_1} (1-p_1)^{n_1-x} \text{ and } p_2(x_2) = \binom{n_2}{x_2} p_2^{x_2} (1-p_2)^{n_2-s_2}$$

the population proportions are P_1 and P_2 and sample proportions are $\hat{P}_1 = x_1/n_1$ and $\hat{P}_2 = x_2/n_2$. The null hypothesis to be tested is Ho : P_1 = P_2. Writing $\hat{P} = (x_1 + x_2)/(n_1) + n_2)$ the pooled estimator of population proportion when P_1 = P_2. We may suggest the test statistic as

$$T = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} . \sqrt{\hat{p}(1-\hat{p})}}$$

which follows asymptotically normal distribution N (0, 1) under Ho. The decision rules for different alternative hypotheses are given below for 100α% level of significance.



Alternative hypotnesis	Decision rule.
Ha : $P \neq P_0$	Reject Ho in favour of Ha if $\mid$ T observed $\mid$ > $Z_{\alpha/2}$; accept Ho otherwise.
Ha : $P > P_0$	Reject Ho in favour of Ha if T observed > Z_α ; accept Ho otherwise.
Ha : $P < P_0$	Reject Ho in favour of Ha if T observed <$-z_\alpha$; accept Ho otherwise.

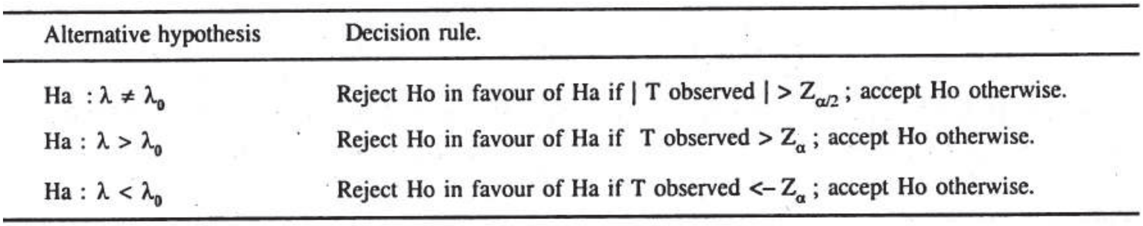
Large sample test for Poisson mean. Let X_1, X_2....X_n be n random sample observations from a Poisson distribution with pmf.

$$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

To test for Ho : $\lambda = \lambda_0$ we consider the sample mean $\bar{X}$ and its asymptotic distribution. The test statistic is

$$T = \frac{\sqrt{n(X - \lambda_0)}}{\sqrt{\lambda_0}}$$

which follows asymptotically normal distribution N (0,1) under Ho. Then the decision rules for different alternative hypotheses will be as follows for 100α% level of significance.



Alternative hypothesis	Decision rule.
Ha : $\lambda \neq \lambda_0$	Reject Ho in favour of Ha if $\mid$ T observed $\mid$ > $Z_{\alpha/2}$; accept Ho otherwise.
Ha : $\lambda > \lambda_0$	Reject Ho in favour of Ha if T observed > Z_α ; accept Ho otherwise.
Ha : $\lambda < \lambda_0$	Reject Ho in favour of Ha if T observed <$-$ Z_α ; accept Ho otherwise.

Large sample test for correlation, let ρ be the population correlation coefficient and let r be the sample correlation coefficient obtained from a pair-wise data of size n. Then, to test Ho : $\rho = \rho_0$ we consider the following statistic.

$$T=\sqrt{n-3\left[\frac{1}{2}\log_c\frac{1+r}{1-r}-\frac{1}{2}\log_e\frac{1+\rho_o}{1-\rho_o}\right]}$$

The asymptotic distribution of which is N (0, 1) under Ho. The decision rules for differemt a alternative hypotheses ae described below for level of significance equal to α.

Alternative hypothesis	Decision rule.
Ha : $\rho \neq \rho_0$	Reject Ho in favour of Ha if $\lvert$ T observed $\rvert$ > $Z_{\alpha/2}$; accept Ho otherwise.
Ha : $\rho > \rho_0$	Reject Ho in favour of Ha if T observed > Z_α ; accept Ho otherwise.
Ha : $\rho < \rho_0$	Reject Ho in favour of Ha if T observed < $-Z_\alpha$; accept Ho otherwise.

A few other large sample tests. As may be noted from the earlier discussions that the decision rule is very similar in nature for all the suggested large sample tests. For the sake of brevity, we present below a few statistics. which may be of use for large sample tests.

- To test for means, μ, we may use statistic

$$T=\frac{\sqrt{n}\left(\bar{X}-\mu\right)}{s}$$

where $\bar{X}$ is the sample mean and s is the sample standard deviation. T has asymptotic normal distribution N (0,1).

To test for variance, σ^2, we may use statistic

$$T=\frac{s^2-\sigma^2}{\sigma^2\sqrt{2/n}}$$

where s^2 is the sample variance. T has asymptotic standard normal distribution.

* To test for normality iri terms of $\beta_1 = 0$ and $\beta_2 = 3$ we may use statistics:

$$T_{124}=g_1\sqrt{\frac{n}{6}\text{ and }T_2}=g_2\sqrt{\frac{n}{24}}$$

where g_1 is the sample measure of β_1, the skewness, and g_2, is the sample measure of (β_2-3) for peakedness. T_1, T_2 are distributed as asymptotic normal deviates.

9.6. Test for goodness of fit and independence.

Pearsonian χ^2 test is used to examine whether the given frequency table is in good agreement with a hypothetical distribution. Let there be k classes with observed fre-quency for the i-th class as f_{0i}. Let f_{ei} be the corresponding expected frequency as per the hypothetical distribution.

The Pearsonian χ^2 for goodness of fit is given by

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{oi} - f_{ei})^2}{f_{ei}}$$

If χ^2 observed $<\chi^2_{\alpha,\ k-1}$ then accept the null hypothesis Ho : Frequency table is in agreement with the hypothetical distribution with level of significane α.

This concept of Pearsonian x^2 can be of help in examining the independence of two attributes. Let Ho : Attributes are independent to be tested against the alternative Ha : Attributes are associated. Writing f_{ij} as the frequency of observations belonging i-th form of attribute A and j-th form of attribute B, i = 1,2, ... $_1$m, j = 1, 2, ... n, we may express the Pearsonian ^{2} as

$$x^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\left(f_{ij} - f_i.f._i / N\right)^2}{\frac{f_i.f._j}{N}}$$

$$\text{where } f_1. = \sum_{j=1}^{n} f_{ij} \text{ m } f._j = \sum_{i=1}^{n} f_{ij} \text{ and } N = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ji}$$

If the χ^2 observed $<\chi^2_\alpha$, (m-1) (n-1) then the Ho is accepted at 100α% level of significance. Otherwise Ho is rejected in favour of Ha.

9.6. Non-parametric tests

Non-parametric testing procedure makes general assumption concerning the distri-bution function and does not prefix the form of the distribution function. Thus, assumptions like normality, homoscedasticity etc. are not made to test the hypothesis of interest. Often variables are only assumped to be coming from a continuous distribution. Ordered observations are of frequent use and median is the more popular measure of central tendency.

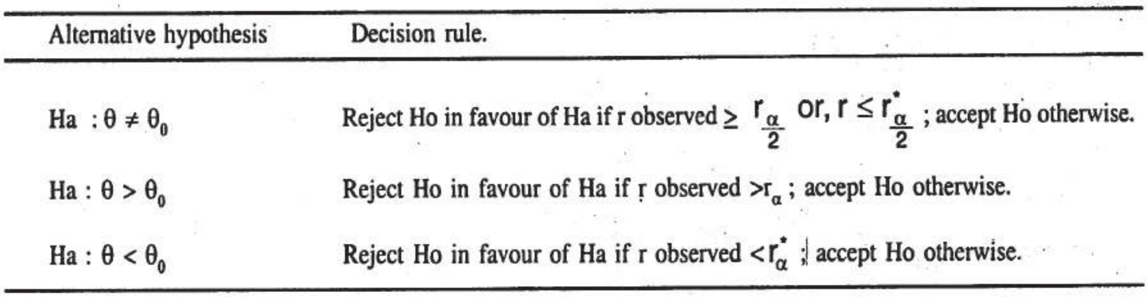
lnter-quartile range is the suggested measure of dispersion. Sample median is the estimator of the population median and sample interef uartile ram is the estimator of the population inter-quartile range. Let us describe a few non-parametric tests which are widely used by' the researchers.

Nor-paramatric test for location. Let the location be measured by median and deno θ. Then, by definition, $P[X \leq \theta] = P[X > \theta] = \frac{1}{2}$

Let X_1, X_y....., X_n be n random sample observations and based on these we want· to test Ho : θ= θ_0. For this, let us examine the sign of $\{X\text{-}\theta_0\}$, i = 1,2 ... , n. Let r be the number of plus signs. Under Ho, r will follow binominal distribution with probability of success 1/2 and probability of failure 1/2, as half of the signs are expected to be positive and remaining half are expected to B negative. The decision rules for different alternative hypotheses are given below in terms of critical values $r_{\frac{\alpha}{2}}$ and $r^{\cdot}_{\frac{\alpha}{2}}$ where $r_{\alpha/2}$ and $r^{\cdot}_{\frac{\alpha}{2}}$ are defined by the following inequalities.

$$\sum_{x=r_{\frac{\alpha}{2}}}^{n} \binom{n}{x}\left(\frac{1}{2}\right)^{n} \leq \alpha/2, \sum_{x=0}^{r^{\cdot}_{\frac{\alpha}{2}}} \binom{n}{x}\left(\frac{1}{2}\right)^{n} \leq \alpha/2$$

and r_α and $r^{\cdot}_\alpha$ are defined accordingly.



Alternative hypothesis	Decision rule.
Ha : $\theta \neq \theta_0$	Reject Ho in favour of Ha if r observed ≥ $r_{\frac{\alpha}{2}}$ or, $r \leq r^{*}_{\frac{\alpha}{2}}$; accept Ho otherwise.
Ha : $\theta > \theta_0$	Reject Ho in favour of Ha if r observed $>r_\alpha$; accept Ho otherwise.
Ha : $\theta < \theta_0$	Reject Ho in favour of Ha if r observed $<r^{*}_\alpha$; accept Ho otherwise.

This test for location is known as sign test. It can be extended to paired sample data $(X_i, Y_i), i = 1, 2, \ldots., n.$ The difference values d_i= X_i - Y_i can be constructed and sign test may be applied on d_i's to test that this difference has the median at θ=θ_0.

Cox and Stuart test for presence. of trend. Let x_1, X_2, ...,X_n, be n chronological observations. Let c = n/2 it n is even or c = $\frac{(n+1)}{2}$ if n is odd. Then the observations can be grouped in pairs as

$$(X_1, X_{C+1}), (X_2, X_{c+2}), ..., (X_{n-c}, X_{n)}.$$

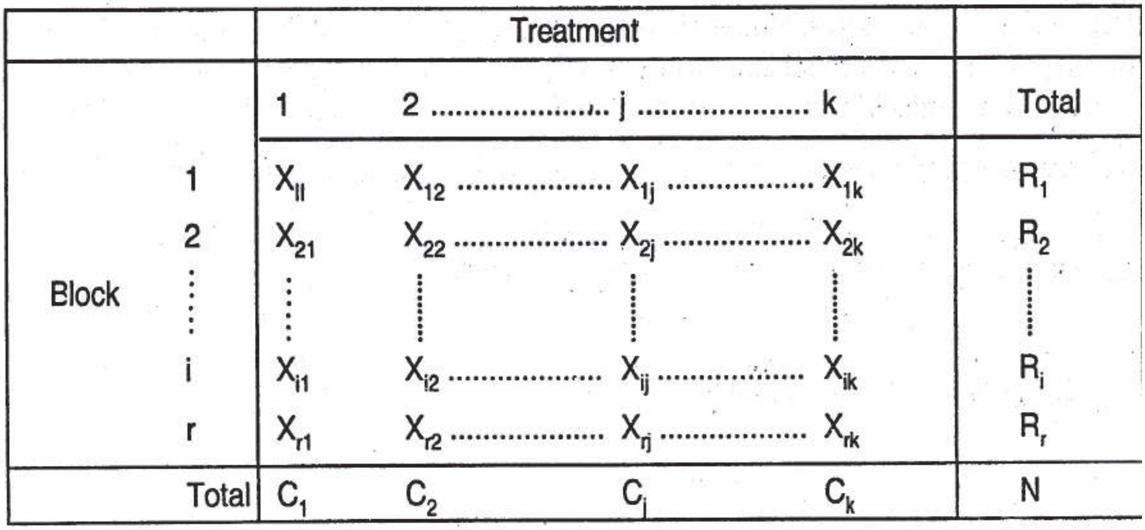
In this process we ignore the middle most number if n is odd. Next, assign + sign if $X_i > X_i > X_{C+i}$ and assign-sign if $X_i < X_{CH}$ and ignore ties, if any. Let r be the number of + signs.

The problem for testing for trend can be studied in terms of Ho : There is no trend in the data i.e.

Ho: $P[X_i<X_{c+1}]=p[X_i>X_{c+i})$

The test procedure based on r will be similar to the sign test as explained earlier with n replaced by n-c.

Cochran test for equality of treatment effects: Suppose k treatments are applied on r blocks with treatment effects measured in terms of binary observations 1 and o, 1 if treatment is effective and o if treatment is in-effective. The scheme of data presentation is given below :



		Treatment				
		1	2	j	k	Total
	1	X_{ll}	X_{12}	X_{1j}	X_{1k}	R_1
	2	X_{21}	X_{22}	X_{2j}	X_{2k}	R_2
Block	⋮	⋮	⋮	⋮	⋮	⋮
	i	X_{i1}	X_{i2}	X_{ij}	X_{ik}	R_i
	r	X_{r1}	X_{r2}	X_{rj}	X_{rk}	R_r
	Total	C_1	C_2	C_j	C_k	N

$$X_{ij} = \begin{cases} 1 & \text{if treatment is effective} \\ o & \text{otherwise} \end{cases}$$

We need to test Ho : Treatment effects are equal against Ha : Treatment effects are not equal. The test statistic suggested by Cochran is

$$T = [k\ (k-1) \sum_{j=i}^{k} \left(C_j - N/k^2\right)^2]/[\sum_{i=1}^{F} R_1\,(k-R_1)]$$

The decision rule will be to reject Ho at 100α% level of significance if

T observed $> \chi^2_{\alpha.\kappa-1}$.

Mann Whitney test for identity of two distribution fufctions. Let X and Y be two random variables with distribution functions F_x (u) and F_y (u). We like to test Ho : F_x (u) = F_x (u) for all u against the alternative that F_x (u) = F_y (u) are not idential for some u. To carry out this test we need to collect random sample obserbations X_1, X_y,..., X_n on x and Y_1, Y_2,...Y_m on Y.

The test procedure involves pooling of (n+m) observations and assignment of ranks to (n + m) observations. Let R (X_i) and R (Y_j) be the rank of X_i of population X and rank of Y_i of population Y respectively. Then the test statistic is

$$T = s - \frac{n(n+1)}{2}$$

where S is the sum of the ranks assigned to observations from population X i.e. S = $\sum_{i=1}^{n} R\,(X_i)$. The decision rule will be to reject Ho in favour of the alternative if T < $w_{\alpha/2}$ or > $w_{1-\alpha/2}$ and accept Ho otherwise, where α is the level of significance and

w_α is the α–th quantile of the Mann-Whitney statistic.

Kruskal Wallis test. This is an extension of Mann-whitney test when number of populations to be studied for identity is k($\geq$3). The test procedure is to collect n, random , sample observations from the i-th population and then rank all the observations in the G pooled sample. Let S_i be the sum of the ranks of the observations belonging to population i, i = 1,2, ... k. Then the appropriate test statistic for testing the identity of the k population distribution functions is

$$T = \frac{12}{N(N+1)} + \sum_{i=1}^{k} \frac{S_i^2}{n_i} - 3(N+1)$$

k where N = $\sum_{i=1}^{k} n_1$

The decision rule is described below :

If k = 3 and n_i < 5 for all i = 1, 2, ..., k reject Ho if the observed value of T is greater than or equal to H_α, the critical value of Kruskal-Wallis test for level of significance α.

If k = 3 or n_i > 5 for all i, then T approximately follows x^2 distribution with (k-1) degrees of freedom under Ho. Then reject Ho if T observed $>_{\chi 2\ \alpha,\ k-1}$.

Kolmogorov-Smirnov test .for goodness of fit. Let the population distribution function be given by F(x) and the hypothesis to be tested is

Ho: F (x)= F_0(x) $\forall \chi$

against the alternative that F(x) is not equal to F_0, (x) for some values of x. With n random sample observations X_1, X_2,...X_n drawn from F(x) one can estimate F(x) by the empirical cdf F_n (x) where

$$F_n\,(x) = \frac{1}{n}\,[\text{no. of observtion} \leq x\,]$$

The test statistic is

$$T = \sup_x\,|F_n\,(x) - F_0\,(x)|$$

and the decision rule at 100α% level of significance is to reject Ho in favour of. the alternative if Tobserved exceeds (1 - α) quartile of Kolmogorov-Smirnov statistic.

9.7. Summary

The basic objective of data collection is to estimate or test for the population characteristics of importance. Estimation of the population characteristics is of two types, according to the nature of the population: finite population or infinite populatin. In case there is a conjecture or hypothesis available with the investigator that hypothesis may be tested in the light of the given data: If probability model is given the hypothesis may be expressed in terms of the population parameters. This type of testing is known as parametric testing. In case the probability model is riot given or assumed the testing problem comes under the domain of non-parametric test. Test procedures also differ according to the size of the sample.

Sample mean is the Sual estimator of the population mean and sample proportion is the usual estimator of the population proportiotin case of simple random sampling without replacement. These are unbiased esimators. Corresponding standard errors are avail-able in explicit form. Sample mean is also an unbiased estimator for the cases of simple random sampling with replacement and systematic sampling. For stratified random sampling weighted sample means of strata is used to unbiasedly estimate the *-d* / ... population mean. Similar is the estimator for two stage sampling plan.

For parametric estimation from an infinite population it is observed that sample mean is the Best Linear Unbiased Estimator (BLUE) for the population mean. Sample mean is the Minimum Variance Unbiased Estimator (MVUE) of the population mean for Poisson, normal and exponential distributions. Sample · mean and sample variance (with divison) are the maximum likelihood estimators of the population mean and population variance of the normal population. The same estimators can be obtained by the method of moments.

Parametric tests have been described under normal setup for single population test for location when dispersion is given and when dispersion is not given. These and the subsequent test procedures are described through critical region. For normal population test for dispersion has been studied when the mean is unknown and when the mean is known.

In case of two normal populations tests for equality of means have been studied under independent and dependent setups covering the cases where variances are known and variances are unknown but equal. Also test for equality of variances has been suggested in terms of F test and test for independence has been suggested in terms of sample corelation coefficient leading to t test.

Large sample tests for proportion and equality of two proportions for binomial setup, test for mean for Poisson setup, test for lack of correlation etc. have been covered.

Tests for goodness of fit and test for independence of attributes have been pre-sented in terms of Pearsonian x^2. Also non-parametric sign test for location, Cox and Stuart test for presence of trend, Cochran's test for equality of treatment effect, Mann-Whitney test for identity of two populations, Kruskal-Wallis test for identity of multiple populations and Kolmogorov-Smirnov test for goodness of fit have been described.

9.8. Questions

1. Suggest an unbiased estimator for the population mean based on SRSWOR explaining the concept of unbiasedness and indicating the estimation procedure for estimating the standard error of the proposed estimator for mean.
2. Using the concept of expectation show that sample mean based on systematic sampling is an unbiased estimator of the population mean. Obtain the expression for standard error in this case.
3. Using the expression for var ($\bar{\bar{X}}$) given under stratified random sampling, examine

the allocation of stratum-wise simple size when var ($\overline{\overline{X}}$) is minimized subject to. restriction that n = $\sum_{i=1}^{k} n_i$. (Ret. 9.2.4.).

4. Prove that sample mean is BL,UE of the population mean.
5. Describe Poisson distribution, exponential distribution and normal distribution and suggest MVUE of their population means.
6. What do you mean by Maximum Likelihood method ? State one important property of mle. Also obtain mle of μ and σ^2 of a normal distribution where is the population mean and σ^2 is the population variance.
7. Describe the statistical testing procedure indicating the errors and their measures.
8. Life of a brand of electric bulb has been measured for 18 prototypes. Assuming life distribution to be normal, test for mean life as 100 hrs using. both small sample test and large sample test. .

   Life (in hrs) : 125, 85, 87, 101, 93, 72, 73, 82,
   91, 91, 93, 81, 89, 78, 82, 67, 78, 99.
9. For the data set given in question no. 8 examine whether the standard deviation is equal to 5 or less.
10. For two normal populations **N** (μ_1, 144) and N ((μ_2, 169) the sample means based in 12 observations each are found to be 105 and 127. Examine the tenability of Ho : $\mu_1 = \mu_2$ against a suitable alternative.
11. For a bivaviate normal ‘distribution sample correlation coefficient based on 20 pairs of observations has been calculated as 0.21. Examine whether the two variables are independent.
12. Suggest a suitable non-parametric test for the problem in question number 8 when normality assumption is questionable and median life. is _same as inean life.
13. For the problem givenunder question number 11 test for Ho : $\rho = 0.25$, where ρ is the population correlation coefficient against the alternative Ha : $\rho < 0.25$.
14. To fit a uniform distribution given by the distribution function F(x) = (x-60)/70, $60 \leq x \leq 130$ to the life data given under question number 8, examine the goodness of fit.
15. Test fr normality of the data set given under question number 8.
16. Describe Kruskal Wallis test clearly. stating the null hypothesis to be tested.

    Suggest a test procedure for two-population case.

17. Apply Cox and Stuart test for presence of trend in the following data set arranged in a chronological way :

    212, 215, 220, 223,230, 237, 242, 250, 255, 259.

Short answer type questions.

1. Explain the terms 'parametric test' and 'no-parametric test'.
2. Explain the basic principle of large sample test.
3. What do you mean by BLUE? Suggest a BLUE for the population proportion.
4. Is mle necessarily unbiased ? If yes, give reasons. If no, give an example of a biased estimator.
5. Explain the concept 'error' in statistical testing of hypotnesis .
6. Explain through an example how decision rule changes with the change in the alternative hypothesis.
7. Is Pearsonian x^2 test for goodness of fit a non-parametric test? Give resons .
8. Will the Cox and Stuart test for presence of trend get changed if a constant is added to all the observations? Give arguments.
9. Suggest a suitable test procedure for examining the presence of trend.
10. Describe Cochran's test for equality of multiple mean treatment effects.

Objective type questions.

Indicate· whether the following statements are true or false.

1. Parametric testing is creation of observation based on parametric information.

   True □ False □

2. Sample mean is an unbiased estimator of population mean ,...

   True □ False □

3. Standard error is the positive square root of variance of the estimator under study

   True □ False □

4. In systematic sampling standard error can not be expressed.

   True □ False □

5. Simple average of the strata means is an unbiased estimator of the population mean when strata are of equal size.

True □ False □

6. BLUE is the MVUE of the population mean when the sample is drawn from a normal population.

True □ False □

7. cr^2 is an unbiased estimator of the variance of a population following normal distribution.

True □ False □

8. Appropriate test statistic for a test for normal mean when variance is unknown follows F distribution under Ho.

True □ Faso □

9. For a normal setup test for independence is equivalent to test for lack of correlation.

True □ False □

10. x^2 statistic for goodness of fit is also of use for testing independence of attributes.

True □ False □

11. Cochran's test is based on binary type information.

True □ False □

12. Empirical cdf is an estimator of the population cdf.

True □ False □

13. Kolmogorov Smirnov test statistic lies between o and 2

True □ False □

Unit - 10 □ Advanced Analytical Tools

Structure

10.1 Introduction

There are some analytical tools of advanced nature, which have been found to be extremely useful in drawing statistical inference in managerial problems. We propose to consider a few of those tools and techniques which will equip the researcher to a great extent. One such tool is Analysis of variance (ANOVA).

It is the major analytical tool for examining the data generated by the experimental design of formal type. We have examined the case where equality of two population means can be tested either by using z test (when variances are known) or by using test (when variances are eqal and unknown). In case we like to consider more than two populations and are interested to test for equality of multiple. mean values, ANOVA technique can provide with an interesting solution to this problem. Thus, ANOVA is having a greater scope than z or t tests. Because of its generic approach it has a wider appeal in the sense that it can be used for testing the suitability of regression analysis, effectiveness of treatment effects and interaction effects, effec-tiveness of classification system according to nature of variation/ and so on. ANOVA *d*/-has been extended to multivariate set up, giving rise to Multivafiate Analysis of Dis-persion (MANOD) also known as. Multivariate Anlysis of variance (MANOVA).

Regression is another widely used technique, mainly used for forecasting purpose. In the multivariate set up, one may have to deal with prediction situations when one dependent variable is affected by multiple independent variables. Mutiple regression analysis is an answer to such problems. The line of attack in multiple regression is similar to the line of attack in ordinary regression. The suitability of multiple regression as a causal model or as a predictor can be easily examined via ANOVA technique.

Some other multivariate tools like Factor analysis, Conjoint - analysis and cluster analysis are also of practical interest.

We propose to present analysis of. variances and multiple regression analysis with Worked out examples. Rest of the tools will. be described more as an invitation to multivariate analysis than detailed prescription.

10.2. Analysis of Variance (ANOVA)

The essence of ANOVA is to split, accorting to causes, the total variation in a set of data. The basic objective to study the nature of variation. Variation may be due to purely chance factor and variations may be due to different causes. In case we can separate out different types of causal variations from the chance variation, a better in sight can be given into the total variation in the data. If we can identify one cause of variation we may rearrange the data according to different levels of that cause. The resultant presentation of data is known as one-way classified data. If two causes can be identified each with multiple levels we may rearrange the data into two-way classified form.

Let us consider the ANOVA for one-way classified data. We may view the levels of the single cause present in the observed system, as different populations. In that case, one-way classified data represent data according to k populations where $k \geq 2$. If we assume the cause of variation to be described by the differential mean values and not by the variances of the population distributions, we may consider variances to be all equal. If there is no variation among the mean values, then the entire data set will represent a single population and let us describe the same by the null hypothesis, alternative hypothesis will be appreciating mean-wise variations among the k populations.

If we denote by X_{ij} the j-th observation on the i-th population, i =1, 2, ..., K, j = 1,2, ..., n_i, then the mathematical model for carrying out subsequant analysis can e presented as

$$x_{ij} = \mu + \beta_i + e_{ij}$$

i = 1, 2, ..., k, j = 1, 2, ..., n_i, where e_{ij} ‘s are independent and normally distributed variables with mean zero and variance σ^2 and

$$\sum_{i=1}^{k} n_i \beta_i = 0$$

We may refer μ as the general effect and β_i as the specific effect of i-th population / class, i = 1,2,..........,k.

The least square estimates of μ andβ_i can be obtained by minimizing sum of squares of error terms, i.e.

$$S = \sum_{i=1}^{k} \sum_{j=1}^{n_i} e_{ij}^2$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \mu - \beta_i \right)^2$$

After differentiating with respect to be and β_i we get the estimating normal euations as

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} = n\mu + \sum_{i=1}^{k} n_i \beta_i$$

and

$$\sum_{j} x_{ij} = n_i \mu + n_i \beta_i, \quad i = 1, 2, \ldots\ldots, k$$

where $n = \sum_{i=1}^{k} n_i$.

Thus, in view of $\sum n_i \beta_i = 0$ we can estimate μ by $\hat{\mu} = \left(\sum_i \sum_j x_{ij} / n \right)$ and β_i by $\hat{\beta}_i = \left(\sum_j x_{ij} / n_i \right) - \hat{\mu}$. Further the total sum of squares

$$TSS = \sum_{i} \sum_{j} \left(x_{ij} - \hat{\mu} \right)^2$$

can be espressed as

$$\sum_{i} \sum_{j} \left(x_{ij} - \hat{\mu} \right)^2 = \sum_{i} \sum_{j} \left(x_{ij} - \bar{x}_i + \bar{x}_i - \hat{\mu} \right)^2$$

$$= \sum_{i} \sum_{j} \left(x_{ij} - \bar{x}_i \right)^2 + \sum n_i (\bar{x}_i - \hat{\mu})^2$$

Writing sum of squares due to class effects as SSB and sum of squares due to error as SSE where

$$SSB = \sum n_i (\bar{x}_i - \hat{\mu})^2 = \sum n_i \left(\bar{x}_i - \bar{\bar{x}} \right)^2$$

$$SSE = \sum_{i} \sum_{j} \left(x_{ij} - \bar{x}_i \right)^2$$

$\bar{\bar{x}} = \hat{\mu}$, the grand mean,

we can finally express the TSS as sum of SSB and SSE. Thus the total variation can be split into two variations one due to class effects and the other due to chance causes :

$$TSS = SSB + SSE.$$

The degrees of freedom of TSS is (n–1) and those of SSB and SSE are (k–1) and $\sum(n_i - 1) = (n - k)$. Thus degrees of freedoms are also additive

$$(n-1) = (k-1) + (n-k).$$

Dividing sum of squares by the corresponding degrees of freedom we obtain mean squares due to class effect and mean squares due to error :

$$MSB = SSB \,/\, (k-1)$$
$$MSE = SSE \,/\, (n-k).$$

Then the hypothesis $H_o : \beta_1 = \beta_2 = \ldots\ldots = \beta_k$, i.e. the class effects are absent, can be tested in terms of the test statistic

$$F = \frac{MSB}{MSE}$$

which follows F distribution under H_o with degrees of freedom (k–1) and (n–k). The decision rule is to reject H_o in favour of the alternative that at least one β_i is different if $F_{observed} > F_{\alpha,\, k-1,\, n-k}$ with 100α% level of significance.

The entire procedure can be summarized in a tabular form, known as ANOVA table, as given below :

ANOVA Table (one-way classified data)

Sources of variation	d.f	sum of squares	mean squares	F ratio	
				observed	Tabulated
Between classes	k–1	$SSB = \sum n_i(\bar{x}_i - \bar{\bar{x}})^2$	$MSB = \frac{SSB}{n-1}$	F =	
Error	n–k	$SSE = \sum\sum(x_{ij} - \bar{x}_i)^2$	$MSE = \frac{SSE}{n-k}$	$\frac{MSB}{MSE}$	$F_{\alpha,\, k-1,\, n-k}$
Total	n–1	$TSS = \sum\sum(x_{ij} - \bar{\bar{x}})^2$			

In case the null hypothesis is rejected we may test for equality of class means in a pair-wise manner. Writing $H_{oij} : \beta_i = \beta_j$ we may carry out usual test for equality of means with the help of t statistic

$$t = \frac{\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j}{\sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \,.\, \sqrt{MSE}}$$

and decision rule of rejecting H_{oij} if $|\, t_{observed} \,|$ is greater than $t_{\frac{\alpha}{2},\, n-k,}$ α being the level of significance.

If $n_1 = n_2 = \ldots\ldots = n_k = n_0$ the pair-wise test can be conducted for each pair in terms of a common critical difference (CD) where

$$CD = t_{\frac{\alpha}{2},k(n_o-1)}\sqrt{\frac{2MSE}{n_o}}$$

and the decision rule is to reject H_{0ij} if $|\bar{x}_i - \bar{x}_j| > CD$, with α level of significance.

Let us explain the entire procedure in terms of the following example.

There are 3 brands of tyres and we need to test for equality of their mean lives.

Brands	**Observed lives (in 1000 miles)**
A	35, 34, 34, 33, 34
B	32, 32, 31, 28, 29
C	34, 33, 32, 32, 33

Note that $n_1 = n_2 = n_3 = n_o = 5$.

Grand mean $\bar{\bar{x}} = 32.4$

Brand means $\bar{x}_1 = 34$, $\bar{x}_2 = 30.4$, $\bar{x}_3 = 32.8$

$$TSS = \sum\sum\left(x_{ij} - \bar{\bar{x}}\right)^2 = 51.6$$

$$SSB = \sum n_i\left(\bar{x}_i - \bar{\bar{x}}\right)^2 = 33.6$$

SSE = TSS – SSB = 51.6 – 33.6 = 18.0

Hence,

MSB = SSB / (k–1) = 33.6 /(3–1) = 16.8

MSE = SSE / (n–k) = 18.0 / (15–3) = 18.0 / 12 = 1.5

Then,

$$F_{observed} = \frac{MSB}{MSE}$$

$$= \frac{16.8}{1.5} = 11.2$$

For a choice of $\alpha = 0.05$ the tabulated value of $F_{\alpha, k-1, n-k}$ is 3.88, where k–1=2, n–k=12.

Thus, H_0 is rejected in favour of the alternative that brand effect is present. The summarized information is given below

ANOVA Table

Sources of variation	d.f	SS	MS	F Ratio	
				observed	tabulated
Between brands	2	33.6	16.8	11.2	$F_{.05, 2, 12}$
Error	12	18.0	1.5		=3.88
Total	14	51.6			

Since H_o has been rejected we have reasons to believe that brands are not all equal in respect of their mean lives. To examine the differences we need to make pair-wise comparison. The critical difference can be calculated as

$$CD = t_{.025,12}\sqrt{\frac{2MSE}{n_o}}$$

$$= 2.179\sqrt{\frac{2(1.5)}{5}}$$

$$= 1.688$$

Pairwise differences, in absolute value, are

$$|\bar{x}_1 - \bar{x}_2| = |34 - 30.4| = 3.6$$

$$|\bar{x}_1 - \bar{x}_3| = |34 - 32.8| = 1.2$$

$$|\bar{x}_2 - \bar{x}_3| = |30.4 - 32.8| = 2.4$$

Thus, $|\bar{x}_1 - \bar{x}_2| > CD$ and $|\bar{x}_2 - \bar{x}_3| > CD$ but $|\bar{x}_1 - \bar{x}_3| < CD$.

Here, we conclude that brand B is of lower mean life and is different singnificantly from the other brands A and C. Mean lives of brand A and brand C are not markedly different.

ANOVA for two-way classified data. Let there be two factors each at various levels. Let these factors be denoted by A and B. Let the levels of A be A_1, A_2, ...A_m, and the levels of B be B_1, B_2, ..., B_n. Let the observed value of a unit under i-th level A_i of A and j-th level B_j of B be denoted by x_{ij} Then the mathematical model can be written as

$$x_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ij}$$

where μ is the general effect, α_i is an effect due to i-th level of factor A, β_j is an effect due to j-th level of factor B, γ_{ij} is an interaction between i th level of factor A and j-th level of factor B and e_{ij} is the error term. We assume errors to be independent and normally distributed with mean zero and common variance σ^2. In case of single observation on any combination of the levels of A and B we assume $\gamma_{ij} = 0\ \forall_{i,j}$ Then the model reduces to

$$x_{ij} = \mu + \alpha_i + \beta_j + e_{ij},$$

with $\sum \alpha_i = 0$, $\sum \beta_j = 0$ and $e_{ij} \sim N(o, \sigma^2)$ $\forall$ i, j The least square estimates of μ,

's and β_j 's can be obtained by minimizing the total of the squared errors i.e. $\sum_i \sum_j e_{ij}^2$. Let us write

$$S' = \sum_{i=1}^{m}\sum_{j=1}^{n}(e_{ij})^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_{ij} - \mu - \alpha_i - \beta_j)^2$$

differentiating S with respect to μ, α_i 's and β_j 's and equating each one with zero we get the normal equations as

$$\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = mn\mu + n\sum \alpha_i + m\sum_{i} \beta_j$$

$$\sum_{j=1}^{n} x_{ij} = n\mu + n\alpha_i + \sum_{j=1}^{n} \beta_j \qquad , i = 1, 2, \ldots\ldots m$$

$$\sum_{i=1}^{m} x_{ij} = m\mu + \textstyle\sum_{i=1}^{m} \alpha_i + m\beta_j \qquad , j = 1, 2, \ldots\ldots\ldots, n$$

Since $\sum \alpha_i = \sum_{j} \beta_j = 0$ we get from the 1st normal equation an estimator of μ as the grand mean $\bar{\bar{x}}$, i.e.

$$\hat{\mu} = \frac{1}{mn} \textstyle\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \bar{\bar{x}}, \text{ say}.$$

Similarly, from the next m equations we get estimators of α_i 's as

$$\hat{\alpha}_i = \frac{1}{n} \textstyle\sum_{j=1}^{n} x_{ij} - \bar{\bar{x}}, \quad i = 1, 2, \ldots\ldots\ldots, m$$

Lastly from the next n equations we get estimators of β_j 's as

$$\hat{\beta}_j = \frac{1}{m} \textstyle\sum_{i=1}^{m} x_{ij} - \bar{\bar{x}}$$

Let us denote by $\bar{x}_{i\cdot} = \sum_{j=1}^{n} x_{ij} / n$ and $\bar{x}_{\cdot j} = \sum_{i=1}^{m} x_{ij} / m$. Then we can write

$$\sum_{i}\sum_{j} \left(x_{ij} - \bar{\bar{x}}\right)^2 = \sum_{i} n\left(\bar{x}_{i.} - \bar{\bar{x}}\right)^2 + \sum_{j} m\left(\bar{x}_{.j} - \bar{\bar{x}}\right)^2 + \sum_{i}\sum_{j} \left(x_{ij} - \bar{x}_{i\cdot} - \bar{x}_{\cdot j} + \bar{\bar{x}}\right)^2$$

Writing

$$\text{TSS} = \sum_{i}\sum_{j} \left(x_{ij} - \bar{\bar{x}}\right)^2$$

$$\text{SSA} = \sum_{i} n\left(\bar{x}_{i\cdot} - \bar{\bar{x}}\right)^2$$

$$\text{SSB} = \sum_{j} m\left(\bar{x}_{\cdot j} - \bar{\bar{x}}\right)^2$$

$$SSE = \sum_i \sum_j \left(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{\bar{x}}\right)^2$$

we have

$$TSS = SSA + SSB + SSE.$$

Hence, the total sum of squares (TSS) can be split into three sum of squares one due to variations in the level of factor A, another due to variations in the level of factor B and the last one due to chance variations also known as error.

The corresponding degrees of freedom can be split as follows.

$$(mn-1) = (m-1) + (n-1) + (m-1)(n-1)$$

We can define mean squares due to A as

$$MSA = SSA / (m-1),$$

mean squares due to B as

$$MSB = SSB / (n-1)$$

and mean square error MSE as

$$MSE = SSE /(m-1)(n-1)$$

To test for the absence of factor effects A, we test the null hypotnesis H_{OA} , $\alpha_1 = \alpha_2 = ... = \alpha_m = 0$ by using the test statistic

$$F_A = \frac{MSA}{MSE},$$

which follows, under H_{OA}, F distribution with degrees of freedoms (m–1) and (m–1)(n–1). Decision rule will be to reject H_{OA} in favour of the alternative that factor effects of A vary according to levels if

F_A obsered > $F_{\alpha, (m-1), (m-1)(n-1),}$

α being the level of significance. Similarly, to test for the absence of factor effects B, we test the null hypothesis $H_{OB} : \beta_1 = \beta_2 = ... = \beta_m = 0$ by using the test statistic

$$F_B = \frac{MSB}{MSE}$$

which follows, under H_{OB}, F distribution with degrees of freedoms (n–1) and (m–1)(n–1). Decision rule will be to reject H_{OB} in favour of the alternative that factor effects of B vary according to levels if

F_B observed > $F_{\alpha, (n-1), (m-1)(n-1)}$

α being the level of significance.

This entire test procedure can be summarized in the following tabular form known as ANOVA table.

ANOVA Table (Two-way classified data)

Sources of variation	Degrees of freedom	Sum of squares	Mean sum of squares	Fratio	
				observed	tabulated
Due to factor A	m–1	$SSA = \sum_i n(\bar{x}_i - \bar{x})^2$	$MSA = \frac{SSA}{m-1}$	$F_A = \frac{MSA}{MSE}$	$F_{\alpha, n-1, (m-1)(n-1)}$
Due to factor B	n–1	$SSB = \sum_j m\left(\bar{x}_{.j} - \bar{\bar{x}}\right)^2$	$MSB = \frac{SSB}{n-1}$	$F_B = \frac{MSB}{MSF}$	$F_{\alpha, n-1, (m-1)(n-1)}$
Due to error	(m-1) (n–1)	$SSE = \sum_i \sum_j (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{\bar{x}})^2$	$MSE = \frac{SSE}{(m-1)(n-1)}$		
Total	mn–1	$TSS = \sum_i \sum_j \left(x_{ij} - \bar{\bar{x}}\right)^2$			

Let us consider one example to show the steps of calculation. Following data present output per manshifit for 12 workers. They were trained according 3 different methods and 4 different periods. The two-way classified data are given below :

Method of training	Training period			
	1 week	2 weeks	3 weeks	4 weeks
Method 1	30	32	33	38
Method 2	41	43	45	47
Method 3	50	52	53	58

It may be observed that m=3 and n=4. For a choice of factor A as method of training and Factor B as the period of training, the grand mean is

$$\bar{\bar{x}} = 43.5$$

the method means are

$$\bar{x}_{1.} = 33.25, \bar{x}_{2.} = 44, \ \bar{x}_{3.} = 53.25$$

and the training period means are

$$\bar{x}_{.1} = 40.33, \bar{x}_{.2} = 42.33, \ \bar{x}_{.3} = 43.67, \ \bar{x}_{.4} = 47.67$$

$$TSS = \sum_{i=1}^{3} \sum_{j=1}^{4} x_{ij}^2 - mn\bar{\bar{x}}^2$$

$$= 23598 - 12(43.5)^2$$

$$= 891.00$$

$$SSA = 4\sum_{i=1}^{3}\left(\bar{x} \quad \bar{\bar{x}}\right)^2$$

$$= 801.5$$

$SSB = 3\sum_{j=1}^{4}\left(\bar{x}_{.j} - \bar{\bar{x}}\right)^2$

$= 87.29$

By subtraction,

$SSE = TSS - SSA - SSB = 2.21$

Thus

$MSA = SSA / (m-1) = 801.5/2 = 400.75,$

$MSB = SSB / (n-1) = 87.29/3 = 29.10$

$MSE = SSE / (m-1)(n-1) = 2.21 / 6 = 0.37$

and hence

$$F_A = \frac{MSA}{MSE} = \frac{400.75}{0.37} = 1083.11$$

$$F_B = \frac{MSB}{MSE} = \frac{29.10}{0.37} = 78.65$$

Since $F_{Aobs} > F_{.05, 2, 6,} = 5.14$ we reject $H_{OA} : \alpha_1 = \alpha_2 = \alpha_3$

and since $F_{Bobs} > F_{.05, 3, 6} = 4.76$ we reject $H_{OB} : \beta_1 = \beta_2 = \beta_3 = \beta_4$

Thus both factors i.e. the method of training and period of training are having significant effects on OMS value. The summarized presentation in tabular form is given below.

ANOVA Table

Sources of variation	d.f	SS	MS	F ratio	
				observed	tabulated (5%)
Due to Method	2	801.5	400.75	$F_A = 1083.11$	5.14
Due to period	3	87.29	29.10	$F_B = 78.65$	4.76
Error	6	2.21	0.37		
Total	11	891.00			

Analysis of variance for two-way classified data with multiple observations per cell

Let there be factors A and B, A having m levels, B having n levels. Let there be p observations per each combination of the levels of A and B The corresponding mathematical model will be

$$x_{ijk} = \mu + \alpha_i + \beta_i + \gamma_{ij} + e_{ijk}$$

$i = 1, 2, \ldots\ldots., m,\ j = 1, 2, \ldots\ldots.., n,\ k = 1, 2, \ldots\ldots.., p.$

μ = general effect,

α_i = effect due to i th level of fector A.

β_j = effect due to j th level of factor B and

γ_{ij} = interaction effect between i-th level of factor A and j-th level of factor B.

e_{ijk} = error following independent normal distribution N(0, σ^2)

Writing $\bar{\bar{x}} = \frac{1}{mnp}\sum_i\sum_j\sum_k x_{ijk}$ as the grand mean

$\bar{x}_{i..} = \frac{1}{np}\sum_j\sum_k x_{ijk}$ as the i-th mean due to factor A

$\bar{x}_{.j.} = \frac{1}{mp}\sum_i\sum_k x_{ijk}$ as the j-th mean due to factor B

$\bar{x}_{ij.} = \frac{1}{p}\sum_k x_{ijk}$ as the i-j-th mean due to interaction

we may define

$$TSS = \sum_i\sum_j\sum_k\left(x_{ijk} - \bar{\bar{x}}\right)^2$$

$$SSA = np\sum_i\left(\bar{x}_{i..} - \bar{\bar{x}}\right)^2$$

$$SSB = mp\sum_j\left(\bar{x}_{.j.} - \bar{\bar{x}}\right)^2$$

$$SS\overline{AXB} = p\sum_i\sum_j\left(\bar{x}_{ij.} - \bar{x}_{i..} - \bar{x}_{.j.} + \bar{\bar{x}}.\right)^2$$

$$SSE = \sum_i\sum_j\sum_k\left(x_{ijk} - \bar{x}_{ij}.\right)^2$$

It can be shown that sum of squares can be split like

$$TSS = SSA + SSB + SS\overline{AXB} + SSE$$

and degrees of freedom can be split like

$$mnp - 1 = (m-1) + (n-1) + (m-1)(n-1) + mn(p-1).$$

We can calculate mean squares as

MSA = SSA / (m–1)

$$MSB = SSB \,/\, (n-1)$$

$$MS\overline{AXB} = SS\overline{AXB} \,/\, (m-1)(n-1)$$

$$MSE = SSE \,/\, mn\,(p-1)$$

To test for the absence of factor effects A we consider null hypothesis H_{OA} : $\alpha_1 = \alpha_2 = \ldots\ldots\ldots = \alpha_m = 0$. The test statistic is

$$F_A = \frac{MSA}{MSE}$$

and the decision rule is to reject H_{OA} if F_A observed > $F_{\alpha,\, m-1,\, mn(p-1)}$, where α is the level of significance. To test for the absence of factor effects-B we consider full hypothesis H_{OB} : $\beta_1 = \beta_2 = \ldots\ldots = \beta_n = 0$ The test statistic is

$$F_B = \frac{MSB}{MSE}$$

and the decision rule is to reject H_{OB} if F_B observed > $F_{\alpha, n-1,\, mn(p-1)}$, where α is the level of significance. T$_o$ test for the absence of interaction effects we consider null hypothesis H_{01} : $\gamma_{ij} = 0\ \forall_{i,\,j}$. The test satistic is

$$F_I = \frac{MS\overline{AXB}}{MSE}$$

and the decission rule is to reject H_{OI} if F_I observed > $F_{\alpha,\,(m-1)(n-1),\, mn(p-1)}$ where α is the level of significance.

10.3. Multiple regression analysis

In regression analysis, as discussed earlier, we make use of a relationship among variables to predict one from the rest. The variable to be predicted is called dependent variable and let it be represented by the symbol y. For example, if we are interested to predict sales in a geographic region we may consider sales (y) as a dependent variable. If it is influenced by the number of customers of that region (x_1) and the price of that product (x_2) we may like to predict y based on a relationship between y and (x_1, x_2). Given that relationship one may consider the future values of x_1 and x_2 to predict the future value of y. Since we are dealing with three variables y, x_1 and x_2 we call it multiple regression analysis.

In general we may have k predictors to predict a dependent variable y. Let the predictors be denoted by $x_1, x_2, \ldots\ldots, x_k$. Usnally, we get a complete set of n observations.

$(y_\alpha, x_{1\alpha}, x_{2\alpha}, \ldots\ldots\ldots, x_{k\alpha})$ $\alpha = 1, 2, \ldots\ldots, n$

based on which a relationship model between y and $(x_1, x_2, \ldots\ldots\ldots, x_k)$ can be developed. The most simple relationship is a linear relationship where

$$y = a_o + \sum_{j=1}^{k} a_j x_j + \epsilon$$

$a_1, \ldots\ldots\ldots, a_k$ being regression coefficients, a_o being a constant and ϵ being the error term following normal distribution in the ideal situation. We may obtain expresions of error terms as follows :

$$y_\alpha = a_o + \sum_{j=1}^{k} a_j x_{j\alpha} + \epsilon_\alpha$$

$$\text{or, } \epsilon_\alpha = y_\alpha - a_o - \sum_{j=1}^{k} a_j x_{j_\alpha}$$

If one employes least square method for estimation of the parameters $(a_o, a_1, \ldots\ldots\ldots a_k)$ one minimizes $S = \sum_{\alpha=1}^{n} \epsilon_\alpha^2$ with respect to $a_o, a_1, \ldots\ldots\ldots, a_k$. The resultant equations are

$$\frac{\partial s}{\partial a_o} = 0,\ \frac{\partial s}{\partial a_1} = 0, \ldots\ldots\ldots, \frac{\partial s}{\partial a_k} = 0$$

These equations are known as normal equations. In this linear regression model the normal equations are

$$\sum_{\alpha=1}^{n} y_\alpha = na_o + \left(\sum_{\alpha=1}^{n} x_{1\alpha}\right) a_1 + \left(\sum_{\alpha=1}^{n} x_{2\alpha}\right) a_2 + \ldots\ldots + \left(\sum_{\alpha=1}^{n} x_{k\alpha}\right) a_k$$

$$\sum_{\alpha=1}^{n} y_\alpha x_{j\alpha} = \left(\sum_{\alpha=1}^{n} x_{j\alpha}\right) a_o + \left(\sum_{\alpha=1}^{n} x_{1\alpha} x_{j\alpha}\right) a_1 + \ldots\ldots\ldots + \left(\sum_{\alpha=1}^{n} x_{k\alpha} x_{j\alpha}\right) a_k \quad j = 1, 2, \ldots\ldots, k$$

The first equation gives

$$\bar{y} = a_o + \bar{x}_1 a_1 + \ldots\ldots\ldots + \bar{x}_k a_k.$$

If we multiply this equation ~~with~~ by $n\bar{x}_j$ / dl- and subtract from the j-th equation of the remaining set, $j = 1, 2, \ldots\ldots\ldots k$, we get the following equation

$$\sum_{\alpha=1}^{n} y_\alpha x_{j\alpha} - n\bar{y}\,\bar{x}_j = a_1 \left[\sum_{\alpha=1}^{n} x_{1\alpha} x_{j\alpha} - n\bar{x}_1\bar{x}_j\right] + \ldots\ldots + a_k \left[\sum_{\alpha=1}^{n} x_{k\alpha} x_{j\alpha} - n\bar{x}_k\bar{x}_j\right] \quad j = 1, 2, \ldots, k,$$

$$\text{or, } S_{yj} = a_1\, s_{1j} + a_2\, s_{2j} + \ldots\ldots + a_k s_{kj} \qquad j = 1, 2, \ldots\ldots\ldots, k.$$

$$\text{where } S_{ij} = \sum_{\alpha=1}^{n} x_{i\alpha} x_{j\alpha} - n\bar{x}_i\bar{x}_j = \sum_{\infty=1}^{n} (x_{i\alpha} - \bar{x}_i)(x_{j\alpha} - \bar{x}_j) \qquad i, j = 1, 2, \ldots, k$$

and $S_{yj} = \sum_{\alpha=1}^{n} y_\alpha x_{j\alpha} - n\overline{y}\overline{x}_j = \sum_{\alpha=1}^{n}(y_\alpha - \overline{y})(x_{j\alpha} - \overline{x}_j), \qquad j = 1,2,\ldots,k.$

Writing $S = ((s_{ij}))$ the matrix of order kxk with (i,j)-th element as S_{ij},

$\underset{\sim}{S}_y = (S_{y1}, S_{y2}, \ldots\ldots\ldots S_{yk})'$ the vector of order kx1

and $\underset{\sim}{a} = (a_1, a_2, \ldots\ldots\ldots, a_k)'$ the vector of order kx1

we have $S\underset{\sim}{a} = \underset{\sim}{S}_y$

Hence $\underset{\sim}{\hat{a}} = S^{-1}\,\underset{\sim}{S}_y$

and $\hat{a}_o = \overline{y} - \hat{\underset{\sim}{a}}' \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_k \end{pmatrix}$

Thus, from (k+1) equations we can estimate the (k+1) parameters $a_o, a_1, \ldots.., a_k$ with the estimated values $\hat{a}_o, \hat{a}_o, \hat{a}_1, \ldots\ldots.. \hat{a}_k$. Then the regression equations can be expressed as

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \ldots\ldots\ldots + \hat{a}_k x_k$$

A measure for goodness of regression is the square of the multiple correlation coefficient, R^2, where

$$R^2 = \frac{\text{Sum of squares due to regression}}{\text{Total sum of Squares}}$$

$$= \frac{\sum_{\alpha=1}^{n}(\hat{y}_\alpha - \overline{y})^2}{\sum_{\alpha=1}^{n}(y_\alpha - \overline{y})^2}$$

R^2 lies between 0 and 1 and a value of R^2 close to 1 indicates goodness of linear regression model. If we standardize the independent and dependent variables the regression equation will be independent of the measuring units. Regression coefficients will then be referred as *beta coefficients.* Since these coefficients are obtained from unit free data, one may compare these beta coefficients to ascertain the inportance of each predictor in the overall regression equation.

To test for the suitability of the muliple linear regression, one may follow the ANOVA method as presented below.

ANOVA Table

Sources of variation	degress of freedom	sum of squares	mean squares	F ratio	
				observed	tabulated
Due to linear regression	k	SSR=$\sum_{j=1}^{k}\hat{a}_j s_{y_j}$	MSR=SSR/k	$F=\frac{MSR}{MSE}$	$F_{\alpha,\,k,\,n\text{-}k\text{-}1}$
Error	n-k-1	SSE=$(y_n-\bar{y})^2-\sum_{j=1}^{k}\hat{a}_j s_{y_i}$	MSE=SSE/(n-k-1)		
Total	n–1	$\sum_{\alpha=1}^{n}(y_\alpha-\bar{y})^2$			

The decision rule is to accept the suitability of linear regression if F observed > $F_{\alpha,k,\ n\text{-}k\text{-}1}$; otherwise we consider the regression coefficients to be all insignificant.

Let us consider one example to explain the calculation for multiple regression analysis. Let y be the sales of a product measured in thousand rupees, x_1 be the expenditure on adveretisement in some unit of measmement and x_2 be expenditure on sales personnel, again measured in thousand rupees for one month in a given locality. The following table gives 18 months' figures on y, x_1 and x_2.

Month	sales	Expenditure on Ad.	Expenditure on sales personnel
1	22	6	9
2	54	7	16
3	33	5	12
4	45	5	22
5	28	4	8
6	21	3	6
7	47	6	11
8	59	8	14
9	41	5	13
10	35	6	14
11	28	4	6
12	13	5	7
13	23	6	5
14	30	6	9
15	16	5	8
17	30	5	7
18	58	9	17

We may calculate the mean values based on n = 18. Mean values are

$$\overline{y} = 35.28, \ \overline{x}_1 = 5.67, \ \overline{x}_2 = 10.94$$

Sum of squares and sum of products can be obtained,using the usual formula and are presented below :

$S_{yy} = 3539.611$, $S_{11} = 35.667$, $S_{22} = 352.944$, $S_{12} = 58.667$
$S_{y1} = 256.667$, $S_{y2} = 858.278$
Then

$$\hat{a}_o = \overline{y} - \underset{\sim}{\hat{a}}' \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \end{pmatrix}$$

$$= 35.28 - \hat{a}_1 \, 5.67 - \hat{a}_2 \, 10.94$$

$$\underset{\sim}{\hat{a}} = S^{-1} \underset{\sim}{S}_y = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}^{-1} \begin{bmatrix} S_{y1} \\ S_{y2} \end{bmatrix}$$

$$\text{or,} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \end{pmatrix} = \begin{bmatrix} 35.667 & 58.667 \\ 58.667 & 352.944 \end{bmatrix}^{-1} \begin{bmatrix} 256.667 \\ 858.278 \end{bmatrix}$$

$$= \begin{bmatrix} 0.039 & -0.006 \\ -0.006 & 0.004 \end{bmatrix} \begin{bmatrix} 256.667 \\ 858.278 \end{bmatrix}$$

$$= \begin{bmatrix} 4.860 \\ 1.893 \end{bmatrix}$$

Thus, $\hat{a}_1 = 4.860$, $\hat{a}_2 = 1.893$ and hence $\hat{a}_0 = -12.986$.

Then the multiple regression equation can be writen as

$$\hat{Y} = -12.986 + 4.860X_1 + 1.893X_2$$

The corresponding R^2 value is

$$R^2 = \frac{\sum_{\alpha=1}^{n} \left(\hat{Y}_\alpha - \overline{Y}\right)^2}{\sum_{\alpha=1}^{n} \left(Y_\alpha - \overline{Y}\right)^2}$$

$$= \frac{2872.136}{3539.611} = 0.811$$

Since R^2 value is close to 1 we may consider the multiple regression to be appropriate. The ANOVA table for testing $H_o : a_1 = 0, a_2 = 0$ against the alternative that the regression equation is effective is given below

ANOVA Table

Sonrces of	d.f.	SS.	Ms.	F ratio	
variation				observed	Tabulated
Due to regression	2	2872.136	1436.068	32.273	F.05,2,15
Error	15	667.475*	44.498		= 3.68
Total	17	3539.611			

* obtained by subtraction

Thus, analysis of variance suggests the rejection of Ho : $a_1 = 0$, $a_2 = 0$ in favour of the alternative that the regression equation is effective at 5% level of significance.

Stepwise regression. It operates in an iterative fashion introducing predictor variables one by one in the regression equation. The first predictor out of k predictors x_1, x_2, ...x_k is the one that predicts best. In terms of R^2 value we may select the best predictor. Thus, we consider k regression equations ; i-th equation deals with prediction of y by x_i. For each equation we calculate R^2 value. The equation that gives rise to maximum R^2 value is appreciated and the corresponding predictor is the first variable to enter into the set of predictors. The second variable is the one which provides the best prediction of y in combination with the first predictor. This process goes on till we arrive at a prefixed high value of R^2. For future prediction only those selected predictors are to be observed. Rest of the variables are of limited use.

10.4 Cluster analysis

The objective of this advanced technique for data analysis is to identify units such as brands, regions, individual customers, which are close to or similar to each other in some sense. This analysis gives rise to natural grouping based on multidimensional observations. Selection of variables is based on their differentiating power in respect of the units of interest. Then each unit is observed in respect of all such selected variables.

Let the vector measure be denoted by $\underset{\sim}{X} = (X_1, X_2, ...X_p)$ which is made of p scalar measures. Let the value of $\underset{\sim}{X}$ as observed for the α-th unit be denoted by $\underset{\sim}{X}_{\alpha,}$ $\alpha = 1, 2, ...n$. Then the Euclidean distance between the αth unit and β–th unit will be expressed as

$$d_{\alpha\beta} = [\,(\underset{\sim}{X}_\alpha - \underset{\sim}{X}_\beta)' \;\; (\underset{\sim}{X}_\alpha - \underset{\sim}{X}_\beta)]^{\frac{1}{2}}$$

A small value of $d_{\alpha,\beta}$ means units α and β are similar in nature in respect of $\underset{\sim}{X}$. A high value of $d_{\alpha\beta}$, on the other hand, means units α and β are dissimilar in respect of $\underset{\sim}{X}$. The final distance matrix is the collection of all the distances so calculated.

***Bottom up approach* of cluster analysis** starts with the hypothesis that initially there are as many clusters as units. The most similar units are then grouped together. Then other units are added or grouped sequentialy so that at the end there will be only one cluster. Calculation of distance matrix when units are grouped together is the main problem to be addressed. There are three different ways of recalculating the distances namely *single linkage, complete linkage* and *average linkage.* Under single linkage, the distance between two groups of units is the minimum of their unit-wise distances. Under complete linkage, the distance between two groups of units is the maximum of their unit-wise distances. Under average linkage the distance between two groups of units is the average of all the unit-wise distances. Any one of the above linkages may be adopted for the calculation of distance matrices at the subsequent stages.

A popular bottom up approach is discribed below. It is known as *Agglomerative hierarchical clustering algorithm.*

1. Start with n clusters each containing a single unit/entity.

2. Calculate the initial distance matrix $D = ((\ d_{\alpha\beta}))_{nxn.}$

3. Search for the nearest pair of entities by examining distances and identifying the minimum distance.

4. Let the most similar cluster be u and v.

5. Merge u and v and label the newly formed cluster by the symbol (uv).

6. Update the entries of the distance matrix by removing the rows and columns corresponding to u and v and by adding a row and a column corresponding to the new cluster (uv).

7. Repeat steps 3 – 6 for (n–1) times.

8. Present the entire work of cluster selection on a *dendrogram* by recording the identity of the clusters that are merged and the levels/distances at which the merger takes place.

Let us explain the entire procedure with a small example. Consider a 4-brand product field, brands being labelled as A, B, C and D. The following is the distance matrix.

$$D = \begin{array}{c|cccc} & A & B & C & D \\ A & 0 & & & \\ B & 3 & 0 & & \\ C & 7 & 9 & 0 & \\ D & 8 & 6 & 5 & 0 \end{array}$$

The minimum distance is 3. It is the distance between brand A and brand B. So merge A and B as (AB). Recalculate the distance matrix in the following way using single linkage method :

$d_{(AB)C} = \text{Min} \{ d_{AC} , d_{BC} \} = \text{Min} \{7, 9\} = 7$

$d_{(AB)D} = \text{Min} \{ d_{AD} , d_{BD}\} = \text{Min} \{8, 6\} = 6$

The new distance matrix becomes :

$$D = \begin{array}{cc} & \begin{array}{ccc} (AB) & C & D \end{array} \\ \begin{array}{c} (AB) \\ C \\ D \end{array} & \begin{bmatrix} 0 & & \\ 7 & 0 & \\ 6 & 5 & 0 \end{bmatrix} \end{array}$$

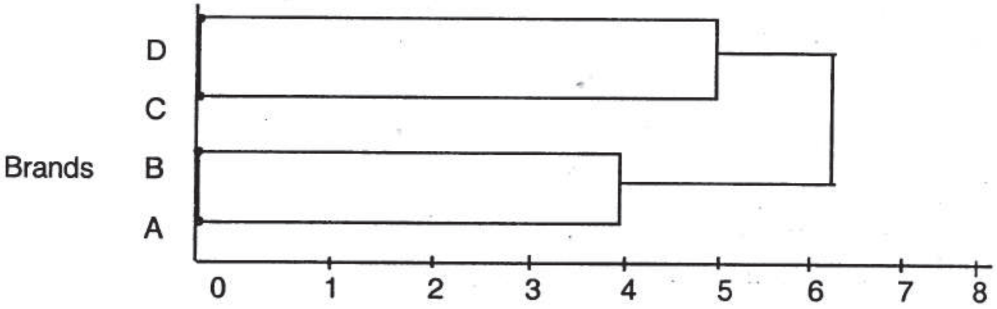
The minimum distance in this matrix is 5. It is the distance between brand C and brand D. So merge C and D as (CD). Recalculate the distance matrix as follows, using single linkage method :

$$\begin{aligned} d_{(AB)\,(CD)} &= \text{Min} \{ d_{(AB)C} , d_{(AB)\,D} \} \\ &= \text{Min} \{ 7, 6 \} \\ &= 6. \end{aligned}$$

The new distance matrix becomes

$$D = \begin{array}{cc} & \begin{array}{cc} (AB) & (CD) \end{array} \\ \begin{array}{c} (AB) \\ (CD) \end{array} & \begin{bmatrix} 0 & \\ 6 & 0 \end{bmatrix} \end{array}$$

Thus, the final merging is between (AB) and (CD) The following dendrogram depicts the scheme of clustering.



Thus, we observe two clusterings, one cluster made of brands A and B and the other cluster made of brands C and D. Most similar brands are A and B.

10.5. Factor analysis

Factor analysis makes an attempt to identify the latent factors that affect the observed variables. It inolves a theory or a model to explain the generation of observation through some common factors and some specific variables. If there are p variables, say X_1, X_2, ...X_p, which are observable, it may be assumed that these variables in turn,can be expressed as a linear combination of a small number of factors F_1, F_2, ..., F_m $(m<p)$ and specific variates e_1, e_2, ..., e_p, e_i being specific to X_i. The full factor analysis model is given by

$$X_i = a_{i1} F_1 + a_{i2} F_2 + \text{..........} + a_{im} F_m + e_i \quad i = 1, 2, \text{..........}, p,$$

where F_j 's are uncorrelated factor effects, uncorrelated with the specific variates and where specific variates are uncorrelated among themselves. A stricter condition is independence of F_j's, e_i's and independence among F_j's and e_i's. The coefficient a_{ij} is the *factor loading* of factor F_j in the variable X_i, i = 1, 2, ... p, j = 1, 2, ...m. The matrix $A = ((a_{ij}))$ is known as the loading matrix.

With var $(F_j) = 1$, j = 1, 2, ..., m and $\text{Var}(e_i) = w_i$, i = 1, 2, ..., p, we refer w_i as the *specific variance* and $\{\text{Var}(X_i) - w_i\}$ as the *communality* of the variable X_i . It is easy to note that

$$\begin{aligned} \text{Var}\ (X_i) &= a_{i1}^2\ \text{Var}\ (F_1) + \text{........} + a_{im}^2\ \text{Var}\ (F_m) + \text{Var}(e_i) \\ &= \{\ a_{i1}^2 + \text{...........} + a_{im}^2\ \} + w_i \end{aligned}$$

and hence communality is measured by

$$\{\ \sum_{j=1}^{m} a_{ij}^2\ \},\ i = 1, 2, \text{........}, p$$

for the variable X_i.

Let us explain the determination of loading, communality and specific variance through a small example studied by Spearman. In a sample of examination marks in classics (X_1), French (X_2) and English (X_3) the correlation matrix has been observed to be R,

$$R = \begin{array}{cccc} \text{Classics} & \text{English} & \text{French} & \\ 1.00 & 0.83 & 0.78 & \text{Classics} \\ 0.83 & 1.00 & 0.67 & \text{French} \\ 0.78 & 0.67 & 1.00 & \text{English} \end{array}$$

If we assume a single factor model we may write

$$\begin{aligned} X_1 &= a_1\ F + e_1 \\ X_2 &= a_2\ F + e_2 \\ X_3 &= a_3\ F + e_3, \end{aligned}$$

where the common factor, F, denotes the general intellectual ability and where e_1,

e_2 and e_3 are variables specific to classics, French and English. To estimate the factor loadings and specific variances we consider the following equations

$$\text{Corr}(X_1, X_2) = a_1a_2 = 0.83$$

$$\text{Corr}(X_1, X_3) = a_1\ a_3 = 0.78$$

$$\text{Corr}(X_2, X_3) = a_2a_3 = 0.67$$

$$w_1 = 1 - a_1^2$$

$$w_2 = 1 - a^2_2$$

$$w_3 = 1 - a_3^2$$

It is easy to obtain

$$a_1a_2a_3 = \sqrt{(a_1a_2)(a_1a_3)(a_2a_3)}$$

$$= \sqrt{(0.83)(0.78)(0.67)}$$

$$= 0.65860$$

Hence $a_1 = (a_1a_2a_3)/(a_2a_3) \quad = 0.65860 / 0.67$

$$= 0.983$$

$$a_2 = (a_1a_2a_3)/(a_1a_3) \quad = 0.65860 / 0.78$$

$$= 0.844$$

$$a_3 = (a_1a_2a_3)/(a_1a_2) \quad = 0.65860 / 0.83$$

$$= 0.793$$

and $w_1 = 1 - (0.983)^2 = 0.034$

$$w_2 = 1 - (0.844)^2 = 0.288$$

$$w_3 = 1 - (0.793)^2 = 0.371$$

Thus, the estimated factor analysis model is

$$X_1 = 0.983F + e_1 \quad , \quad \text{Var}(e_1) = 0.034$$

$$X_2 = 0.844F + e_2 \quad , \quad \text{Var}(e_2) = 0.288$$

$$X_3 = 0.793F + e_3 \quad , \quad \text{Var}(e_3) = 0.371$$

In a multifactor model the column vectors of the loading matrix can be obtained from the normalized eigen vectors of the dispersion matrix of $(X_1, X_2, \ldots.., X_p)$. Since there are p such eigen vectors we select those eigen vectors which correspond to high eigen values, say $\lambda_1, \lambda_2, \ldots\ldots., \lambda_m$ with $\lambda_1 \geq \lambda_2 \geq \lambda_m \geq \ldots\ldots\ldots.. \geq \lambda_p$.

10. 6. Conjoint analysis

Study of consumer behariour reveals that consumers make a complex trade-off among different product attributes and their levels and make the final purchasing decision. Conjoint analysis is an indirect way of arriving at the utility values of each attribute at each level so as to throw light on the overall utility of a product and its impact on the consumer's choice process. This technique is based on preference data obtained from respondents' assessment of varions levels of different attributes. For example, one may think of different TV brands to be examined in respect of different price levels, different years of guarantee and different levels of image-quality. Respondents may be requested to indicate his/her preference for each level combination of two attributes taken at a time. Let us consider, for example, price and years of guarantee as the two attributes. Let there be three levels of price low, medium and high: and three levels of year of guarantee:5 years,7 years and 9 years. Thus, there are in total,9 combinations in respect of these two attributes. A respondent may be asked to rank these level combinations. The rank observations are the preference data that work as inputs to the conjoint analysis. Since we have three altributes of importance we can think of ${}^3C_2 = 3$ combinations of two altributes. If there are 500 respondents then the total number of preference data will be 1500.

An example of preference data is given below :

Years of guarantee	Price		
	Low	Medium	High
5 Years	4	7	9
7 Years	3	6	8
9 Years	1	2	5

Here preference 1 indicates the best choice and preference 9 indicates the worst choice.

It may be observed that customers are trading off between price and years of guarantee. In view of this, preference data are known as trade off data. It is easy to collect trade off data specially when levels are only a few. An alternative to trade-off data is known as full-profile data where respondents are given a complete profile consisting of all the attributes at different level combination. Respondents are asked to indicate their preferences for varions alternative full-profiles. For example, one full-profile is medium price with high image quality and 7 years guarantee.

Once the preference data are collected the investigator attempts to determine utilities / partworths. Software packages are available for determination of utilities. one such package is MONANOVA. It involves an interation procedure where utilities are assigned to each level of each attribute and overall utility is calculated based on a prefixed rule. Given the overall utility so calculated based on arbitrary assignment of partworths, we examine the closeness between utility based ranking and ranking based on preference data. If matching is not satisfactory, part worths are changed and the entire exercise gets repeated. The iteration procedure stops when the matching between the utility based ranks and preference data based ranks becomes unidirectional.

Conjoint analysis is very useful for evaluating a product concept.During the stage of product design it may help the concept planner in arriving at a d϶sirable configuration that will meet the customers' need. However, conjoint analysis is dependent on the levels and attributes selected at the start of the study. Any change in the later stage in the number of attributes or their levels may change the entire result and a fresh study must be taken up.

10.7. Summary

To draw statistical infererce in some managerial problems an investigator may have to employ advanced analytical tools Analysis of variance (ANOVA) is one method that works on the principle of division of total variation into chance variation and causal variation. In case of a single assignable cause we consider ANOVA for one way classified data. For two assignable causes of variation ANOVA for two-way classified data will be appropriate. The test procedure in terms of F ratio has been described for both one-way classified data and two-way classified data with single observation per cell and multiple observations per cell. Multiple observations per cell provide opportunity to test for presence of interaction among the two causes/factors.

Under multiple regresion analysis the linear regression model has been described along with the least square method for estimating the regression coefficients and intercept. To measure the goodness of regression the concept of multiple correlation coefficient has been introduced. A value close to 1 is an indication of goodness of linear regression. The ANOVA procedure for testing the suitability of the multiple linear regression has been included. Also the step-wise regression procedure to include the predictors in sequence and according to importance has been described as this saves cost of information when used for future prediction.

Cluster analysis, an interesting multivariate technique, has been presented next to indicate how natural groupings can be identified using multidimensional observations and resultant distance matrix. The bottom up approach starts with as many clusters as units. The most similar units are then grouped together. Then other units are added or grouped sequentially so that at the end there is only one large group covering all the units. For the calculation of distances at a subsequent stage one may either go by single linkage based on minimum distance, or by complete linkage based on maximum distance, or by average linkage based on average distance. The entire procedure of cluster selection may be presented graphically to indicate the sequence of groupings and distances. This diagram is known as dendrogram.

Factor analysis, another widely used advanced data analysis tool, aims at identifying the latent factors that affect the observed variables. The full factor analysis model considers a linear combination of factor variables and specific variate to describe an observed variable. The contribution of factors towards the variability in the observed variable is known as communality and the contribution of specific variable is known as specific variance. To determine these measures one has to estimate the coefficients of factors in each linear combination. These coefficients are known as factor loadings. Factor loadings can be determined by various methods and one such method has been described in terms of correlation matrix.

Conjoint analysis is a multivariate approach that attempts to describe the complex trade off that takes place in the mind of a consumer while making a purchasing decision. It deals with preference data of full profile type or trade-off type. Utilities are assigned to different levels of the attributes under considenation and through an iterative procedure a choice of utilities is reached which matches with the observed preference pattern.

10.8. Questions

Long answer type

1. How does ANOVA technique work? For a one-way classified data describe the model, the estimation procedure for the model parameters and the tabular form of test.

2. What should be the requirement of data if interaction is present in a two-factor model ? Describe the test procedure.

3. How can you make use of ANOVA technique for examining the suitability of a multiple linear regression ?

4. In a multiple linear regression setup describe the estimation procedure for estimating parameters of the regression equation.

5. Explain single linkage, complete linkage and average linkage. Also describe the bottom up approach of clustering with special reference to agglomerative hirerarchical clustering algorithm.

6. Explain with an example of your choice how the correlation matrix can be used in estimating the parameters of a single factor factor analysis model.

7. What is the need of conjoint analysis ? Describe the principle based on which MONANOVA package works for carrying out conjoint analysis.

8. Carry out analysis of variance to test whether linear regression is a good fit.

9. Below are given the sales (in Rs 00) per month for three brands of soap.

Brand A	Brand B	Brand C
77	110	74
63	107	48
84	138	41
70	80	62
95	135	66
88	79	72
81	127	71
101	99	47

Test if brands differ significantly among themselves. Write down the analysis of variance table.

10. The following figures relate to demand (in 000 units) of three varieties A, B and C of a product of a company for six weeks. Is there any significant difference in demand when analysed accoding to brand and according to week ?

Week

Brand	1	2	3	4	5	6
A	44	38	47	20	25	22
B	52	41	55	39	37	39
C	40	35	43	22	21	20

11. Following information are available on the sales and related issues for a compancy for monitoring sales.

Year

	2004	2003	2002	2001	2000
Sales (Rs 000)	375	350	341	332	321
Expenditre on Ad (Rs. 000)	82	75	67	66	60
Extent of competition (on a 6 point scale)	5	5	4	3	2

Develop a multiple linear regression equation for sales forecasting. How much should be spent on Ad if the company wants to have sales around Rs 400,000 given that the extent of competition is as in 2004 ?

12. Carry out cluster analysis based on the following distance matrix. Also draw the dendrogram.

Distance matrix

Brand

$$\begin{array}{cc|ccccc|} & & A & B & C & D & E \\ & A & 0 & & & & \\ \text{Brand} & B & 15 & 0 & & & \\ & C & 13 & 16 & 0 & & \\ & D & 11 & 14 & 17 & 0 & \\ & E & 12 & 18 & 10 & 19 & 0 \end{array}$$

Short Answer Type Questions

1. Explain the construction of critical differerce. When is it needed ?

2. Write down the models for ANOVA for a two-factor situation. Comment on interaction effects and their estimation.

3. Define multiple correlation coefficient. Indicate one of its uses

4. What do you mean by beta coefficient ?

5. Explain the step-wise regression method.

6. Using the following information calculate distance between brand A and brand B.

Item	Brand A	Barnd B
Awareness	6	3
Quality level	5	4
Esteem value	5	3
Use value	4	4

7. Using the distance matrix given in the text for four-brand problem,carry out cluster analysis based on complete linkage.

8. Explain the terms communality and specific variance.

9. What do you mean by full-profile data ? How does it differ from trade-off data ?

Objective type Questions :

Indicate whether the following statements are true or false.

1. One-way classified data deals with only one population.

True □ False □

2. Average of SSB and SSE is equal to TSS

True □ False □

3. Critical difference can be calculated if number of observations is same for each population.

True □ False □

4. For ANOVA, error components are assumed to be normally distributed.

True □ False □

5. Normal equations are those equations which are normally distributed.

True □ False □

6. R^2 Value always lies between 0 and 1/2.

True □ False □

7. If there are k predictors, d.f. of SS due to regression will be (k – 1)

True □ False □

8. Stepwise regression reduces the information requirement.

True □ False □

9. Single linkage provides lower distance than average linkage.

True □ False □

10. Clustering of units does not depend on the choice of linkage.

True □ False □

11. Dendrogram presents the scheme of clustering.

True □ False □

12. In factor analysis it is desirable to have specific variance more than the communality value.

True □ False □

13. Eigen vectors can be used to develop factor loading.

True □ False □

14. Conjoint analysis may be useful in evaluating a product concept.

True □ False □

15. MONANOVA is a monotonic analysis of variance tool.

True □ False □

10.9. Report Writing

The end task of any research activity is the formal or informal reporting on the entire work. **Informal reports** follow short format which is appropriate when the research problem is well defined but of limited scope and the research methodology is simple. For example, **Interim Reports** are usually short reports where progress and problems are presented to draw the attention of the planners. Short reports run into 10–15 pages with brief discussions on the problem studied and its interrelated issues, conclusions drawn and recommendations made. **Formal reports,** on the other hand, are long reports describing in detail reasons for the choice of the problem, statement of the problem, background, information, methodology used, research findings, conclusions and recommendations and additional supportive information. Formal reports are of two types **Technical report** and **Management report.** Technical reports are of use for current and future researchers because they present complete documentation of the current work along with original data set and details of the research design. Management reports are meant for quick understanding of the research findings and their basic style of presentation in non-technical style.

Followings are the major components of a report . Depending on the type of report some components may be skipped as indicated earlier and some components may be expanded too.

Components of a report :

1. Prepatory information
   - (a) Letter of invitation
   - (b) Title page
   - (c) Authorization
   - (d) Preview of the work
   - (e) Contents
2. Introduction
   - (a) Statement of the problem
   - (b) Objectives of the study
   - (c) Current status
3. Methodology
   - (a) Sampling design
   - (b) Observational design
   - (c) Data collection
   - (d) Data analysis
   - (e) Limitations.
4. Findings
5. Conclusions
   - (a) Summary & Conclusions
   - (b) Recommendations.
6. Annexures
7. References.

Let us briefly describe these components. **Letter of invitation** assumes importance when the research is conducted as per client's request. Under a formal relationship between the client and the researcher, a copy of the letter of invitation should be added at the beginning of the report.

Title page includes the specific fitle of the report, the identification of the researcher and the organization / client for whom the research has been carried. It should also indicate the year of publication of the report. Regarding the title, we may suggest that the title should indicate the variables included in the study and the population on which the study has been conducted.

Letter of authority is the proof of assignment of the work. Mostly for public organization or for confidential nature of the work a formal letter of authority is needed. It should be issued before the start of the work. The same should also be included in the formal report as a proof of authority for undertaking the work.

Preview of the work provides a bird's eye view of the entire work along with the summary of the important findings and conclusions. Generally, preview of the work is written after the completion of the rest of the report writing. Style of presentation of the preview should be a popular one where technical details are to be arrided. This is basically an invitation to the readers for taking interest in the detailed report. **Contents** in a sketchy guide about the coverage of the work and is useful especially when the report is a long one.

Statement of the problem is the starting point of the detailed report as it explains the need for undertaking the particular research activity. The entire report should be consistent with the stated problem and try to suggest a solution to the problem. **Objectives of the study** aim at defining the critical variables and describing their interrelationships keeping in mind not only the problem under study but also the purpose of the study. In fact, purpose describes the philosophic aspects of objectives. Current status presents the background information to facilitate the movement from known to unknown. It includes review of the existing literature on the subject, presentation of secondary data, preliminary results of exploration from pilot survey and or experience survey.

Sampling desing and observational design describe, in sequence, the target population, sampling unit, sampling method and measuring instruments. Along with these procedural details, the rationale for using one type of design inplace of their competing designs must be spelt out. Merits and demerits of the chosen design should be idenfified. **Data Collection**, reports training and management of investigators employed for that purpose, period for data collection, field situation, handling of irregularities, if any, and other special observations. These are all related to collection of primary data. In case of collection of secondary data, report should indicate the relevance of secondary data.

Data analysis describes the methodology of data handling, initial analysis, detailed analysis, Statistical testing of hypothesis and similar such technical information.

Limitations are parts of every research study, mainly in the field of social science. Indication of implementational problems must be given to ensure transparency in the work procedure and give an idea about the extent of validity of the analysis,

Findings is the most important part of the report. The aim, however, is to provide insight into the data without drawing any conclusion at this stage. To make the presentation affractive and simple charts, graphs, diagrams and tables must be added. Findings that favor the hypothesis should be presented along with those which do not support the hypothesis.

Summary & Conslusions are the points of convergence of the study. While summary describes the brief statements of the major findings, the conclusions cover the inferences drawn thereform. Findings are factual information. But conclusions are inductive inferences. Sometimes, conclusions are presented in tabular form for presentational compactness. This tabular presentation also helps the future researchers to easily grasp the results.

Recommendations are of importance in case of applied research where the purpose of undertaking the research is to take some definite steps for development and or growth of any system. However, for pure research recommendations may draw the attention of the researchers on further studies.

Annexurces include detailed tables, complex tables, statistical tests and copies of questionaires etc.

References are the end notes that present the complete information on sources of secondary data, earther results, web sites etc, For referring to formal articles, books and edited volumes standard citation procedure is to be followed.

10.10. References

R. Bennet : *Management Research,* ILO.

D.R. Cooper and P.S. Schindler : *Business Research Methods,* Tata McGraw-Hill Publishing company limited.

R.J.R. Brent : *Roles of Research in the Management Pocess,* MSU Bnsiness Topics, 1976,pp 13-22.

P. Fitzroy : *Analytical Methods in Maketing,* Mcgraw Hill.

A. M. Goon, M. K. Gupta and B. Das Gupta : *Fundamentals of Statistics,* Vol I and II, World Press. *Kolkata*

P. E. Green and D. S. Tull : *Research for Marketing Decisions,* Prentice Hall of India. Private Limited.

R. A. Johnson and D. W. Wicherin : *Applied Multivariate Statistical Analysis,* Prentice Hall of India Private Limited.

Appendix–I

Decision theoretic approach

In two different ways one may examine the value of information. The first approach is search for perfect information and valuation of the perfect information in terms of effects on the decision-making process and the resultant expected pay off. The second approach is incorporation of sample information in the decision-making process and valuation of the resultant decision, again,in terms of expected pay off.

Let there be n states of the nature of the environment with E_j denoting the j-th state of the environment. Let there be m options or courses of actions with S_i denoting the i-th course of action. There will be, in total,mn situations with i-j-th combination as course of action S_i, against the state of the environment as E_j, where i ranges from 1 to m and j ranges from 1 to n. Let p_j be the chance of happening of the j-th state of the environment, p_j being positive with total sum of p_j 's as one. Under the above setup let us examine the choice of option and effect of that choice under perfect information. When perfect information is not available one may take help of expectation principle and opt for a course of action which yields the maximum expected pay off. It we denote by π_{ij} the pay off of i-th course of action when the state of environment is E_j, then the expected pay off of the i-th option will be given by

$$EP_i = \sum_{j=1}^{n} \pi_{ij} p_j$$

where EP_i stands for the expected pay off of the i-th option. It is natural to maximize the pay off and hence, if direct maximization of pay off is not possible, one must maximize the expected pay off. Thus, if i_o is the value of i for which EP_i is maximum for all choices of i = 1, 2,...., m, then S_{i_o} is the ideal choice of the course of action with expected pay off

$$EP_{i_o} = \text{Max } EP_i ,$$

maximization being done with respect to i, i=1,2,...,m..

In case an agency or an individual is willing to sell the perfect information the decision maker can incorporate perfect information in the decision making process. If we know that j-th state of the environment is going to happen, then the choice of option should be such that π_{ij} is maximized with respect to i. For each j we get the pay off as ($\underset{i}{\text{Max}} \ \pi_{ij}$) under the perfect information. Hence, the expected pay off under perfect information can be expressed as EPPI where

$$EPPI = \sum_{j=1}^{n} \left(\underset{1 \le j \le M}{\text{Max}} \ \pi_{ij} \right) p_j$$

The additional expected pay off is, obviously, due to perfect information and hence the expected value of perfect information (EVPI) can be expressed as

$$EVPI = EPPI - EP_{i_o}$$

$$= \sum_{j=1}^{n} p_j \left\{ \max_{1 \le i \le m} \pi_{ij} \right\} - \max_{1 \le i \le m} \sum_{j=1}^{n} p_j \pi_{ij}$$

Let us consider the following example to explain the above-mentioned theory. Suppose that a company has entered into the stage of maturity of the product life cycle curve and wants to undertake product modification. Let there be three possible courses of action. The company may go for major product modification, moderate product modification or minor product modification. There may be high acceptance of the modified version. The corresponding probability is 0.6. The chance of low acceptance of the modified version is 0.4. The pay off matrix per year in Rs 000 is given below :

Pay off matrix

	High acceptance (j=1)	Low acceptance (j=2)
Major Modification (i=1)	6000 (π_{11})	-1000 (π_{12})
Moderate Modification (i=2)	3000 (π_{21})	1500 (π_{22})
Minor Modification (i=3)	2000 (π_{31})	1200 (π_{32})

Thus, m = 3, n = 2 and there are 6(=3x2) possible situations. Let us denote

E_1 = High acceptance

E_2 = Low acceptance

S_1 = Major modification

S_2 = Moderate modification

S_3 = Minor modification

$p_1 = 0.6$

$p_2 = 0.4$.

Then, EP_1 = Expected pay off under S_1

$= 0.6\ (6000) + 0.4 \times (-1000)$

$= 3200$

EP_2 = Expected pay off under S_2

$= 0.6 \times (3000) + 0.4 \times (1500)$

$= 2400$

EP_3 = Expected pay off under S_3

= 0.6 × (2000) + 0.4 (1200)

= 1680

Max of EP_1, EP_2 and EP_3 is EP_1 =3200. Thus i_o = 1 and

EP_{i_o} = Max EP_i=3200.

Under perfect information when we know the high acceptance of modification we opt for option S_1 with maximum pay off

$\underset{1 \le i \le 3}{\text{Max}} \pi_{i1}$ = Maximum of 6000, 3000, and 2000 = 6000.

Under perfect information when we know the low acceptance of modification, we opt for option S_2 with maximum pay off

$\underset{1 \le i \le 3}{\text{Max}} \pi_{i2}$ = Maximum of -1000, 1500, and 1200 = 1500.

Then the EPPI will be given by

EPPI = 0.6 ×(6000)+0.4×(1500) = 4200

Hence, the expected value of perfect information can be obtained as

EVPI = EPPI - EP_{i_o}

= 4200 - 3200

= 1000.

At the most, one can spend Rs 10,00,000 for obtaining the perfect information as one gains in payoff, in expectation, by Rs 10,00,000.

Let us now consider the case of sample information available for incorporation in the decision making process. Sample information may be in terms of an indicator, correctness of which is stochastic in nature. Let there be q such indicators or reports. For the k-th indicator I_k let the probability of occurrence, given the j-th state of the environment, be α_{jk} where j varies between 1 to n and k varies between 1 to q. To incorporate the available sample information, say I_k, in the decision-making process let us follow Bayesian posterior analysis. Prior probability of happening of the state E_j was p_j. But given I_k this prior probability is to be modified by Bayesian rule. According to Bayes' theorem, the posterior probability $p_j^*(I_k)$ for happening of E_j given the report as I_k is a

$p_j^*(I_k)$= Probability of occurrence E_j given the report as I_k

= $P(E_j \mid I_k)$

= $P(E_j \cap I_k)/P(I_k)$

= $P(I_k \mid E_j)\ P(E_j)/\ \{\sum_{j=1}^{n} P(I_k \mid E_j)\ P(E_j)\}$, from Bayes' theorem.

$$= \alpha_{jk}p_j / \{\textstyle\sum_{j=1}^{n} \alpha_{jk}p_j\}$$

Also, from above $P(I_k) = \sum_{j=1}^{n} \alpha_{jk}\, p_j.$

Writing the positerior expected pay off as $EP_i^*\ (I_k)$ for the ith-course of action S_i given the report I_k, we have

$$EP_i\ ^*(I_k) = \textstyle\sum_{j=1}^{n} \pi_{ij}\ P_j\ ^*\ (I_k)$$

$$= \textstyle\sum_{j=1}^{n} \pi_{ij} \left[\alpha_{jk}p_j / \left\{\sum_{j=1}^{n} \alpha_{jk}p_j\right\}\right]$$

The corresponding optimum option will be the one that maximizes $EP_i\ ^*(I_k)$ with respect to i and the pay off will be

$$\underset{1\le i\le m}{\text{Max}}\ EP_i\ ^*(I_k).$$

But the chance of getting a report I_k is $P\ (I_k)$. Hence, the expected pay off under sample information (EPSI) is given by

$$EPSI = \textstyle\sum_{k=1}^{q} P(I_k)\ \underset{1\le i\le m}{\text{Max}}\ EP_i\ ^*(I_k).$$

$$= \textstyle\sum_{k=1}^{q} \left[\sum_{j=1}^{n} \alpha_{jk}p_j\right] \underset{1\le i\le m}{\text{Max}} \left[\sum_{j=1}^{n} \pi_{ij}[\alpha_{jk}p_j / \left\{\sum_{j=1}^{n} \alpha_{jk}p_j\right\}\right].$$

Then this addition in the expected pay off over the expected pay off without sample information can be attributed to the sample information. Thus, the expected value of sample information on (EVSI) can be expressed as

$$EVSI = EPSI - EP_{i_o}.$$

One may spend at the most EVSI amount for getting the sample information.

Let us refer to the example of product modification to explain this concept of EVSI. Let I_1 and I_2 be two reports/indicators where I_1 is a favorable report indicating high acceptance of the modified product and I_2 is an unfavorable report indicating low acceptance of the modified product. The information so provided in terms of indicators are expected to be highly reliable but may not be cent percent perfect. The following table describes the extent of reliability, i.e., the conditional probabilities (α_{jk} values) of the indicators.

Event	Chance of receiving the report	
j	I_1 (favorable)	I_2 (unfavorable)
	(k = 1)	(k = 2)
E_1 (j = 1)	$\alpha_{11} = 0.90$	$\alpha_{12} = 0.10$
E_2 (j = 2)	$\alpha_{21} = 0.05$	$\alpha_{22} = 0.95$

The calculation for posterior probabilities of happening of E_1 and E_2 is given below when the indicator is I_1 :

$$p_1^*(I_1) = p_1\alpha_{11} / \{ p_1\alpha_{11} + p_2\alpha_{21} \}$$
$$= (0.90)(0.60)/\{(0.90)(0.60)+(0.05)(0.40)\}$$
$$= 0.54/0.56$$
$$= 0.964$$

$$P_2^* (I_1) = p_2\alpha_{21} / \{ p_1\alpha_{11} + p_2\alpha_{21} \}$$
$$=(0.05)(0.40)/\{(0.90)(0.60)+(0.05)(0.40)\}$$
$$=0.02/0.56$$
$$=0.036$$

Also, $P(I_1) = 0.54 + 0.02 = 0.56$

The calculation for expected pay off under posterior probabilities is

$$EP_1^* (I_1) \sum_{j-1}^{2} \pi_{1j} p_j^* (I_1)$$
$$= (6000)(0.964) + (- 1000) (0.036)$$
$$= 5748$$

$$EP_2^* (I_1) = \sum_{j-1}^{2} \pi_{2j} \; p_j^* (I_1)$$
$$= (3000)(0.964) + (1500) (0.036)$$
$$= 2946$$

$$EP_3^* (I_1) = \sum_{j-1}^{2} \pi_{3j} \; p_j^* (I_1)$$
$$= (2000)(0.964)+(1200)(0.036)$$
$$= 1971.2$$

Thus, if the report is a favorable one, i.e., I_1, the best course of action will be S_1 with the corresponding expected pay off as

Maximum of {5748, 2946, 1971.2}

= 5748

Next, consider the report as I_2, i.e., an unfavorable report. The corresponding calculation for posterior probabilities of happenings of E_1 and E_2 is given below under I_2:

$$P_1^*(I_2) = p_1\alpha_{12} / \{ p_1\alpha_{12} + p_2\alpha_{22} \}$$
$$= (0.10)(0.60)/\{(0.10)(0.60)+(0.95)(0.40)\}$$
$$= 0.06/0.44$$
$$= 0.136$$

$$P_2^*(I_2) = p_2\alpha_{22} / \{ p_1\alpha_{12} + p_2\alpha_{22} \}$$
$$= (0.95)(0.40)/\{(0.10)(0.60)+(0.95)(0.40)\}$$
$$= 0.38/0.44$$
$$= 0.864.$$

Also, $P(I_2) = 0.06 + 0.38 = 0.44$.

The calculation for expected pay off under posterior probabilities is as follows :

$$EP_1^*(I_2) = \sum_{j=1}^{2} \pi_{1j}\ p_j^*(I_2)$$
$$= (6000)(0.136) + (-1000)(0.864)$$
$$= -48$$

$$EP_2^*(I_2) = \sum_{j=1}^{2} \pi_{2j}\ p_j^*(I_2)$$
$$= (3000)(0.136) + (1500)(0.864)$$
$$= 1704$$

$$EP_3^*(I_2) = \sum_{j=1}^{2} \pi_{3j} p_j^*(I_2)$$
$$= (2000)(0.136) + (1200)(0.864)$$
$$= 1308.8$$

Thus, if the report is an unfavorable one, i.e., I_2, the best course of action will be S_2 with the corresponding expected pay off as

$$\underset{1 \le i \le 3}{\text{Max}} EP^*(I_2) = \text{Maximum of } \{-48, 1704, 1308.8\}$$
$$= 1704$$

Thus, when the report is I_1,the corresponding strategy is S_1 with expected pay off as 5748 and when the report is I_2,the corresponding strategy is S_2 with expected pay off as 1704. Since the chance of occurrence of I_1 and I_2 are respectively 0.56 and 0.44 we have the expected pay off under sample information as

$$\begin{aligned} \text{EPSI} &= \sum_{k=1}^{2} P(I_k) \max_{1\le i\le m} EP_i^{*}(I_k) \\ &= P(I_1)\ (5748) + P(I_2)\ (1704) \\ &= 0.56\ (5748) + 0.44\ (1704) \\ &= 3968.64 \end{aligned}$$

Since $\max\limits_{1\le i\le 3} EP_i = 3200$, under prior analysis we have finally obtained the value of sample information in terms of expectation as

$$\begin{aligned} \text{EVSI} &= \text{EPSI} - \max_{1\le i\le 3} EP_i \\ &= 3968.64 - 3200 \\ &= 768.64 \end{aligned}$$

Thus, at the most one can spend Rs. 7,68,640 for obtaining the sample information as one gains, in expectation, by an amount of Rs. 7,68,640.

Appendix–2

Some Common Mistakes

In organising, storing and analysing of the data, the researcher often makes some mistakes that should be checked. A good researcher should maintain a check-list so that such mistakes could be avoided. Needless to say, no research can be foolproof, there would remain errors and omissions. However, there could be some errors which are very costly. A researcher would avoid such errors, by scrupulously checking whether the research contains such mistakes. In this section we highlight some of such mistakes. The list,however,is not exhaustive :

(1) *The data :* An empirical research must have the necessary data with the help of which the analysis would be done. Unless the researcher knows which are the necessary variables, the researcher would fail to address the issue properly. A researcher should maintain a checklist of the variables that are absolutely necessary to perform the analysis. If after a proper search, it is found that some attributes (some variables) remains missing, he would go back to the theory and try for a proxy variable that may serve the research purpose with respect to the atribute that has not been captured previously. One should, however, mention that such a round about way should be avcided, as far as practicable. One should try to identify the variables quite cautiously and solve every problem pertaining to field data before going for a research on the issue.

(2) *Tabulation and Compilation :* Quality of an empirical research depends on the quality of data. One may draw a wrong set of conclusions simply because there exits fundamental mistakes in the data. There should be scrutiny at every stage. As a precautionary step, a small sample should be taken up randomly to check the robustness of the data before tabulation and compilation of the data. The implication of the attributes with respect to which the set of information has been collected must also be understood properly so that the data can capture these attributes properly. The common mistake is that a variable is defined in terms of certain attribute and while getting the empirical mapping, the information is collected in terms of certain measures which do not contain the proper implication of the variable, as discussed in the theory.

(3) *Estimation & Testing :* Another common mistake that a researcher with his insufficient training in statistical theory,applies all the advanced statistical tools to judge the validity of his research hypothesis without being sure whether such tools could at all be applied to his data. Thus a researcher very often applies the tools of testing a statistical hypothesis on his data set even though the data had not been collected from a sample frame that adheres to the statistical theory of randomness. Often the researcher fails to realize that the robustness of statistical estimation depends on the robustness of sample frame which he might often be unable to comply. No tool of estimation or

testing a statistical hypothesis can be applied unless the basic data pertains to a sample, which has been randomly selected from a population.

In fact, the empirical research often fails to adhere to the very first step in statistical research. Before someone goes for collecting the primary data one must understand the 'Population' from which the 'Sample' is being drawn. Some research may even be based on complete enumeration of the elements in sample universe. If the study is based on complete enumeration there cannot be any 'Sampling Error'. Consequently, there does not exist any estimation and setting a 'confidence interval' for a 'statistic' to represent the true value of the parameter. All these are redundant because we are studying here the population itself. One should know that all such problems crop up as population on the basis of the results that we get by analysing a part of it which is known as the 'sample'. Assessing something about the whole from the knowledge of the part, is something like making a generalisation from some particular observation. This is a problem of induction. The theory of 'statistics' is basically a science related to induction. The theory would help one make an induction about the nature and behaviour of trhe population itself when a proper sample is chosen.

The whole theory maintains its usefulness by developing certain devices to understand whether the sample is truly drawn from the population. If suppose, we have taken a sample from the field study withot knowing the population, the errors in estimation will exist and this will be error due to sampling. If the sample has not been drawn on the basis of what is known as random sampling this error canont be assessed with the help of the statistical tools and the power of induction of this estimated values cannot be assessed properly. The idea is that the errors will exist even if the sample is drawn randomly (i.e. by adhering to the statistical principle of drawing the random sample). Naturally error related to a sample which is non-random, cannot be assessed with the help of these techniques. A researcher must know this point very clearly before he goes for applying these tools to his research.

(4) *Misconception about regression analysis :* Now-a-days, the empirical research very often utilises the tools of regression thanks to the development of sophisticated statistical packages and pathbreaking research in regression empirical exercise. But then this often gives rise to some problems the consequences of which might even be dangerous. A researcher may often be tempted to utilise such techniques without knowing the basic limitations of a so-called powerful tool and the field to which the tool should be applied. A researcher must remain aware of two basic facts : first a tool can reveal a truth which is not contained in the data set and secondly, a tool may lead to a wrong conclusion about the universe unless it is used properly. A high regression value does not necessarily mean thaτ the phenomenon has really been explained satisfactorily by the explanatory variable, unless a good and

sensible theory can justify the results. The empirical research is useless unless the research is founded only on a robust theoretical base. The researcher, while going for the regression analysis must also be aware of some basic facts about the theoretical basis of the tools that he is applying. For example, the researcher must know that he cannot go for any regression analysis if the degrees of freedom (defined as number of observation – the number of explanatory variables less one) is not sufficiently high.

The other important point is that the researcher must know what are the 'irrelevant variables' included in the model and what are the relevant variables that have been omitted from the model. The inclusion of irrelevant variables decreases the precision of R^2 (R^2 adjusted for degrees of freedom) and results in higher standard errors and lower 't' scores. The effect of omission of relevant variables creates bias and inconsistency in the estimated co-efficients of the included variables. The researcher should go back to the elements of theory behind the empirical analysis immediately and check the robustners of his research hypothesis.

(5) *Interpreting the results :* After going through all the statistical operations the researcher would take up the task of interpreting the results. The elementary mistake that the researcher must not make while interpreting the results, is that the interpretations must maintain consistency and aim at unfolding the inner truth of the data, step by step. A researcher must not over stretch the findings : he must also avoid wrong interpreation by defending the results in a wrong way. A result must not be defended by advocating wrong interpretations. A better course of action would be to rerun the regression with a proper software package that might suggest which variable should be included and which are to be dropped.

Interpreting a regression from the computer results often leads to mistakes that should be avoided by taking proper caution. Such mistakes often crop up because of insufficient exposure of the user to the theory of econometrics. It is advisable that the researcher, before taking up a regression analysis, will browse over the text book on econometrics and prepare a regression user's checklist and guide for regression users, so that he can fall back upon the econometric theory, while facing problems in interpreting the regression results derived by utilizing software packages. Finally, one must remember that the ethical researcher does not revise the findings to suit his purpose, the findings coming as the trugh after adhering to rigour in very step is something that an ethical researcher would never suppress or avoid.