TAMIL NADU OPEN UNIVERSITY

Research Methods

Master of Library and Information Science

MLS-05

RESEARCH METHODS

Master of Library and Information Science

MLS-05

Tamil Nadu Open University

SYLLABI-BOOK MAPPING TABLE

Research Methods

Syllabi	Mapping in Book
UNIT I **Elements of Research:** Definition, Characteristics - Types of Research: Historical, Fundamental/Pure, Applied, Scientific Method, Formulation of Problem - Sources of Identification, Factors Influencing Selection of Problems. Hypothesis: Meaning, Definition, Types - Formulation and Testing.	**Unit 1:** Elements of Research **(Pages 3-37)**
UNIT II **Research Methods and Techniques:** Research Methods - Survey, Census, Case Study, Experimental, Focused Groups. Method of Data Collection: Observation, Interview and Questionnaires - Advantages and Disadvantages. Sampling: Introduction - Definition of Universe, Population, Sample; Sampling Techniques - Probability and Non-probability.	**Unit 2:** Research Methods and Techniques **(Pages 39-87)**
UNIT III **Design of Research:** Definition and Importance of Research Design - Types: Exploratory, Description, Experimental - Content Analysis - Socio-metric Techniques, Constructive Typology, Projective Techniques, Statistical Survey, Evaluation Studies.	**Unit 3:** Design of Research **(Pages 89-109)**
UNIT IV **Data Analysis and Presentation:** Problem Measurement - Reliability, Validity, Measures of Central. Tendency - Average, Types of Averages - Measures of Dispersion - Correlation Analysis - Regression Analysis - Time Series - Measurement of Trends - Testing of Hypothesis: Statistical Testing - SPSS - Chi-square Test - Report Writing - Organization Report, Style and Presentation - Characteristics - Tables and Charts, Figures-style Manuals.	**Unit 4:** Data Analysis and Presentation **(Pages 111-224)**

CONTENTS

INTRODUCTION

NOTES

Research is the quest for knowledge or a systematic investigation in order to establish facts. It helps to solve problems and to increase knowledge. The basic aim of research is to discover, interpret and develop methods and systems to advance human knowledge on diverse scientific matters. There are different types of research, such as exploratory, descriptive and experimental. Exploratory research is done when few or no previous studies of the subject exist. Descriptive research is used to classify and identify the characteristics of a subject. Experimental research suggests or explains why or how something happens. Thus, one of the primary aims of research is to explain new phenomena and generate new knowledge.

Before conducting any research, a specific approach is to be decided; this is called research methodology. Research methodology refers to the way research can be conducted. It is also known as the process of collecting data for various research projects and helps to understand both the products as well as the process of scientific enquiry. A research process involves selection and formulation of a research problem, research design, sample strategy or sample design, as well as the interpretation and preparation of research report.

A few important factors in research methodology include the validity and reliability of research data and the level of ethics. A job is considered half done if the data analysis is conducted improperly. Formulation of appropriate research questions and sampling probable or non-probable factors are followed by measurement using survey and scaling techniques. A research design is a systematic plan for collecting and utilizing data so that the desired information can be obtained with sufficient accuracy. Therefore, research design is the means of obtaining reliable, authentic and generalized data. Research methodology is a very important function in today's business environment. There are many new trends in research methodology through which an organization can function in this dynamic environment.

This book, *Research Methods*, is divided into four units and has been written in a simple and easy-to-understand manner. Each unit begins with an 'Introduction' to the topic followed by an outline of the 'Unit Objectives'. Thereafter, the detailed content is presented in an organized manner. A 'Summary' along with a list of 'Key Terms' and a set of 'Questions and Exercises' is also provided at the end of each unit. A special feature of the book is 'Check Your Progress' questions, which are designed for effective recapitulation. Answers to these questions have been provided at the end of each unit. Relevant examples/illustrations have been included for better understanding of the topics.

UNIT 1 ELEMENTS OF RESEARCH

NOTES

Structure

1.0 INTRODUCTION

Research, in the layman's terms, means the search for knowledge. Scientific research is a systematic and objective way of seeking answers to certain questions that require inquiry and insight or that have been raised on a particular topic. The purpose of research, therefore, is to discover and develop an organized body of knowledge in any discipline. Research is a journey of discovery. It is a solution-oriented inquiry that must be objective and repeatable. It should inspire and guide further studies and should foster applications.

Research will provide practical benefits if it can provide advanced understanding of a discipline or suggest ways to handle some situations that we confront. We may sometimes wonder how researchers come up with ideas for a research project. They do so mostly when they face problems in the field. Since most researchers are engaged in social, human or health service programmes, they would automatically take up such issues that help them improve their fields of activity. Another source for research ideas is when researchers regularly update themselves by reading available literature and then extrapolating ideas from current researches in their respective fields of study. Government agencies and even private organizations often bring out 'request for proposals' for researchers. These are basically descriptions of problems that the agency would like researchers to work on. Sometimes, researchers come up with their own ideas of research which are influenced by their educational backgrounds, upbringing, culture, geographical influences, and so on.

NOTES

Every researcher should have the necessary training in gathering data, organizing materials suitably and engaging in field or laboratory work. He should also have the competence in using statistics for treating the data and the ability to interpret the collected data meaningfully. Research needs discipline, the right mental makeup, the ability to manage time effectively, objectivity, logical thinking, the capacity to evaluate the results of the research and the ability to carefully assess the findings of the research.

Research data allows people to make informed decisions by extrapolating the findings from the field or laboratory on to real life situations. This is the practical application of the findings generated by research. Research is also a way of preparing the mind to look at things in a fresh or different way. Out of such an orientation would come new and innovative observations about everyday events and occurrences. This is how originality comes about in research. Some of the most outstanding discoveries have been made in the most serendipitous manner. Some outstanding results have been obtained by researchers who had kept their minds open and free of clutter. This enabled them to see startlingly new connections.

This unit will give you the meaning of research and the definition of research. It will familiarize you with the possible types of research orientation and also explain the research process.

In this unit, you will also be taught about hypothesis. You will learn its meaning, need and nature. The unit will also discuss the criteria for hypothesis construction. Finally, you will also learn about the various types of statistical hypothesis.

1.1 UNIT OBJECTIVES

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

- Define research and explain its meaning
- Discuss the various types of research
- Examine the process of formulation of a problem
- Identify the sources of identification and list the factors influencing selection of social problems
- Discuss the meaning and types of hypothesis

1.2 DEFINITION AND CHARACTERISTICS

Research in common parlance refers to search for knowledge. One can also define research as a scientific and systematic search for pertinent information on a specific topic. In fact, research is an art of scientific investigation. According to the *Advanced Learner's Dictionary of Current English*, 'research is a careful investigation or enquiry, especially a thorough search for new facts in any branch of knowledge.' Well-known authors on research methodology, L.V. Redman and A.V.H. Mory (1923) defined research as a 'systematized effort to gain new knowledge.' Some people consider research as a voyage of discovery that involves movement from the known to the unknown.

Research in a technical sense is an academic activity. Renowned author Clifford Woody defined research as an activity that comprises defining and redefining problems, formulating a hypothesis; collecting, organizing and evaluating data; making deductions and reaching conclusions; and carefully testing the conclusions to determine if they support the formulated hypothesis. Eminent authors D. Slesinger and M. Stephenson, in the *Encyclopaedia of Social Sciences*, defined research as 'the manipulàtion of things, concepts or symbols for the purpose of generalizing, extending, correcting or verifying the knowledge, whether that knowledge aids in the construction of theory or in the practice of an art'. Research is, thus, an original contribution to the existing stock of knowledge making for its advancement.

Principles of Research

The basic principles of research include a systematic process to identify a question or problem, set forth a plan of action to answer the question or resolve the problem, and meticulously collect and analyse data. In conducting any research, it is crucial to choose the right method and design for a specific researchable problem. All research is different. However, the following factors are common to all good pieces of research:

- It is based on empirical data.
- It involves precise observations and measurements.
- It is aimed at developing theories, principles and generalizations.
- There are systematic, logical procedures involved.
- It is replicable.
- The findings of the research need to be reported.

Objectives of Research

The objective of any research is to find answers to questions through the application of scientific procedures. The main aim of any research is exploring the hidden or undiscovered truth. Even though each research study has a specific objective, the research objectives in general can be categorized into the following broad categories:

- **Exploratory or Formulative Research Studies:** These are aimed at gaining familiarity with a particular phenomenon or at gaining new insights into it.
- **Descriptive Research Studies:** These are aimed at accurately portraying the characteristics of a particular event, phenomenon, individual or situation.
- **Diagnostic Research Studies:** These studies try to determine the frequency with which something occurs.
- **Hypothesis Testing Research Studies:** These studies test a hypothesis and determine a causal relationship between the variables.

Characteristics and purpose of research

The following are the characteristics of research:

- It is a systematic and critical investigation into a phenomenon.
- It uses scientific methods.
- It is objective and logical.
- It requires emperial evidence.
- It focuses on finding answers to questions and solution to problems.

The following points will help in understanding the purpose of research:

- Research helps in extending the knowledge of human beings, the environment and natural phenomenon to others.

NOTES

- It brings out the information which is not developed fully during the ordinary course of life.
- It verifies the existing facts and identifies the changes in these existing facts.
- It helps in developing facts for critical evaluation.
- It analyses the interrelationship between variables and derives casual explanations.
- It develops new tools and techniques for those who study unknown phenomenon.
- It helps in planning and development.

1.2.1 Research: Significance and Approach

Research involves developing a scientific temperament and logical thinking. The significance of research-based answers can never be underestimated. The role of research is specially important in the fields of Economics, Business, Governance, and so on . Here research helps in finding solutions to problems encountered in real life. Decision-making is facilitated by applied research. Research is also of special significance in the operational and planning processes of business and industry. Here logical and analytical techniques are applied to business problems to maximize profits and minimize costs. Motivational research is another key tool in understanding consumer behaviour and health related issues. Responsible citizenship concerns can all be addressed through good research findings. Social relationships involving issues such as attitudes, interpersonal helping behaviour; environmental concerns such as crowding, crime, fatigue, productivity and; other practical issues are all capable of being addressed well by scientific research.

Social science research is extremely significant in terms of providing practical guidance in solving human problems of immediate nature.

Research is also important as a career for those in the field of academics. It could be a career option for professionals who wish to undertake research to gain new insights and idea generation. Research also fosters creative thinking and new theorizations.

Research for its own sake and for the sake of knowledge, and for solving different problems, all that is required is formal training in scientific methodology.

Approaches to Research

Quantitative approach and qualitative approach are the two basic approaches to research. These two paradigms are based on two different and competing ways of understanding the world. These competing ways of comprehending the world are reflected in the way research data is collected (for example, words versus numbers) and the perspective of the researcher (Perspectival versus Objective). The perspectives of the participants are very critical.

(i) **Quantitative Approach:** If there has been one overwhelming consensus among academic psychologists on a single point over the past few decades, it is that the best empirical research in the field is firmly grounded in quantitative methods. In this approach, data is generated in quantitative form, and then that data is subjected to rigorous quantitative analysis in a rigid and formal fashion. Inferential, experimental and simulation approaches are the sub-classifications of quantitative approach. Inferential approach to research focuses on survey research where databases are built studying samples of population and then these databases are used to infer characteristics or relationships in populations. In experimental approach, greater control is exercised over the research environment and often, some independent variables are controlled or manipulated to record their effects on dependent variables. In simulation approach, an artificial environment is constructed within which relevant data and information is generated. This way, the dynamic behaviours of a system are observed under controlled conditions.

(ii) Qualitative Approach: This approach to research is concerned with subjective assessment of attitudes, opinions and behaviour. Research in such a situation is a function of researcher's insight and impressions. Such an approach to research generates results either in non-quantitative form or in the forms which are not subjected to rigorous quantitative analysis.

NOTES

Table 1.1 provides us with types of research, methods employed and techniques used by these types of research.

***Table 1.1** Types of Research*

	Type	Methods	Techniques
1.	*Library Research*	(i) Analysis of historical records (ii) Analysis of documents	Recording of notes, content analysis, tape and film listening and manipulations, reference and abstract guides, content analysis.
2.	*Field Research*	(i) Non-participant direct observation	Observational behavioural scales, use of score cards, etc.
		(ii) Participant observation	Interactional recording, possible use of tape recorders, photographic techniques.
		(iii) Mass observation	Recording mass behaviour, interview using independent observers in public places.
		(iv) Mail questionnaire	Identification of social and economic background of respondents.
		(v) Opinionnaire	Use of attitude scales, projective techniques, use of goniometric scales.
		(vi) Personal interview	Interviewer uses a detailed schedule with open and closed questions.
		(vii) Focused interview	Interviewer focuses attention upon a given experience and its effects.
		(viii) Group interview	Small groups of respondents are interviewed simultaneously.
		(ix) Telephone survey	Used as a survey technique for information and for discerning opinion; may also be used as followup questionnaire.
		(x) Case study and life history	Cross-sectional collection of data for intensive analysis, longitudinal collection of data of intensive character.
3.	*Laboratory Research*	Small group study of random behaviour, play and role analysis	Use of audio-visual recording devices, use of observers, etc.

Methods versus Methodology

Research Methods: They refer to all the methods the researchers use while studying the research problems and while conducting research operations. In general, research methods can be categorized into the following three groups:

(i) The first group includes the methods that are concerned with the data collection.

(ii) The second group includes the statistical techniques needed for mapping relationships between the unknowns and the data.

(iii) The third group contains the methods necessary to evaluate the accuracy of the results obtained.

Research Methodology: It is the procedure that helps to systematically proceed in steps to solve a research problem. Research methodology is a broader concept that includes not the research methods but also the logic behind the research methods in the context of a particular research study; and it explains the reasons for using particular research methods and statistical techniques. Research mythology also defines how the data should be evaluated to get the appropriate results.

Applications of Research in Business Decisions

The discussion so far points out the role and significance of research in aiding business decisions. The question one might ask here is about the critical importance of research in different areas of management. Is it most relevant in marketing? Do financial and production decisions really need research assistance? Does the method or process of research change with the functional area? Figure 1.1 explains the complete research process.

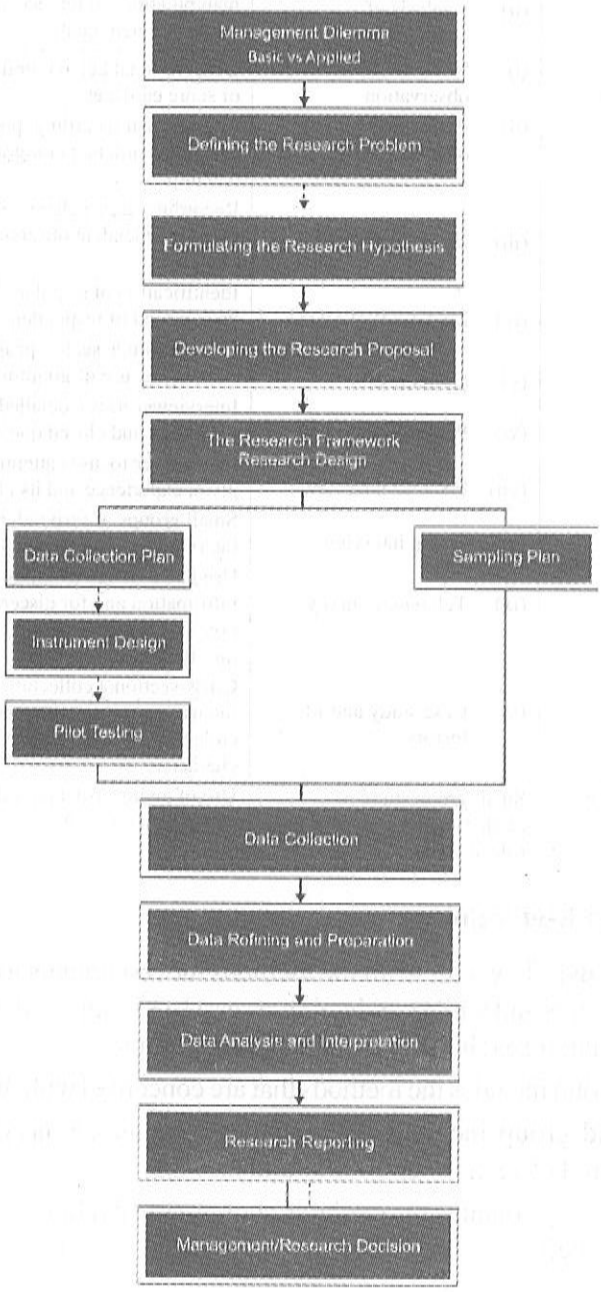

Management Dilemma
Basic vs Applied
Defining the Research Problem
Formulating the Research Hypothesis
Developing the Research Proposal
The Research Framework
Research Design
Data Collection Plan
Sampling Plan
Instrument Design
Pilot Testing
Data Collection
Data Refining and Preparation
Data Analysis and Interpretation
Research Reporting
Management/Research Decision

Fig. 1.1 The Process of Research

The answer to all the above questions is NO. Business managers in each field—whether human resources or production, marketing or finance—are constantly being confronted by problem situations that require effective and actionable decision-making. Most of these decisions require additional information or information evaluation, which can be best addressed by research. While the nature of the decision problem might be singularly unique to the manager, organization and situation, broadly for the sake of understanding, it is possible to categorize them under different heads.

Marketing Function

This is one area of business where research is the lifeline and is carried out on a vast array of topics, and is conducted both in-house by the organization itself and outsourced to external agencies. Broader industry- or product-category-specific studies are also carried out by market research agencies and sold as reports for assisting in business decisions. Studies like these could be:

- Market potential analysis; market segmentation analysis and demand estimation
- Market structure analysis which includes market size, players and market share of the key players
- Sales and retail audits of product categories by players and regions as well as national sales; consumer and business trend analysis—sometimes including short- and long-term forecasting

However, it is to be understood that the above-mentioned areas need not always be outsourced; sometimes they might be handled by a dedicated research or new product development department in the organizations. Other than these, an organization also carries out researches related to all four functions of marketing, such as:

- **Product Research:** This would include new product research; product testing and development; product differentiation and positioning; testing and evaluating new products, and packaging research; and brand research—including equity to tracks and imaging studies.
- **Pricing Research:** Price determination research; evaluating customer value; competitor pricing strategies, and alternative pricing models and implications.
- **Promotional Research:** Includes everything from designing of the communication mix to design of advertisements, copy testing, measuring the impact of alternative media vehicles and impact of competitors' strategy.
- **Place Research:** Includes locational analysis, design and planning of distribution channels, and measuring the effectiveness of the distribution network.

These days, with the onset of increased competition and the need to convert customers into committed customers, Customer Relationship Management (CRM), customer satisfaction, loyalty studies and lead user analysis are also areas in which significant research is being carried out.

1.2.2 Personnel and Human Resource Management

Human Resources (HR) and organizational behaviour is an area which involves basic or fundamental research as a lot of academic, macro level research may be adapted and implemented by organizations into their policies and programmes. Applied HR research by contrast is more predictive and solution oriented. Though there are a number of

NOTES

academic and organizational areas in which research is conducted, yet some key contemporary areas which seem to attract more research are as follows:

- **Performance Management:** Leadership analysis development and evaluation; organizational climate and work environment studies; talent and aptitude analysis and management; and organizational change implementation, management and effectiveness analysis.
- **Employee Selection and Staffing:** This includes pre and on-the-job employee assessment and analysis; and staffing studies.
- **Organizational Planning and Development:** Culture assessment—either organization specific or the study of individual, and merged culture analysis for mergers and acquisitions; and manpower planning and development.
- **Incentive and Benefit Studies:** These include job analysis and performance appraisal studies; recognition and reward studies, hierarchical compensation analysis; and employee benefits and reward analysis, both within the organization and industry best practices.
- **Training and Development:** These include training need gap analysis; training development modules; monitoring and assessing impact; and effectiveness of training.
- **Other Areas:** Other areas include employee relationship analysis; labour studies; negotiation and wage settlement studies; absenteeism and accident analysis; turnover and attrition studies; and work-life balance analysis.

Critical success factor analysis and employer branding are some emerging areas in which HR research is being carried out. The first is a participative form of management technique, developed by American Organizational theorist John F. Rockart (1981) in which the employees of an organization identify their critical success factors, and help in customizing and incorporating them in developing the mission and vision of their organization. The idea is that a synchronized objective will benefit both the individual and the organization, and which will lead to a commitment and ownership on the part of the employees. Employer branding is another area which is being actively investigated as the customer perception (in this case, it is the internal customer, i.e., the employee) about the employer or the employing organization has a strong and direct impact on his intentions to stay or leave. Thus, this is a subjective qualitative construct which can have hazardous effect on organizational effectiveness and efficiency.

Financial and Accounting Research

The area of financial and accounting research is so vast that it is difficult to provide a pen sketch of the research areas. In this section, we are providing just a brief overview of some research topics:

- **Asset Pricing, Corporate Finance and Capital Markets:** The focus here is on stock market response to corporate actions (IPOs or Initial Public Offerings, takeovers and mergers), financial reporting (earnings and firm specific announcements) and the impact of factors on returns, e.g., liquidity and volume.
- **Financial Derivatives and Interest Rate and Credit Risk Modelling:** This includes analysing interest rate derivatives, development and validation of corporate credit rating models and associated derivatives; analysing corporate-decision making and investment risk appraisal.

- **Market Based Accounting Research:** Analysis of corporate financial reporting behaviour; accounting-based valuations; evaluation and usage of accounting information by investors and evaluation of management compensation schemes.
- **Auditing and Accountability:** This includes both private and public sector accounting studies, analysis of audit regulations; analysis of different audit methodologies; governance and accountability of audit committees.
- **Financial Econometrics:** This includes modelling and forecasting in volatility, risk estimation and analysis.
- **Other Areas:** Other related areas of investigation are in merchant banking and insurance sector, and business policy and economics areas.

Considering the nature of the decision required in this area; the research is a mix of historical and empirical research. Behavioural finance is a new and contemporary area in which, probably, for the first time, subjective and perceptual variables are being studied for their predictive value in determining consumer sentiments.

NOTES

Production and Operation Management

This area of management is one in which quantifiable implementation of the research results takes on huge cost and process implications. Research in this area is highly focused and problem specific. The decision areas in which research studies are carried out are as follows:

- **Operation Planning:** These include product/service design and development; resource allocation and capacity planning.
- Demand forecasting and decision analysis.
- **Process Planning:** Production scheduling and material requirement management; work design planning and monitoring.
- Project management and maintenance management studies.
- Logistics and supply chain, and inventory management analysis.
- **Quality Estimation and Assurance Studies:** These include Total Quality Management (TQM) and quality certification analysis.

This area of management also invites academic research which might be macro and general but helps in developing technologies, such as JIT (Just-In-Time) technology and EOQ (Economy Order Quantity)—an inventory management model), which are then adapted by organizations for optimizing operations.

Cross-Functional Research

Business management being an integrated amalgamation of all these and other areas sometimes requires a unified thought and approach to research. These studies require an open orientation where experts from across the disciplines contribute to and gain from the study. For example, an area, such as new product development, requires the commitment of the marketing, production and consumer insights team to exploit new opportunities. Other areas requiring cross functional efforts are as follows:

- Corporate governance and ethics—the role of social values and ethics, and their integration into a company's working is an area that is of critical significance to any organization.

NOTES

- Technical support systems, enterprise resource planning systems, knowledge management, and data mining and warehousing are integrated areas requiring research on managing coordinated efforts across divisions.
- Ecological and environmental analysis; legal analysis of managerial actions; human rights and discrimination studies.

1.3 TYPES OF RESEARCH

Although research is a vast subject and is difficult to categorize, it can be classified according to its intent or as per the methods of study.

On the basis of intent, research can be classified as follows:

- **Pure research:** It is done only for the sake of knowledge. The intention is not to apply it in regular practice. Pure research is also called basic or fundamental research. It is not focused on specific problems, but instead it focuses on the extension of knowledge. New theory or refinements of an existing theory are developed with the help of pure research. It lays the foundation for applied research. It helps in finding the critical factors in a problem. It helps in generating alternative solutions and choosing the best one amongst them.
- **Applied research:** When real-life problems require some solution and decision-making, applied research is carried out. This means that applied research is problem oriented and action directed. It brings immediate and practical results; for example, marketing research carried on for identifying customer habits to purchase something. Though it is problem oriented and action directed, it can contribute to the development of theoretical knowledge by leading to the discovery of new facts.
- **Exploratory research:** It is also called formulative research. When a researcher has no knowledge or little knowledge about an unfamiliar problem, they do a preliminary study. The objective of this research is to generate new ideas, gather new facts, precise formulation of problem and increasing familiarity of the researcher to the unfamiliar problem. Renowned theorist Katz conceptualizes two levels of exploratory research. At the first level is the discovery of significant variables in particular situations; at the second, the discovery of relationship between variables.
- **Descriptive research:** In this research, facts are analysed in detail for clear understanding. This research is simple in nature and in its application. It is more specific than exploratory research. It focuses on the problem under study and also aims at a classification of the range of elements comprising the subject matter of study. Empirical observations are used to conceptualize the problems and facts. It highlights methods of data collection and interpretation.
- **Diagnostic research:** It is just like descriptive research but with a different focus. It is aimed towards in depth approaches to reach the basic casual relations of a problem and possible solutions for it. Prior knowledge of the problem is required for this type of research. Problem formulation, defining the population correctly for study purposes, proper methods for collecting accurate information, correct measurement of variables, statistical analysis and tests of significance are essential in diagnostic research.

Check Your Progress

1. State the objectives of research.
2. What is the qualitative approach to research?

NOTES

The classification of research can be done as per methods of study in the following manner:

- **Fundamental:** This type of research is mainly concerned with identifying certain important principles in a specific field. It intends to find out information that has a broad base of application. Examples of fundamental research are John Robinson's imperfect competition theory in Economics and Maslow's hierarchy of needs theory in motivation.
- **Applied:** This type of research aims at finding a solution to an immediate problem, faced by a society or an industrial organization. It is supposed to discover a solution to some basic practical problems. Applied research suggests corrective methods to minimize a social or business problem.
- **Historical:** Historical research studies the social effects of the past that may have given rise to current situations, i.e., past incidents are used to analyse the present as well as the future conditions. The study of the current state of Indian labour based on past labour union movements in the Indian economy to formulate the Indian Labour Policy is an example of this type of research.
- **Formulative or exploratory:** It helps examine a problem with suitable hypothesis. This research, on social science, is mainly significant for clarifying concepts and innovations for further researches. The researchers are mainly concerned with the principles of developing hypothesis and testing with statistical tools.
- **Experimental:** The experimental type of research enables a person to calculate the findings, employ the statistical and mathematical devices, and measure the results, thus, quantified.
- **Ex post facto:** This type of research is the same as experimental research, which is conducted to deal with the situations that occur in or around an organization. Examples of such a research are market failure of an organization's product being researched later, and research into the causes for a landslide in the country.
- **Case study:** This method undertakes intensive research that requires a thorough study of a particular unit.

1.4 FORMULATION OF A PROBLEM

The crux of the scientific approach to identifying and pursuing a research path is to identify the 'what', i.e., what is the exact research question to which you are seeking an answer. The second important thing is that the process of arriving at the question should be logical and follow a line of reasoning that can lend itself to scientific enquiry. However, we would like to sound a note of caution here. The challenge for a business manager is not only to identify and define the decision problem; the bigger challenge is to convert the decision into a research problem that can lend itself to scientific enquiry. As Powers *et al.* (1985) have put it, 'Potential research questions may occur to us on a regular basis, but the process of formulating them in a meaningful way is not at all an easy task.' One needs to narrow down the decision problem and rephrase it into researchable terms. Well-known authors Bonnie L. Yegidis and Robert W. Weinbach (1991) have also referred to the complexity of phrasing the decision in research terms.

The second concern in formulating business research problems is the fact that more often than not, managers become aware of problems, seek information and arrive at decisions under conditions of bonded rationality. A concept formalized by organizational theorists James C. March and Herbert A. Simon which implies that managers do not

Check Your Progress

3. When is applied research carried out?
4. What do you mean by ex post facto research?

NOTES

always work and take decisions in a perfectly rational sequence. The model says that information search or problem recognition phase like any other behaviour has to be motivated. Unless the manager is driven by present levels of dissatisfaction or by high expected value of outcomes, the process does not start. The next implication of the model is that in most instances, a manager does not have access to complete and perfect information. Further, the manager might try to seek reasonably convenient and quick information that meets minimal rather than optimal standards.

Scientific Thought

The real requirement is not the identification of the decision situation but applying a thought process that can take a panoramic view of the business decision. One needs to reason logically and effectively to cover all the probable alternatives that need to be addressed in order to arrive at any concrete basis for decision-making. This reasoning approach could be deductive or inductive or a combination of both.

Deductive thought: This kind of logic is a culmination, a conclusion or an inference drawn as a consequence of certain reasoned facts. The reasons cited have to be real and not a figment of the researcher's judgement, and second, the deductions or conclusions must essentially be an outcome of the same reasons. For example, if we summarize for Ms Dubey's problem that:

All well-executed projects have well-integrated teams. (Reason 1)

The ABC project has many shortfalls. (Reason 2)

The ABC project team is not a very cohesive and integrated team. (Inference)

A note of caution here is that the above could be only two probable reasons; this inference is justified if we look at only these facts. Thus, unless all probable reasons have been isolated and identified, the nature of the inference is incomplete.

Inductive thought: On the other end of the continuum is inductive thought. Here, there is no strong and absolute cause and effect between the reasons stated and the inference drawn. Inductive reasoning calls for generating a conclusion that is beyond the facts or information stated. In the same example of the ABC project, we might begin by asking a question, 'What is the reason for the ABC project not being executed on time?' A probable answer could be that the project team is not making a coordinated effort. Again, this is only one explanation and there could be other inductive hypotheses as well, for example:

The vendors and suppliers are ineffective in maintaining and managing the raw material and supplies.

or

The local authorities are extremely corrupt. At each stage, they deliberately put an official spoke in the wheel and do not let the next phase of the project to be achieved till their 'rightful' share is negotiated and delivered.

or

The workers union in the area is very strong and is on a go-slow call which prevents the execution of work on time.

Thus, the fact of the matter is that inductive thought draws assumptions and hypothesis which could explain the phenomena observed and yet there could be other propositions which might explain the event as well as the one generated by the manager/researcher. Each one of them has a potential truth in it. However, we have more confidence in some over the others, so we select them and seek further information in order to get confirmation.

In practice, scientific thought actually makes use of both inductive and deductive reasoning in a chronological order. We might question the phenomena by an inductive hypothes, and then collect more facts and reasons to deduct that the hypothesized conclusion is correct.

Defining the Research Problem

The first and the most important step of the research process is to identify the path of enquiry in the form of a research problem. It is like the onset of a journey; in this instance, the research journey, and the identification of the problem gives an indication of the expected result being sought. A research problem can be defined as a gap or uncertainty in the decision-makers' existing body of knowledge which inhibits efficient decision-making. Sometimes, it may so happen that there might be multiple reasons for these gaps, and identifying one of these and pursuing its solution might be the problem. According to professor Fred N. Kerlinger, a well respected positivist, 'If one wants to solve a problem, one must generally know what the problem is. It can be said that a large part of the problem lies in knowing what one is trying to do.' The defined research problem might be classified as simple or complex (Hicks, 1991). Simple problems are those that are easy to comprehend, and their components and identified relationships are linear and easy to understand, e.g., the relation between cigarette smoking and lung cancer. Complex problems, on the other hand, talks about interrelationship between antecedents and subsequently with the consequential component. Sometimes the relation might be further impacted by the moderating effect of external variables as well, e.g., the effect of job autonomy and organizational commitment on work exhaustion, at the same time considering the interacting (combined) effect of autonomy and commitment. This might be further different for males and females. These kinds of problems require a model or framework to be developed to define the research approach.

Thus, the significance of a clear and well-defined research problem cannot be overemphasized, as an ambiguous and general issue does not lend itself to scientific enquiry. Even though different researchers have their own methodology and perspective in formulating the research topic, a general framework which might assist in problem formulation is given below.

Problem identification process

The problem recognition process invariably starts with the decision-maker and some difficulty or decision dilemma that he/she might be facing. This is an action oriented problem that addresses the question of what the decision-maker should do. Sometimes, this might be related to actual and immediate difficulties faced by the manager (applied research) or gaps experienced in the existing body of knowledge (basic research). The broad decision problem has to be narrowed down to information oriented problem which focuses on the data or information required to arrive at any meaningful conclusion. Given in Figure 1.2 is a set of decision problems and the subsequent research problems that might address them.

Management decision problem

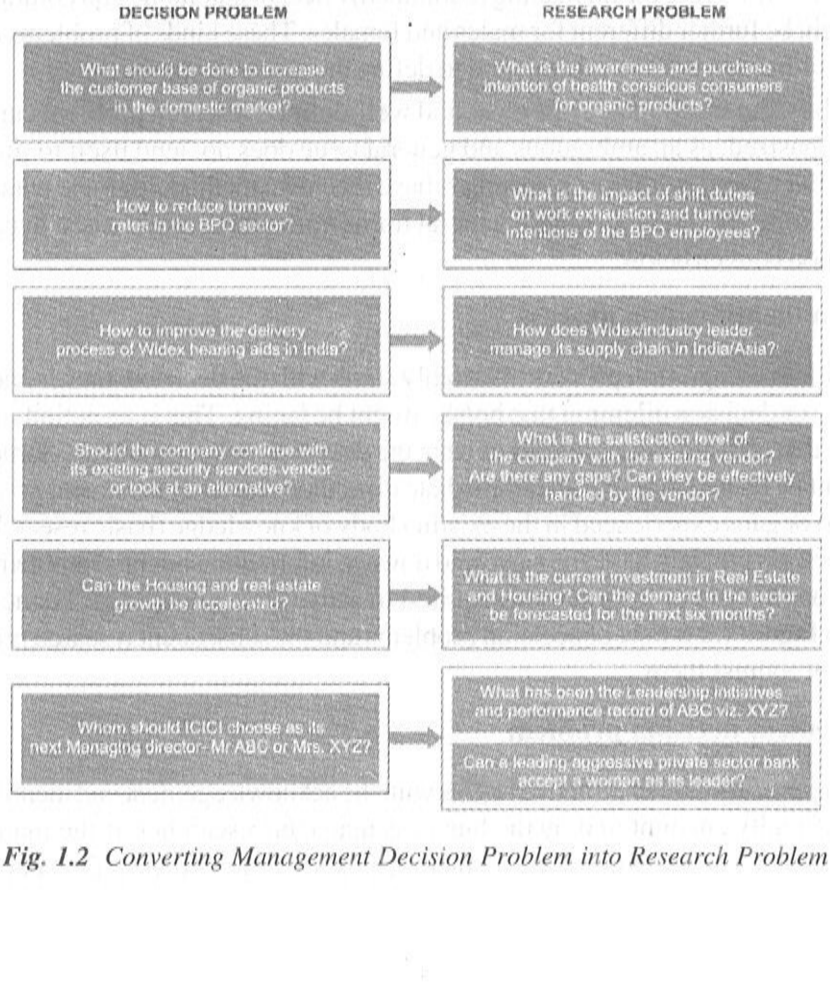
The entire process explained above begins with the acknowledgement and identification of the difficulty encountered by the business manager/researcher. If the manager is skilled enough and the nature of the problem requires to be resolved by him or her alone,

NOTES

the problem identification process is handled by him or her else he or she outsources it to a researcher or a research agency. This step requires the author to carry out a problem appraisal, which would involve a comprehensive audit of the origin and symptoms of the diagnosed business problem. For illustration, let us take the first problem listed in the Figure 1.2. An organic farmer and trader in Uttarakhand, Nirmal farms, wants to sell his organic food products in the domestic Indian market. However, he is not aware if this is a viable business opportunity and since he does not have the expertise or time to undertake any research to aid in the formulation of the marketing strategy, he decides to outsource the study.

Discussion with subject experts

The next step involves getting the problem in the right perspective through discussions with industry and subject experts. These individuals are knowledgeable about the industry as well as the organization. They could be found both within and outside the company. The information on the current and probable scenario required is obtained with the assistance of a semi-structured interview. Thus, the researcher must have a predetermined set of questions related to the doubts experienced in problem formulation. It should be remembered that the purpose of the interview is simply to gain clarity on the problem area and not to arrive at any kind of conclusions or solutions to the problem. For example, for the organic food study, the researcher might decide to go to food experts in the Ministry for Food and Agriculture or agricultural economists, or retailers stocking health food as well as doctors and dieticians. This data, however, is not sufficient in most cases, while in other cases, accessibility to subject experts might be an extremely difficult task as they might not be available. The information should, in practice, be supplemented with secondary data in the form of theoretical as well as organizational facts.



DECISION PROBLEM	RESEARCH PROBLEM
What should be done to increase the customer base of organic products in the domestic market?	What is the awareness and purchase intention of health conscious consumers for organic products?
How to reduce turnover rates in the BPO sector?	What is the impact of shift duties on work exhaustion and turnover intentions of the BPO employees?
How to improve the delivery process of Widex hearing aids in India?	How does Widex/industry leader manage its supply chain in India/Asia?
Should the company continue with its existing security services vendor or look at an alternative?	What is the satisfaction level of the company with the existing vendor? Are there any gaps? Can they be effectively handled by the vendor?
Can the Housing and real estate growth be accelerated?	What is the current investment in Real Estate and Housing? Can the demand in the sector be forecasted for the next six months?
Whom should ICICI choose as its next Managing director- Mr ABC or Mrs. XYZ?	What has been the Leadership initiatives and performance record of ABC viz. XYZ? Can a leading aggressive private sector bank accept a women as its leader?

Fig. 1.2 Converting Management Decision Problem into Research Problem

Review of existing literature

A literature review is a comprehensive compilation of the information obtained from published and unpublished sources of data in the specific area of interest to the researcher. This may include journals, newspapers, magazines, reports, government publications, and also computerized databases. The advantage of the survey is that it provides different perspectives and methodologies to be used to investigate the problem, as well as identify possible variables that may need to be investigated. Second, the survey might also uncover the fact that the research problem being considered has already been investigated and this might be useful in solving the decision dilemma. It also helps in narrowing the scope of the study into a manageable research problem that is relevant, significant and testable.

Once the data has been collected from different sources, the researcher must collate all information together in a cogent and logical manner instead of just listing the previous findings. This documentation must avoid plagiarism and ensure that the list of earlier studies is presented in the researcher's own words. The logical and theoretical framework developed on the basis of past studies should be able to provide the foundation for the problem statement.

The reporting should cite clearly the author and the year of the study. There are several internationally accepted forms of citing references and quoting from published sources. The *Publication Manual of the American Psychological Association* (2001) and the *Chicago Manual of Style* (1993) are academically accepted as referencing styles in management.

To illustrate the significance of a literature review, given below is a small part of a literature review done on organic purchase.

> Research indicates organic is better quality food. The pesticide residue in conventional food is almost three times the amount found in organic food. Baker *et al.* (2002) found that on an average, conventional food is more than five times likely to have chemical residue than organic samples. Pesticides toxicity has been found to have detrimental effects on infants, pregnant women and general public (National Research Council, 1993; Ma *et al.*, 2002; Guillete et al., 1998). Major factors that promote growth in organic market are consumer awareness of health, environmental issues and food scandals (Yossefi and Willer, 2002).

This paragraph helps justify the relevance and importance of organic versus non-organic food products as well as identify variables that might contribute positively to the growth in consumption of organic products.

Organizational analysis

Another significant source for deriving the research problem is the industry and organizational data. In case the researcher/investigator is the manager himself/herself, the data might be easily available. However, in case the study is outsourced, the detailed background information of the organization must be compiled, as it serves as the environmental context in which the research problem has to be defined. It is to be remembered at this juncture that the organizational context might not be essential in case of basic research, where the nature of study is more generic.

This data needs to include the organizational demographics—origin and history of the firm; size, assets, nature of business, location and resources; management philosophy and policies as well as the detailed organizational structure, with the job descriptions.

Qualitative survey

Sometimes the expert interview, secondary data and organizational information might not be enough to define the problem. In such a case, an exploratory qualitative survey might be required to get an insight into the behavioural or perceptual aspects of the problem. These might be based on small samples and might make use of focus group discussions or pilot surveys with the respondent population to help uncover relevant and topical issues which might have a significant bearing on the problem definition.

In the organic food research, focused group discussions with young and old consumers revealed the level of awareness about organic food and consumer sentiments related to purchase of more expensive but a healthy alternative food product.

Management research problem

Once the audit process of secondary review and interviews and survey is over, the researcher is ready to focus and define the issues of concern, that need to be investigated further, in the form of an unambiguous and clearly-defined research problem. Once again it is essential to remember that simply using the word 'problem' does not mean there is something wrong that has to be corrected; it simply indicates the gaps in information or knowledge base available to the researcher. These might be the reason for his inability to take the correct decision. Second, identifying all possible dimensions of the problem might be a monumental and impossible task for the researcher. For example, the lack of sales of a new product launch could be due to consumer perceptions about the product, ineffective supply chain, gaps in the distribution network, competitor offerings or advertising ineffectiveness. It is the researcher who has to identify and then refine the most probable cause of the problem and formalize it as the research problem. This would be achieved through the four preliminary investigative steps indicated above.

Last, the researcher must be able to isolate the underlying issues from the symptoms of the problem. For example, in the organic food study, the manufacturer has an outlet in an up market area in Delhi, and is constantly doing some attractive sales promotion but there is no substantial increase in sales. Here, the real problem is lack of awareness and motivation on the part of the consumer about the benefits of organic food. Thus, the low sales are primarily a consequence of lack of awareness and purchase intention.

To address the problems of clarity and focus, we need to understand the components of a well-defined problem. These are as follows:

The unit of analysis: The researcher must specify in the problem statement the individual(s) from whom the research information is to be collected and on whom the research results are applicable. This could be the entire organization, departments, groups or individuals. In the organic food study, for example, the retailer who has to be targeted for stocking the product as well as the end consumer could be the unit of analysis. Thus, the information required for decision might sometimes require investigation at multiple levels.

Research variables: The research problem also requires identification of the key variables under the particular study. To carry out an investigation, it becomes imperative to convert the concepts and constructs to be studied into empirically testable and observable variables. A variable is generally a symbol to which we assign numerals or values. A variable may be dichotomous in nature, that is, it can possess only two values such as male-female or customer–non-customer. Values that can only fit into prescribed number of categories are discrete variables, for example, very important (1) to very unimportant (5). There are still others that possess an indefinite set, e.g., age, income and production data.

Variables can be further classified into five categories, depending on the role they play in the problem under consideration.

- *Dependent variable:* The most important variable to be studied and analysed in research study is the dependent variable (DV). The entire research process is involved in either describing this variable or investigating the probable causes of the observed effect. Thus, this in essence has to be reduced to a measurable and quantifiable variable. For example, in the organic food study, the consumer's purchase intentions and the retailers stocking intentions as well as sales of organic food products in the domestic market could all serve as the dependent variable.

  A financial researcher might be interested in investigating the Indian consumers' investment behaviour, post the recent financial slow down. In another study, the HR head at Cognizant Technologies would like to study the organizational commitment and turnover intentions of short and long tenure employees in the company.

  Hence, as can be seen from the above examples, it might be possible that in the same study, there might be more than one dependent variable.

- *Independent variable:* Any variable that can be stated as influencing or impacting the dependent variable is referred to as an independent variable (IV). More often than not, the task of the research study is to establish the causality of the relationship between the independent and the dependent variable(s). The proposed relations are then tested through various research designs.

  In the organic food study, the consumers' attitude towards healthy lifestyle could impact their organic purchase intention. Thus, attitude becomes the independent and intention the dependent variable. Another researcher might want to assess the impact of job autonomy and role stress on the organizational commitment of the employees; here job autonomy and role stress are independent variables.

- *Moderating variables:* Moderating variables are the ones that have a strong contingent effect on the relationship between the independent and dependent variables. These variables have to be considered in the expected pattern of relationship as they modify the direction as well as the magnitude of the independent–dependent association. In the organic food study, the strength of the relation between attitude and intention might be modified by the education and the income level of the buyer. Here, education and income are the moderating variables (MVs).

  In a consulting firm, the management is looking at the option of introducing flexi-time work schedule. Thus, a study might need to be taken to see whether there will be an increase in productivity of each individual worker (DV) subsequent to the introduction of a flexi-time (IV) work schedule.

  In real time situations and actual work settings, this proposition might need to be revised to take into account other impacting variables. This second independent variable might need to be introduced because it has a significant contribution on the stated relationship. Thus, we might like to modify the above statement as follows:

There will be an increase in productivity of each individual worker (DV) subsequent to the introduction of a flexi-time (IV) work schedule, especially amongst women employees (MV).

There might be instances when confusion might arise between a moderating variable and an independent variable.

Consider the following situation:

Proposition 1: Turnover intention (DV) is an inverse function of organizational commitment (IV), especially for workers who have a higher job satisfaction level (MV).

While another study might have the following proposition to test:

Proposition 2: Turnover intention (DV) is an inverse function of job satisfaction (IV), especially for workers who have a higher organizational commitment (MV).

Thus, the two propositions are studying the relation between the same three variables. However, the decision to classify one as independent and the other as moderating depends on the research interest of the decision-maker.

- *Intervening variables:* An intervening variable (IVV) has a temporal connotation to it. It generally follows the occurrence of the independent variable and precedes the dependent variable. Research theorist Tuckman (1972) defines it as 'that factor which theoretically affects the observed phenomena but cannot be seen, measured, or manipulated; its effects must be inferred from the effects of the independent variable and moderator variables on the observed phenomenon.'

  For example, in the previous case, there is an increase in job satisfaction (IVV) of each individual worker, subsequent to the introduction of a flexi-time (IV) work schedule, which eventually affects the individual's productivity (DV), especially amongst women employees (MV). Another example would be the introduction of an electronic advertisement for the new diet drink (IV), which will result in increased brand awareness (IVV). This, in turn, will impact the first quarter sales (DV).This would be significantly higher amongst the younger female population (MV).
- *Extraneous variables:* Besides the moderating and intervening variables, there might still exist a number of extraneous variables (EVs) which could affect the defined relationship but might have been excluded from the study. These would most often account for the chance variations observed in the research investigation. For example, a tyrannical boss; family pressures or nature of the industry could impact the flexi-time impact, but since these would be applicable to individual cases, they might not heavily impact the direction of the findings. However, in case the effect is substantial, the researcher might try to block their effect by using an experimental and a control group. (This concept will be discussed later in the section on experimental designs.)

At this stage, we can clearly distinguish between the different kinds of variables discussed above. An independent variable is the prime antecedent condition which is qualified as explaining the variance in the dependent variable; the intervening variable follows the occurrence of the independent variable and may in turn impact the dependent variable; the moderating variable is a contributing variable which might impact the defined relationship; the extraneous variables are outside the domain of the study and responsible for chance variations, but in some instances, their effect might need to be controlled.

Theoretical foundation and model building

Having identified and defined the variables under study, the next step requires operationalizing the stated relationship in the form of a theoretical framework. This is an outcome of the problem audit conducted prior to defining the research problem; it can be best understood as a schema or network of the probable relationship between the identified variables. Another advantage of the model is that it clearly demonstrates the expected direction of the relationships between the concepts. There is also an indication of whether the relationship would be positive or negative.

This step, however, is not mandatory as sometimes the objective of the research is to explore the probable variables that might explain the observed phenomena (DV) and the outcome of the study helps to theorize and propose a conceptual model.

The theoretical framework, once formulated, is a powerful driving force behind the research process and ought to be comprehensively developed. It requires a thorough understanding of both theory and opinion.

Given below is a predictive model for turnover intentions developed to explain the high rate of attrition amongst BPO professionals. Once validated, it is of course possible to test it in different contexts and differing respondent population.

The turnover intention model

The proposed model to predict turnover intention is specified as mentioned below:

$$TI = f(WE, OC, A, MS, TWE) \qquad ...(1)$$

Where, TI = Turnover intention

WE = Work exhaustion

OC = Organizational commitment

A = Age

MS = Marital status

TWE = Total work experience

The theoretical construct of work exhaustion is influenced by Perceived Workload (PWL), Fairness of Reward (FOR), Job Autonomy (JA) and Work Family Conflict (WFC) [Adapted from Ahuja, Chudoba and Kacman, 2007] . This can be mathematically written as:

$$WE = f(PWL, FOR, JA, WFC) \qquad ...(2)$$

Similarly, Organizational Commitment depends upon Job Autonomy, Work–Family Conflict, Fairness of Reward and Work Exhaustion (WE) [Adapted from—Ahuja, Chudoba and Kacman, 2007]. Therefore, this can be stated mathematically as:

$$OC = f(JA, WFC, FOR, WE) \qquad ...(3)$$

The model is diagrammatically represented in Figure 1.3.

The formulated framework has been explained verbally as a verbal model. The flowchart of the relationship between independent and intervening variables has been demonstrated in graphical form as a graphical model and the same have been also reduced to three mathematical equations specifying the relationship between the same in the form of a mathematical model. What needs to be understood is that all three compliment each other and are basically representatives of the same framework.

NOTES

Statement of research objectives

Next, the research question(s) that were formulated need to be broken down and spelt out as tasks or objectives that need to be met in order to answer the research question.

Based on the framework of the study, the researcher has to numerically list the thrust areas of research. This section makes active use of verbs such as 'to find out', 'to determine', 'to establish' and 'to measure' so as to spell out the objectives of the study. In certain cases, the main objectives of the study might need to be broken down into sub-objectives which clearly state the tasks to be accomplished.

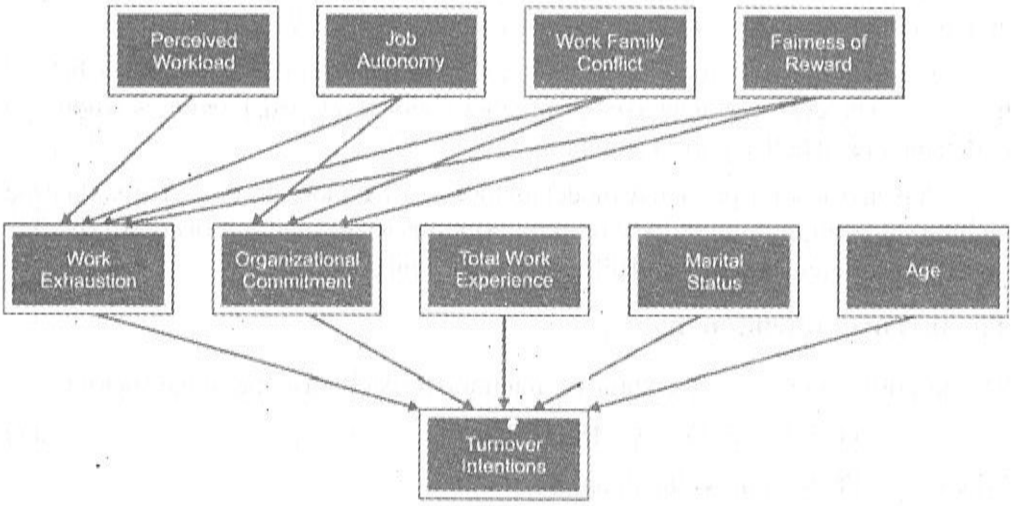


Fig. 1.3 Proposed Model for Turnover Intention

In the organic food research, the objectives and sub-objectives of the study were as follows:

1. *To study the existing organic market*: This would involve:
   - Categorizing the organic products available in Delhi into grain, snacks, herbs, pickles, squashes and fruits and vegetables
   - Estimating the demand pattern of various products for each of the above categories
   - Understanding the marketing strategies adopted by different players for promoting and propagating organic products
2. *Consumer diagnostic research:* This would entail:
   - Studying the existing consumer profile, i.e., perception and attitudes towards organic products and purchase and consumption patterns
   - Studying the potential customers in terms of consumer segments, level of awareness, perception and attitude towards health and organic products
3. *Opinion survey:* To assess the awareness and opinions of experts such as doctors, dieticians and chefs in order to understand organic consumption and propagation
4. *Retail market:* This would involve:
   - Finding the gap between demand and supply for existing retailers
   - Forecasting demand estimates by considering the existing as well as potential retailers

1.4.1 Sources of Identification and Factors Influencing Selection of Social Problems

Research is a systematic, objective and scientific study done to collect the research data related to current problems. Research enables a company to exploit the opportunities available in the environment. If the research has not been planned systematically, it is difficult for a firm to achieve the desired objectives. The research process can be described as follows:

Defining the problem and objectives
↓
Developing data source
↓
Data collection
↓
Data analysis
↓
Presenting the findings

Defining the Problem and Objectives

The defined objectives should be SMART.

S – Specific
M – Measurable
A – Attainable
R – Realistic
T – Time Bound

The first step in research is definition of a problem. Selection of a problem is itself a difficult decision. The success of research depends on right selection of the problem. If the problem has not been identified in right manner, it is very difficult for the researcher to find the right solution of the issue.

The following sources can help a researcher identify the research problems:

- **Brainstorming:** A researcher can learn new dimensions of a problem by discussing ideas, thoughts, facts and data with other people who have knowledge of the subject.
- **Consultations:** By consulting others, the researcher identifies new dimensions of a problem.
- **Daily experience:** Daily experience develops the evaluative thinking in a researcher.
- **Academic experience:** Academic experience helps the researcher develop critical thinking towards the happenings.
- **Field situations:** Research is done because every field today is developing and, hence, changing constantly.

Objectives of formulating the problem

'A well-defined problem is half solved.' This statement reveals the fact that how important it is to formulate or define a problem. The primary objective of a research is to collect

NOTES

relevant data and analyse this data to get answers to the research problem. This means that the success of research depends upon accuracy of data and information required for investigation. Right formulation solves this purpose. Proper definition of the problem, its analysis, identifying questions for data collection and the formulation of hypothesis to be tested are key steps which are required for formulation of the problem. Once the exact and accurate data is known to the researcher, he can plan the other steps without wastage of resources. Thus, right formulation of the problem gives the right direction to the entire research and limits the approach towards pertinent facts out of the large variety of facts. It helps us in determining statistical methods to be used for research.

Criteria for formulation of the problem

Criteria for formulating one problem out of identified problems can be grouped into internal criteria and external criteria. These are discussed in detail in the following section:

(i) Internal criteria

Internal criteria consist of the following:

- **Interest of the researcher:** The problem should be from the subject of interest of the researcher and can be challenging to him. Without interest in the problem, it becomes very difficult for the researcher to sustain continuity in the research. A researcher's interest depends on his experience, educational background, sensitivity, and so on.
- **Own resources of the researcher:** Research requires a lot of money. If the researcher does not have enough money and he is unable to manage external finance, the researcher should not go in for research. Moreover, time resource is more important than money. Research requires more time and, hence, it should be utilized properly.
- **Competence of the researcher:** A mere interest in research is not enough. The researcher must be competent enough to plan and carry out a study of the problem. He should have sufficient knowledge of the subject matter, relevant methodology and statistical procedures.

(ii) External criteria

External criteria consist of the following:

- **Potential for research:** Very narrow or extremely vague problems should be avoided. In order to be researched, a problem must be one for which observation or other data collection in real world can provide the answer.
- **Importance and urgency:** Issues that require investigation are unlimited but available research efforts are very limited. Therefore, relative importance and significance of the problem is required. Important and urgent issues should be given priority over an unimportant one.
- **Novelty of the problem:** A problem on which a lot of research work has been done should not be considered for research as there are fewer chances of throwing light on any new factor.
- **Feasibility:** Novelty of the problem is not sufficient if it is not feasible to conduct the study on problem in real world, i.e., it should contain facts which can be analysed. Even if the problem is novel, we should make a small feasibility study first and proceed only after this if study allows.

- **Facilities:** A well-equipped library, proper guidance in data analysis, and so on, are basic facilities which are required to carry on any research.
- **Research personnel:** Availability of adequate research personnel including investigators and research officers is very important for data collection, which is a major issue in many developing countries like India.

Techniques involved in formulating the problem

Defining a research problem properly and clearly is a crucial part of the research study and must, in no case, be done hurriedly. The technique for this purpose involves undertaking the following steps:

- **Statement of the problem in a general way:** The problem should be carefully worded. The problem statement should indicate nature of the problem and intention of researcher.
- **Understanding the nature of the problem:** The best way to understand the nature of the problem is to discuss with those who have prior experience in the same kind of research. This will ensure that the origination of problem and the objectives in view are correct. If the marketer has stated the problem himself, he should consider all the facts that induced him to make a general statement concerning the problem.
- **Developing the ideas through discussions:** Many new ideas are developed by discussing them with others. This discussion provides useful information for research. Discussion is done with those people who have enough experience in the concerned field.
- **Rephrasing the research problem:** After going through the given four steps, the researcher gets a clear idea about the environment in which the problem is to be studied. Now rephrasing the problem into analytical or operational terms is not a difficult task. Through rephrasing, the researcher puts the research problem in as specific terms as possible so that it may become operationally viable and may help in the development of working hypothesis.

1.5 HYPOTHESIS: MEANING AND TYPES

According to Theodorson, 'a hypothesis is a tentative statement asserting a relationship between certain facts'. Kerlinger describes it as 'a conjectural statement of the relationship between two or more variables'.

Need of hypothesis

The following points help in understanding the importance of hypothesis:

- A hypothesis is a proposal intended to explain a fact or an observation.
- A hypothesis specifies the sources of data which shall be studied and in what context they shall be studied.
- It determines the data needs.
- Hypothesis suggests the type of research which is likely to be the most appropriate.
- A hypothesis contributes to the development of the theory.

Check Your Progress

5. Define the term research problem.
6. What do you mean by literature review?
7. What is the first step in research?
8. What is the best way to understand the nature of a research problem?

Nature of hypothesis

Hypothesis is more useful when stated in precise and clearly defined terms. A good hypothesis implies that hypothesis which fulfils its intended purposes and is up to the mark. The following are some important points to be kept in mind:

(a) A good hypothesis should be stated in the simplest possible terms. It is also called the principle of the economy or business. It should be clear and precise.

(b) A good hypothesis is in agreement with the observed facts. It should be based on original data derived directly.

(c) It should be so designed that bits test will provide an answer to the original problem which farms the primary purpose of the investigation.

(d) Hypothesis should state relationship between variables, if, it happens to be a rational hypothesis.

1.5.1 Criteria for Hypothesis Construction

Once the investigative questions are set up for each of the objectives, the researcher should identify the anticipated or possible answers to the investigative questions. A survey of related theories and earlier studies and discussions with co-scientists will facilitate this process. He, then, should write down those answers as appropriate types of hypotheses—descriptive, relational or causal—as the case may be. He should evaluate these tentative hypotheses in terms of the characteristics of a good hypothesis and refine and record them into logical and testable hypotheses, keeping in mind the rules given below:

Rules for constructing hypotheses

According to research theorist Smith, there are certain rules for constructing good hypotheses. These are as follows:

(i) Search for variable measurements with the most quantitative characteristics available: Precise quantitative measurements are more critical in testing theory than qualitative characteristics. Another theorist Hage gives four techniques to search for and create variables from non-variable concepts. First, the researcher can search for implied dimensions underlying non-variable concepts. Campbell provides a good example of this method in the study of the non-variable concept 'social group'. He identifies four underlying dimensions of degrees of proximity, similarity, perceived common fate and perceived spatial pattern. Second, one can create new variables by comparing conceptual synonyms or analogies. Price used the synonym technique in his study of organizational measurement concepts like 'participation in decision making', 'organization control', 'power' and 'influence' refer to the degree of organizational centralization. Third, one can search the literature for rarely occurring associations between phenomena. The cognitive dissonance theory in sociology got its start in this manner. Fourth, one can generate new variables through ordering many concepts from more or less abstract extension of the research variables and their applicability in the new order of society, say virtual or network organization.

(ii) Make the variable scale properties explicit by stating all of the variable's mutually exclusive and totally inclusive categories by degrees. For example, a variable like 'income' may be categorized into (1) up to ₹ 5000 per month, (2) ₹ 5001 to 10000, (3) ₹ 10001 to 20000, (4) ₹ 20001 and above.

(iii) Describe the means used to sort observations into your variables categories in sufficient detail so that your methods may be evaluated and replicated by others. 'Personality disintegration' is a good example of a poorly operationalised variable. It is an unreliable measure and cannot be replicated.

(iv) Always consider alternative operations that might be more appropriate for a given variable.

(v) Analyse variables through their relationships. Non-ratio uni- or multi-variable distribution is arbitrary, since it has no intrinsic lower boundary.

(vi) Link two or more formal propositions through a shared independent or dependent variable where possible. For example, from following concrete observations:

(a) 'Married persons are less likely than unmarried persons to commit suicide.'

(b) 'Married persons with children are less likely than married persons without children to commit suicide.'

The following abstract formal hypotheses may be inductively produced:

(i) 'Sucide rates vary directly with the degree of individualism.'

(ii) 'Suicide rates vary indirectly with the degree of group cohesion.'

1.5.2 Types of Statistical Hypothesis

In context of statistical analysis, we generally consider two types of hypothesis:

- Null Hypothesis
- Alternative Hypothesis

When comparing the superiority of both the methods A and B, if we assume that both the methods are equally good, then the assumption is known as 'null hypothesis'. On the other hand, if we consider method A to be better, it is alternative hypothesis. These may be, symbolically presented as:

Null Hypothesis = Ho

Alternative Hypothesis = Ha

Difference between Proposition, Hypothesis and Theory

(i) A proposition is a logical statement of relationship between two or more variables which has, generally, been confirmed by empirical research. (A proposition should be distinguished from a hypothesis which is a logical statement of an assumed relationship between two or more variables which must be empirically tested, replicated and elaborated before being accepted as confirmed.)

(ii) Proposition is a broad statement drawn from a theory, whereas a hypothesis takes this one step further and formulates a more specific statement that is empirically testable. Proposition states a relationship between two concepts, and a hypothesis operationalizes this relationship and puts it in an empirically testable form.

(iii) The term 'hypothesis' is used to refer to an explanation of things that occur. In some cases, it may refer to a simple guess. In other instances, it may be a well-developed set of propositions that are crafted to explain the detailed workings of some occurrence or occurrences. One definition states specifically that it is the antecedent to a conditional proposition.

(iv) The hypothesis is formed and tested within the scientific process. One may develop the hypothesis while observation is occurring, but that may also be considered premature. The act of observation (outside of experimentation) may actually present opportunity to disprove a hypothesis. The hypothesis though is necessarily well defined and inclusive of details. This allows for accurate testing. It also, in many cases, distinguishes it from a theory.

(v) The term 'theory' is one of a rather scientific nature, but of a less limited nature. Some uses can refer to explanations of occurrences; some do include usage as referencing a simple guess. There is more though. Theory is used to refer to a branch of study that is focused on the general and conceptual, as compared to the practical and the applied of the same subject. It is significant that a theory is conjectural in nature.

(vi) A hypothesis is a proposed explanation for something. We call it a theory when that hypothesis has been tested with considerable evidence. As a result, a theory is usually a much larger set of statements than a hypothesis because a theory can grow with every new piece of evidence it explains. In other words, a theory can explain far more than the phenomenon it originally was proposed to explain.

(vii) A hypothesis attempts to answer questions by putting forth a plausible explanation that has yet to be rigorously tested. A theory, on the other hand, has already undergone extensive testing by various scientists and is generally accepted as being an accurate explanation of an observation. This does not mean the theory is correct; only that current testing has not yet been able to disprove it, and the evidence as it is understood, appears to support it. A theory will often start out as a hypothesis — an educated guess to explain observable phenomenon. The more a hypothesis is tested and holds up, the better accepted it becomes as a theory.

1.6 FORMULATION AND TESTING

A claim or hypothesis about the values or population parameters is known as the Null Hypothesis and is written as H_0. In the case of the above discussed situation, our assumption that a butler is innocent would form the null hypothesis and would be stated as follows:

$$H_0 = \text{The butler is innocent}$$

This hypothesis is then tested with the available evidence and the decision is made whether to accept this hypothesis or reject it. If this hypothesis is rejected, then we accept the alternate hypothesis which is that the butler is not innocent. This alternate hypothesis is denoted as H_1 and is stated as:

$$H_1 = \text{The butler is not innocent}$$

The process involves testing of the null hypothesis. If the null hypothesis is rejected, then the alternate hypothesis is accepted. It should be noted that the acceptance of the alternate hypothesis does not mean that it is correct. It simply means that there is not enough evidence to be reasonably sure that the null hypothesis is acceptable.

As already explained, there are two types of errors that can be used in making decisions regarding accepting or rejecting the null hypothesis. The first type of error, known as Type I error, is used when the null hypothesis is rejected even if it is true. The second type of error, known as Type II error, is used when a null hypothesis is accepted even if it was not true and should have been rejected.

Check Your Progress

9. State the importance of hypothesis.
10. Define proposition.

In statistical hypothesis testing and decision-making about the values of population parameters as defined by the sample statistics, the null hypothesis asserts that there is no true difference between the sample statistics and the corresponding population parameter under consideration and if indeed there is any visible difference, it is considered to be due to natural fluctuations in sampling.

To conclude we say that:

- *Null Hypothesis* H_0: An assertion about the population parameter that is being tested by the sample results
- *Alternate Hypothesis* H_1: A claim about the population parameter that is accepted when the null hypothesis is rejected
- *Type I Error*: An error made in rejecting the null hypothesis, when in fact it is true
- *Type II Error:* An error made in accepting the null hypothesis, when in fact it is false

Type I error is denoted by α (Alpha) and is expressed as a probability of rejecting a true hypothesis. It is also known as the level of significance. $1 - \alpha$ expresses the level of confidence. For example, $\alpha = 0.05$ means that the confidence level is 95% or 0.95.

Type II error is denoted by β (Beta) and is expressed as the probability of accepting a false hypothesis. It is desirable to have the β value as low as possible for its value reflects the power of the test being performed and a low β value indicates that the test of significance is powerful and reliable. (Type I and Type II errors have been discussed in detail in 1.6.2).

1.6.1 Procedure for Hypothesis Testing

The general procedure for hypothesis testing consists of the following steps:

1. ***State the Null Hypothesis as well as the Alternate Hypothesis:*** This means stating the assumed value of the population parameter which is to be tested. For example, suppose we want to test the hypothesis that the average IQ of our college students is 130. Then this would become our null hypothesis and the alternate hypothesis would be that this average IQ is not 130. These statements are expressed as follows:

$$H_0 : \mu = 130$$
$$H_1 : \mu \neq 130$$

2. ***Establish a Level of Significance Prior to Sampling:*** The level of significance signifies the probability of committing Type I error α and is generally taken as equal to 0.05, which really means that after the hypothesis has been tested and a decision is made, we will still be making an error in rejecting the null hypothesis when in fact it is true, i.e., 5% of the time. Sometimes the value α is established as 0.01, but it is at the discretion of the investigator to select its value, depending upon the sensitivity of the study.
3. ***Determine a Suitable Test Statistic:*** This means the choice of appropriate probability distribution to use with the particular available information under consideration. The normal distribution using the Z-score table or the *t*-distribution is most often used.
4. ***Define the Rejection (Critical) Regions:*** The critical region will be established on the basis of the choice of the value of the level of significance α. For example, if we select the value of $\alpha = 0.05$, and we use the standard normal

NOTES

distribution as our test statistic for testing the population parameter μ, then as we have discussed before, the difference between the assumption of null hypothesis, assumed value of this population parameter and the value obtained by the analysis of sample results is not expected to be more than $\pm 1.96\sigma_{\bar{x}}$ at $\alpha = 0.05$. This relationship is shown in the Figure 1.4.

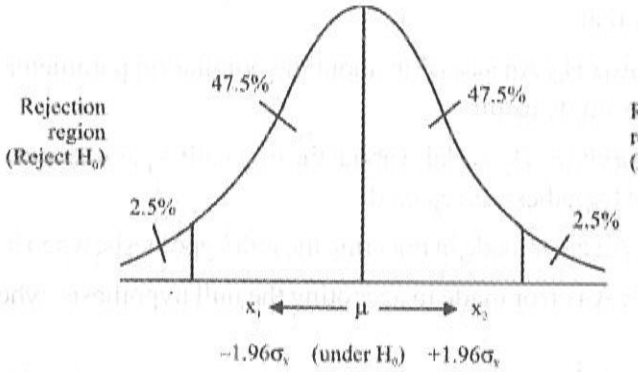

47.5%
47.5%
Rejection region (Reject H₀)
Rejection region (Reject H₀)
2.5%
2.5%
x₁
μ
x₂
−1.96σₓ (under H₀) +1.96σₓ

Fig. 1.4 Rejection Region

In the above figure, if the sample $\overline{X}$ statistic falls within $1.96\sigma_{\bar{x}}$ of the assumed value of μ under the assumption of null hypothesis H_0, then we accept the null hypothesis as being correct at 95% confidence level (or 0.05 level of significance). The difference between $\bar{X}$ and μ which may be any value between X_1 and μ or X_2 and μ is considered to be accidental or due to chance element and is not considered significant enough or real enough to reject null hypothesis, so that for all practical purposes the value of $\bar{X}$ is considered equal to μ even though $\bar{X}$ can have any value between X_1 and X_2 as shown above. However, if the value of $\bar{X}$ falls beyond X_2 on the upper side or beyond X_1 on the lower side, then this difference between the values of $\bar{X}$ and μ would be considered significant and it will lead to rejection of null hypothesis. Since 5% of the time, this difference between the values of $\bar{X}$ and μ would be significant with 2.5% of the time $\bar{X}$ being too far above μ (beyond X_2) and 2.5% of the time being too far below μ (below X_1), the area of rejection will be on both sides of the mean extending into the tail sections of the curve. This area of rejection is known as the *critical region.*

5. ***Data Collection and Sample Analysis:*** This involves the actual collection and computation of the sample data. A sample of the pre-established size *n* is collected and the estimate of the population parameter is calculated. This estimate is the value of the test statistic. For example, if we are testing a hypothesis about the value of population mean μ, then the test statistic would be the sample mean $\bar{X}$. Then we test this statistic to check whether it falls in the critical region or in the acceptance region. For example, if we want to test for the average IQ of the college students to be 130, then in that case, we have to see that our population mean μ must be tested. We take a random sample of a given size *n* and calculate its mean $\bar{X}$, and then test it to see if the value of this $\bar{X}$ falls in the area of acceptance or in the area of rejection at a given level of significance.

6. ***Making the Decision:*** Before the statistical decision is made, a decision rule must be established. Such decision rule will form the basis on which the null hypothesis will be accepted or rejected. This decision rule is really a formal statement of the obvious purpose of the test. For example, this rule could be stated as follows:

   Accept the null hypothesis if the value of sample statistic $\bar{X}$ falls within the area of acceptance, otherwise reject the null hypothesis.

   Based upon this established decision rule, a decision can be made whether to accept or reject the null hypothesis.

1.6.2 Committing Errors: Type I and Type II

Types of Errors: There are two types of errors in statistical hypothesis, which are as follows:

- **Type I Error:** In this type of error, you may reject a null hypothesis when it is true. It means rejection of a hypothesis, which should have been accepted. It is denoted by α (alpha), and is also known alpha error.
- **Type II Error:** In this type of error, you are supposed to accept a null hypothesis when it is not true. It means accepting a hypothesis, which should have been rejected. It is denoted by β (beta), and is also known as beta error.

Type I error can be controlled by fixing it at a lower level. For example, if you fix it at 2%, then the maximum probability to commit Type I error is 0.02. However, reducing Type I error has a disadvantage when the sample size is fixed as it increases the chances of Type II error. In other words, it can be said that both types of errors cannot be reduced simultaneously. The only solution of this problem is to set an appropriate level by considering the costs and penalties attached to them or to strike a proper balance between both types of errors.

In a hypothesis test, a Type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

In a hypothesis test, a Type II error occurs when the null hypothesis H_0 is not rejected when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; that is H_0: there is no difference between the two drugs on average. A Type II error would occur if it were concluded that the two drugs produced the same effect, that is, there is no difference between the two drugs on average, when in fact, they produced different ones.

In how many ways can we commit errors?

We reject a hypothesis when it may be true. This is Type I Error.

We accept a hypothesis when it may be false. This is Type II Error.

NOTES

We accept a hypothesis when it is true. We reject a hypothesis when it is false.

	Accept H_0	Reject H_0
H_0 True	Accept True H_0 Desirable	Reject True H_0 Type I Error
H_1 False	Accept False H_0 Type II Error	Reject False H_0 Desirable

The level of significance implies the probability of Type I error. A 5% level implies that the probability of committing a Type I error is 0.05. A 1% level implies 0.01 probability of committing Type I error.

Lowering the significance level and, hence, the probability of Type I error is good, but unfortunately, it would lead to the undesirable situation of committing Type II error.

To Sum Up:

- **Type I Error:** Rejecting H_0 when H_0 is True.
- **Type II Error:** Accepting H_0 when H_0 is False.

Note: The probability of making a Type I error is the level of significance of a statistical test. It is denoted by α.

Where, α = Prob. (Rejecting H_0 / H_0 True)

$1 - \alpha$ = Prob. (Accepting H_0 / H_0 True)

The probability of making a Type II error is denoted by β.

Where, β = Prob. (Accepting H_0 / H_0 False)

$1 - \beta$ = Prob. (Rejecting H_0 / H_0 False) = Prob. (The test correctly rejects H_0 when H_0 is false.)

$1 - \beta$ is called the power of the test. It depends on the level of significance α, sample size n and the parameter value.

1.6.3 Null and Alternative Hypotheses

Hypothesis is usually considered as the principal instrument in research. The basic concepts regarding the testability of a hypothesis are as follows:

Null Hypothesis and Alternative Hypothesis

In the context of statistical analysis, while comparing any two methods, the following concepts or assumptions are taken into consideration:

- **Null Hypothesis:** While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good, it is called null hypothesis. It is also known as statistical hypothesis and is symbolized as H_0.
- **Alternate Hypothesis:** While comparing two different methods, regarding their superiority, wherein stating a particular method to be good or bad as compared to the other one, it is called alternate hypothesis. It is symbolized as H_1.

Comparison of Null Hypothesis with Alternate Hypothesis

Following are the points of comparison between null hypothesis and alternate hypothesis:

- Null hypothesis is always specific, while alternate hypothesis gives an approximate value.
- The rejection of Null hypothesis involves great risk, which is not in the case of alternate hypothesis.

Null hypothesis is more frequently used in statistics than alternate hypothesis because it is specific and is not based on probabilities.

The hypothesis to be tested is called the null hypothesis and is denoted by H_0. This is to be tested against other possible states of nature called alternative hypothesis. The alternative is usually denoted by $H_{1.}$

The null hypothesis implies that there is no difference between the statistic and the population parameter. To test whether there is no difference between the sample mean X and the population μ, we write the null hypothesis as:

H_0: $\bar{X} = \mu$

The alternative hypothesis would be:

H_1: $\neq \mu$

This means $> \mu$ or $< \mu$. This is called a two-tailed hypothesis.

The alternative hypothesis H_1: $> \mu$ is right tailed.

The alternative hypothesis H_1: $< \mu$ is left tailed.

These are one-sided or one-tailed alternatives.

Note 1: The alternative hypothesis H_1 implies all such values of the parameter, which are not specified by the null hypothesis H_0.

Note 2: Testing a statistical hypothesis is a rule, which leads to a decision to accept or reject a hypothesis.

A one-tailed test requires rejection of the null hypothesis when the sample statistic is greater than the population value or less than the population value at a certain level of significance.

1. We may want to test if the sample mean exceeds the population mean μ. Then the null hypothesis is:

   H_0: $> \mu$

2. In the other case, the null hypothesis could be:

   H_0: $< \mu$

Each of these two situations leads to a one-tailed test and has to be dealt with in the same manner as the two-tailed test. Here the critical rejection is on one side only, right for $> \mu$ and left for $< \mu$. Both the Figures 1.5 and 1.6 here show a 5% level of test of significance.

For example, a minister in a certain government has an average life of 11 months without being involved in a scam. A new party claims to provide ministers with an average life of more than 11 months without scam. We would like to test if, on the average, the new ministers last longer than 11 months. We may write the Null Hypothesis H_0: = 11 and Alternative Hypothesis H_1: > 11.

NOTES

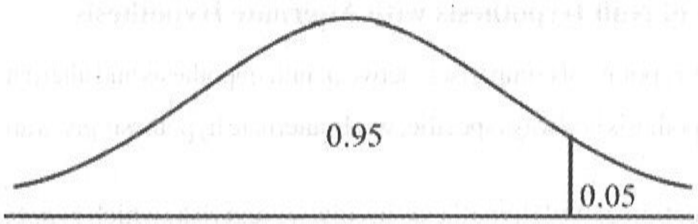


Fig. 1.5 H_0: $\overline{X} > \mu$

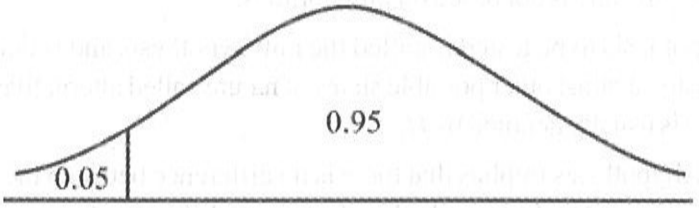


Fig. 1.6 H_0: $< \mu$

1.7 SUMMARY

- Research in common parlance refers to search for knowledge. One can also define research as a scientific and systematic search for pertinent information on a specific topic.
- The basic principles of research include a systematic process to identify a question or problem, set forth a plan of action to answer the question or resolve the problem, and meticulously collect and analyse data.
- Quantitative approach and qualitative approach are the two basic approaches to research. These two paradigms are based on two different and competing ways of understanding the world.
- Human Resources (HR) and organizational behaviour is an area which involves basic or fundamental research as a lot of academic, macro level research may be adapted and implemented by organizations into their policies and programmes.
- The crux of the scientific approach to identifying and pursuing a research path is to identify the 'what', i.e., what is the exact research question to which you are seeking an answer. The second important thing is that the process of arriving at the question should be logical and follow a line of reasoning that can lend itself to scientific enquiry.
- The first and the most important step of the research process is to identify the path of enquiry in the form of a research problem. It is like the onset of a journey, in this instance the research journey, and the identification of the problem gives an indication of the expected result being sought.
- The problem recognition process invariably starts with the decision-maker and some difficulty or decision dilemma that he/she might be facing. This is an action oriented problem that addresses the question of what the decision-maker should do.

Check Your Progress

11. What do you mean by null hypothesis?
12. Define alternate hypothesis.

- Research is a systematic, objective and scientific study done to collect the research data related to current problems. Research enables a company to exploit the opportunities available in the environment.
- The primary objective of a research is to collect relevant data and analyse this data to get answers to the research problem. This means that the success of research depends upon accuracy of data and information required for investigation. Right formulation solves this purpose.
- A good hypothesis implies that hypothesis which fulfils its intended purposes and is up to the mark.
- In context of statistical analysis, we generally consider two types of hypothesis: Null Hypothesis and Alternative Hypothesis.
- In statistical hypothesis testing and decision-making about the values of population parameters as defined by the sample statistics, the null hypothesis asserts that there is no true difference between the sample statistics and the corresponding population parameter under consideration, and if indeed there is any visible difference, it is considered to be due to natural fluctuations in sampling.
- Before the statistical decision is made, a decision rule must be established. Such decision rule will form the basis on which the null hypothesis will be accepted or rejected.
- There are two types of errors in statistical hypothesis, which are: Type I Error and Type II Error.

1.8 KEY TERMS

- **Data collection:** It is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research question hypotheses, and evaluate outcomes.
- **Pilot testing:** It is a small-scale trial, where a few examinees take the test and comment on the mechanics of the test.
- **Research proposal:** It is a document that provides a detailed description of the intended program.
- **Asset pricing:** It is the study of how financial assets are priced.
- **Financial derivatives:** These are financial instruments that are linked to a specific financial instrument or indicator or commodity, and through which specific financial risks can be traded in financial markets in their own right.
- **Job autonomy:** It is a practice or set of practices involving the delegation of responsibility down the hierarchy so as to give employees increased decision-making authority in respect to the execution of their primary work tasks.
- **Work family conflict:** It occurs when there are incompatible demands between the work and family roles of an individual that makes participation in both roles more difficult.
- **Two-tailed hypothesis:** It is the standard test of significance to determine if there is a relationship between variables in either direction.

NOTES

1.9 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The objective of any research is to find answers to questions through the application of scientific procedures. The main aim of any research is exploring the hidden or undiscovered truth.
2. Qualitative approach to research is concerned with subjective assessment of attitudes, opinions and behaviour. Research in such a situation is a function of researcher's insight and impressions. Such an approach to research generates results either in non-quantitative form or in the forms which are not subjected to rigorous quantitative analysis.
3. When real-life problems require some solution and decision-making, applied research is carried out. This means that applied research is problem oriented and action directed. It brings immediate and practical results.
4. Ex post facto research is the same as experimental research, which is conducted to deal with the situations that occur in or around an organization. Examples of such a research are market failure of an organization's product being researched later and research into the causes for a landslide in the country.
5. A research problem can be defined as a gap or uncertainty in the decision-makers' existing body of knowledge which inhibits efficient decision-making.
6. A literature review is a comprehensive compilation of the information obtained from published and unpublished sources of data in the specific area of interest to the researcher. This may include journals, newspapers, magazines, reports, government publications, and also computerized databases.
7. The first step in research is definition of a problem. Selection of a problem is itself a difficult decision. The success of research depends on right selection of the problem. If the problem has not been identified in right manner, it is very difficult for the researcher to find the right solution of the issue.
8. The best way to understand the nature of the problem is to discuss with those who have prior experience in the same kind of research. This will ensure that the origination of problem and the objectives in view are correct.
9. The following points help in understanding the importance of hypothesis:
   (a) A hypothesis is a proposal intended to explain a fact or an observation.
   (b) A hypothesis specifies the sources of data which shall be studied and in what context they shall be studied.
   (c) It determines the data needs.
   (d) Hypothesis suggests the type of research which is likely to be the most appropriate.
   (e) A hypothesis contributes to the development of the theory.
10. A proposition is a logical statement of relationship between two or more variables which has, generally, been confirmed by empirical research.
11. While comparing two different methods in terms of their superiority, wherein the assumption is that both the methods are equally good, it is called null hypothesis. It is also known as statistical hypothesis and is symbolized as H_0.

12. While comparing two different methods, regarding their superiority, wherein, stating a particular method to be good or bad as compared to the other one, it is called alternate hypothesis. It is symbolized as H_1.

1.10 QUESTIONS AND EXERCISES

Short-Answer Questions

1. List the characteristics of research.
2. Differentiate between research methods and research methodology.
3. Briefly describe the marketing function of research.
4. Write a short note on cross functional research.
5. State the components of a well-defined research problem.
6. State the criteria for the formulation of a research problem.
7. Briefly describe the nature of hypothesis.
8. State the criteria for hypothesis construction.
9. Differentiate between proposition, hypothesis and theory.
10. List the points of comparison between null hypothesis and alternate hypothesis.

Long-Answer Questions

1. Discuss the significance and approach to research methods.
2. Explain the different types of research.
3. Describe the importance and the procedure followed for the formulation of a research problem.
4. Explain the sources of identification and factors influencing the selection of social problems.
5. What are the rules for constructing hypothesis? Also, identify the types of statistical hypothesis.
6. Examine the procedure of hypothesis testing. Support your answer with a suitable diagram.

1.11 FURTHER READING

Booth, Wayne. 2008. *The Craft of Research*, Third edition. Illinois: University of Chicago Press.

Creswall, John W. 2008. *Research Designs: Quantitative, Qualitative and Mixed Methods Approaches*. London: Sage Publications.

Christenson, Larry B. *et al.* 2010. *Research Methods, Design and Analysis*, Eleventh edition. New Jersey: Allyn and Bacon.

Kothari, C. R. 2008. *Research Methodology: Methods and Techniques*. New Delhi: New Age International.

UNIT 2 RESEARCH METHODS AND TECHNIQUES

NOTES

Structure

2.0 INTRODUCTION

Research is a systematic, organized attempt to find answers to meaningful questions, using predefined methods, procedures or techniques which are clearly documented. It should be possible for other people to understand exactly what the researchers did to arrive at their conclusions. In this way, the results and conclusions can be assessed and analysed in terms of relevance and accuracy, bearing in mind any limitations or factors which the researchers may have highlighted.

In this unit, you will be familiarized with the meaning, kinds and stages of survey. You will learn about meaning, types and features of interviews, along with advantages and disadvantages. Apart from survey and interviews, other methods of research such as observations, schedules and questionnaire, and case study method as well as their advantages and disadvantages are also discussed in detail in the unit.

In this unit, you will also be taught about sampling and sampling design. You will learn about the various methods of sampling such as probability or random sampling, and non-probability or non-random sampling. The unit will also discuss the various sources of data. This includes the collection of primary and secondary data, and qualitative and quantitative data.

NOTES

2.1 UNIT OBJECTIVES

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

- Describe the meaning, types and stages of survey
- Explain the different types of interviews
- Discuss the importance of schedules and questionnaires
- Identify the various methods of sampling
- Examine the characteristics and types of observation

2.2 RESEARCH METHODS

A research method refers to the various specific tools or ways through which data can be collected and analysed, such as, survey, questionnaire, case study, interview, and so on. In this section, we will discuss the various methods of research.

2.2.1 Survey

Survey is an important tool in research. No research can be performed without them. Survey can be defined in various ways. Some of the common definitions of survey are as follows:

- To view with a scrutinizing eye; to examine
- To inspect, or take a view of; to view with attention, as from a high place; to overlook; as to stand on a hill and survey the surrounding country
- To determine the form, extent, position, and so on, of as a tract of land, a coast, harbour, or the like, by means of linear and angular measurements, and the application of the principles of geometry and trigonometry; as to survey land or a coast
- A particular view; an examination, especially an official examination, of all the parts or particulars of a thing, with a design to ascertain the condition, quantity, or quality; as a survey of the stores of a ship; a survey of roads and bridges; a survey of buildings
- To examine with reference to condition, situation, value, and so on to examine and ascertain the state of; as to survey a building in order to determine its value and exposure to loss by fire

Types of Survey

There are basically two types of surveys:

(i) **Descriptive:** These surveys generally collect information on what people think and do.

(ii) **Analytic:** These surveys are generally used to either test hypotheses or to answer particular research questions.

While the most common method of collecting survey data is the 'questionnaire', the means by which you gather the information that goes into the survey responses may vary. If the survey makes use of a questionnaire, the measuring instruments must have demonstrable reliability and validity, especially with regard to sampling, questioning and mode of questioning.

Some examples of collecting survey data include self-administered posted questionnaires, web-based forms, telephonic question and answer interviews or face-to-face interviews. There are advantages and disadvantages of each approach, primarily to do with sample size and open versus closed questions. In order to make a judgement, the key areas for consideration include the cost, co-ordination, size of the sample, rate of return, nature and quality of the data obtained, and the ability to clarify questions or responses. The success of using surveys depends strongly on the design of appropriate body of questions and the skill of the interviewer.

Other important methods of collecting survey data are Interview, Observation and Case Study. All these methods have been discussed in detail later.

Stages in Survey Method

Surveys go through the following seven stages:

(i) Planning and designing the survey

In this stage, you must define the goals and objectives of your survey. You should write down the outline of your research and also establish a budget for the project. You are also required to plan your schedule, define the population and estimate the required sample size. The method of data collection and the method for determination of the results should also be decided at this stage. Finally, you must write down the questions and design and pretest the questionnaire.

(ii) Collecting data

In this stage, first you have to decide on the survey method that will suit your research needs. There is no best method of collection of data and you must gather the required data keeping your resources in mind. You must also decide what steps to take in case sufficient data is not collected from the respondents.

(iii) Accessing data

The only purpose of this stage is to transfer the data into the analytical software for processing it further.

(iv) Preparing and managing data

The main aim of this stage is to get the data ready for analysis. In this step, you are required to formulate a 'codebook'. This codebook must include variable names, variable formats and descriptive variable labels. You should also set up multiple item indices and scales, i.e., multiple variables that have exactly the same answer set. In this step, you should transform your data, which will help to get the data in the form and structure required for analysis. Also, the missing data values should be replaced with estimates so that better summary statistics are obtained.

(v) Analysing data

In this stage, you take out all the useful information that you require from the data that you have collected. This helps you make informed decisions.

(vi) Reporting

After analysing the data, the results need to be reported. The main aim of reporting is to produce results from the data analysis which can be easily understood by others who can use this information.

(vii) Deployment

You must tailor your results according to the needs of the target audience. This will ensure the effectiveness of the results.

2.2.2 Case Study

We explore and analyse the life of a social entity, whether it be a family, a person, an institution or a community, with the help of a case study. The purpose of case study method is to identify the factors and reasons that account for particular behaviour patterns of a sample chapter and its association with other social or environmental factors. Generally, social researchers use the case study method to understand the complex social phenomenon and to identify the factors related to this phenomenon. The case study provides the clues and ideas to a researcher for further research study. By adopting the case study method, a researcher gets to know about happenings in the past, which could be related to the research studies and analyse the problem with better perspectives.

Assumptions of the case study method

The assumptions made in a case study method are as follows:

- Case study depends on the imagination of the investigator who is analysing the case study. The investigator makes up his procedure as he goes along.
- History related to the case is complete and as coherent as it could be.
- It is advisable to supplement the case data by observational, statistical and historical data, since these provide standards for assessing the reliability and consistency of the case material.
- Efforts should be made to ascertain the reliability of life history data by examining the internal consistency of the material.
- A judicious combination of techniques of data collection is a prerequisite for securing data that is culturally meaningful and scientifically significant.

Advantages and Disadvantages of Case Study Method

Case study ensures several advantages to the researcher for his research work. The key advantages of the case study method are as follows:

- Provides the basis for understanding complex social phenomenon and all related factors affecting the social phenomenon.
- Provides clues and ideas for exploratory research. When the researcher is not able to get a fair idea about the research, past happenings mentioned in a case study help the researcher get clues and ideas.
- Case study helps in generating objectives for exploratory research.
- It suggests the new courses of inquiry.
- Case study helps in formulating research hypothesis.

Some important disadvantages of case study method are as follows:

- **Reliability:** Data collected through case study may not be reliable or it can be difficult to verify the reliability of data in the current scenario.
- **Adequacy:** Data collected through case studies may not be adequate for research work as data is not pertinent to the research conditions.
- **Representative:** Data presented by case studies represents the happenings with unknown circumstances to a researcher. Hence, it cannot be the true representation of events to a researcher.

Making Case Study Effective

The criteria for evaluating the adequacy of case history is of central importance for case study. American psychologist and social scientist John Dollard has proposed seven criteria for evaluating such adequacy. They are as follows:

(i) The subject must be viewed as a specimen in the cultural series, i.e., the case drawn out from its total context for the purpose of study must be considered as a member of the particular cultural group or community. The scrutiny of life histories of people must be done with a view to identify community values, standards and their shared way of life.

(ii) The organic motto of action must be socially relevant, i.e., the action of individual cases must be viewed as a series of reactions to social stimuli or situations. In other words, the social meaning of behaviour must be taken into consideration.

(iii) The strategic role of the family group in transmitting the culture must be recognized, i.e., in case of the individual being the member of a family in shaping his behaviour must never be overlooked.

(iv) The specific method of elaboration of organic material onto social behaviour must be clearly shown, i.e., the case history that portrays in detail how basically a biological organism, the man, gradually blossoms forth into a social person, is especially fruitful.

(v) The continuous related character of experience for childhood through adulthood must be stressed. In other words, the life history must be a configuration depicting the interrelationships between the persons with various experiences.

(vi) The social situation must be carefully and continuously specified as a factor. One of the important criteria for the life history is that a person's life must be shown as unfolding itself in the context of and partly owing to specific social situations.

(vii) The life history material itself must be organized according to some conceptual framework. This, in turn, would facilitate generalizations at a higher level.

Case study as a method of business research

A detailed case study helps the researcher identify the reasons behind business-related problems. As it can be possible that that particular incident has happened in past, so the current issues can be sorted out by referring to the same case. In-depth analysis of selected cases is of particular value to business research when a complex set of variables may be at work in generating observed results and intensive study is needed to unravel the complexities. The exploratory investigator should have an active curiosity and willingness to deviate from the initial plan, when the finding suggests a new course of

NOTES

enquiry, which might prove more productive. With the help of case study method, the risk can be minimized in any decision-making process.

2.2.3 Census

Census Survey: Census survey means gathering pertinent information about all the units of population, viz., people, institutions, householders, and so on. As you know, population may consist of persons, institutions, objects, attributes, qualities, families among others. A population is a well defined group of many of these. For instance, the Census Survey of India, which takes place once in ten years, gathers benchmark data about each and every household of India. Since it concentrates on each and every household, it restricts its scope to certain surface level demographic data such as age, sex, income, education, lands possessed, cattle, nature of house, domestic facilities available, and so on. The studies are conducted through a quick survey in a stipulated period. However, coverage of units is very exhaustive. The census survey as a method of research in education can be employed to understand educational problems and make policy decisions.

***Strength of Census Survey*:** The strength of the census survey is associated with generalized characteristics of data. Description of population data acts as a major source of identifying several pertinent issues and questions for research. It is very useful in making a trend analysis of different events. Moreover, hard database system of the entire population is very useful in the development of strategic planning and policy-making of education at the micro level as well as at the macro level.

***Limitations of Census Survey*:** As discussed, each and every unit of population is covered under the census survey. However, data is gathered only under limited headings. Also, this data is only surface level information. Through a census survey, one can gather nominal data. Thus, the researcher cannot ask questions in depth.

Many times, such data is gathered mechanically where the investigators are not well trained about cross examining the evidence at the field level. In such cases, the probability of getting valid data is also minimized. Census surveys involve employment of huge manpower and monetary resources. This method is also time consuming. Getting each respondent to cooperate for data collection is very difficult. Hence, the feasibility of conducting census studies is very limited. Moreover, because of sample surveys, many questions can be well answered by saving time, money and human resources; hence, one may look for census studies with limited focus of research.

2.2.4 Experimental Research

Experimental research refers to the research activity wherein the manipulation of variables takes place and the resultant effect on other variables is studied. It provides a logical and structured basis for answering questions. The experimental researchers manipulate the environment, stimuli or applications and observe the impact of this manipulation on the condition or behaviour of the subject. The manipulation that they undertake is deliberate and systematic.

Experimentation is the testing of hypotheses. Once the experimenters have defined a situation or issue, they formulate a preliminary solution or hypothesis. They then apply their observations of the controlled variable relationships in order to test, and then confirm or reject the hypothesis.

Experimentation is the classic method of experimenting in a science laboratory where elements are manipulated and effects observed can be controlled. It is the most sophisticated, exacting and powerful method for discovering and developing an organized body of knowledge.

According to J. W. Best, (Butler University, Emeritus), *Experimental research is the description and analysis of what will be or what will occur under carefully controlled conditions.*

Characteristics of Experimental Research

Experimental research is based on highly rigorous procedures, and aims at producing reliable and valid conclusions. By looking at the various designs and procedures used, one can formulate some essential characteristics of experimental research which distinguish it from other types of research methods like survey and historical.

- **Pre-Experimental Statistical Equivalence of Subjects in Different Groups:** This pre-condition is achieved by random selection and assignment of subjects to different groups. This procedure is essential to meet the threat of selection differences to the internal validity of the results.
- **Use of At Least Two Groups or Conditions that can be Compared:** An experiment cannot be conducted with one group of subjects or one condition at a time. The intent of the experimenter is to compare the effect of one condition on one group with the effect of a different condition on another equivalent group. An experiment may take the shape of a comparison of the effect of one condition on a group of subjects and the effect of another condition on the same group.
- **Manipulation of the Independent Variable:** It is perhaps the most distinct feature of experimental research. Manipulation stands for the process of assignment of different values or magnitudes or conditions or levels of the independent variable to different groups.
- **Measurement of Dependent Variable in Quantifiable Form:** This distinguishes experimental research from descriptive, qualitative or analytical research.
- **Use of Inferential Statistics:** This is done to make probability statements about the results, and, thus, meet the requirements of imperfect measurements on which the behavioural sciences base their generalization.
- **Control of Extraneous Variables:** Though applicable to any other type of research, control of extraneous variables is the *sine qua non* of true experimental designs and the experimenter makes a determined effort to achieve it. It helps the experimenter to eliminate the possibility of any other plausible rival hypothesis claiming to explain the result.

Steps in Experimental Research

The steps in experimental research are as follows:

(i) **Survey of the Literature Relating to the Problem:** In experimentation, the researcher needs to acquire up-to-date information relating to the problem.

(ii) **Selection and Definition of the Problem:** It needs a rigorous logical analysis and definition of the problem in precise terms. The variables to be studied are defined in operational terms clearly and unambiguously. It helps the researcher to convert the problem into a hypothesis that can be verified or refuted by the experimental data.

(iii) **Statement of Hypotheses:** Hypotheses are the heart of experimental research. They suggest that an antecedent condition or phenomenon is related to the occurrence of another condition, phenomenon, event or effect. To test a hypothesis, the researcher attempts to control all the conditions except the independent variable. Therefore, he should give sufficient attention to the formulation of hypotheses. The experimental plant and statistical procedures help him in the testing of hypotheses and contribute little in the development of theories or advancement of knowledge. However, the hypotheses developed or derived from existing theories contribute to the development of new theories and knowledge.

(iv) **Construction of Experimental Plan:** Experimental plan refers to the conceptual framework within which the experiment is to be conducted. According to well-known theorist Van Dalen, an experimental plan represents all elements, conditions, phenomena and relations of consequences so as to:

- Identify the non-experimental variables.
- Identify the most appropriate research design.
- Identify a sample of subjects that will suitably represent the target population, form groups of these subjects and decide on the experiments which will be conducted on each group.
- Choose or develop an instrument that can be deployed to measure the results of the experiment.
- Lay out the data collection process and conduct a pilot study to test the instrument and the research design and state the hypotheses.

Variables

A **variable** is any feature or aspect of an event, function or process that, with its presence and nature, affects some other event or process which is being studied. According to Professor Fred N. Kerlinger, *Variable is a property that takes on different value.*

Types of Variables

The various types of variables are as follows:

- ***Independent Variables*:** These are conditions or characteristics that are manipulated by the researcher in order to identify their relationship to observed phenomena. In the field of educational research, for instance, a specific teaching method or a variety of teaching material are types of independent variables.

  The two kinds of independent variables are as follows:

  (i) ***Treatment Variables*:** These are variables which can be manipulated by the researcher and to which he assigns subjects.

  (ii) ***Organism or Attribute Variables*:** These are factors, such as age, sex, race, religion, and so on, which cannot be manipulated.

- ***Dependent Variables*:** Dependent variables represent characteristics that alter, appear or vanish as a consequence of introduction, change or removal of independent variables. The dependent variable may be a test score or achievement of a student in a test, the number of errors or measured speed in performing a task.

- ***Confounding Variables*:** A confounding variable is one which is not the subject of the study but is statistically related with the independent variable. Hence, changes in the confounding variable track the changes in the independent variable. This creates a situation wherein subjects in a particular condition differ unintentionally from subjects in another condition. This is not a good result for the experiment which is attempting to create a situation wherein there is no difference between conditions other than the difference in the independent variable. This phenomenon enables us to conclude that the manipulation undertaken directly causes differences in the dependent variable. However, if there is another variable besides the independent variable that is also changing, then the confounding variable is the likely cause of the difference. An example of a common confounding variable is that when the researcher has not randomly assigned participants to groups, and some individual difference such as ability, confidence, shyness, height, looks, and so on, acts as a confounding variable. For instance, any experiment that involves both men and women is naturally afflicted with confounding variables, one of the most apparent being that males and females operate under diverse social environments. This should not be confused to mean that gender comparison studies have no value, or that other studies in which random assignment is not employed have no value; it only means that the researcher must apply more caution in interpreting the results and drawing conclusions.

Let us consider an instance wherein an educational psychologist is keen to measure how effective is a new learning strategy that he has developed. He assigns students randomly to two groups and each of the students study materials on a specific topic for a defined time period. One group deploys the new strategy that the psychologist has developed, while the other uses any strategy that they prefer. Subsequently, each participant takes a test on the materials. One of the obvious confounding variables in this study would be advance knowledge of the topic of the study. This variable will affect the test results, no matter which strategy is used. Because of an extraneous variable of this nature, there will be a level of inconsistency within and between the groups. It would obviously be the preferred situation if all students had the exact same level of pre-knowledge. In any event, the experimenter, by randomly assigning the groups, has already taken an important step to ensure the likelihood that the extraneous variable will equivalently affect the two groups.

Let us imagine an experiment being undertaken to measure the effect that noise has on concentration. Assume that there are 50 subjects each in quiet and noisy environments. Table 2.1 below illustrates the ideal or perfect version of this experiment. 'IV' and 'EV' represent the independent variable and external variables, respectively. Note that (as shown in the table) the only difference between the two conditions is the IV, which indicates that the noise level varies from low to high in the two conditions. All the other variables are controlled and are exactly the same for the two conditions. Therefore, any difference in the concentration levels of subjects between the two conditions must have been caused by the independent variable.

Table 2.1 *Determining the Impact of Internal and External Variables*

Variables	Quiet Condition N = 50	Noisy Condition N = 50
Noise Level (IV)	Low	High
IQ (EV)	Average	Average
Room temperature (EV)	68 degrees	68 degrees
Sex of subjects (EV)	60 per cent F	60 per cent F
Task difficulty (EV)	Moderate	Moderate
Time of day (EV)	All different times between 9–5	All different times between 9–5
Etc. (EV)	Same as noisy environ.	Same as quiet environ.
Etc. (EV)	Same as noisy environ.	Same as quiet environ.

An Ideal Experiment

Now consider another version of this experiment wherein some of the other variables differ across conditions. These are confounding variables (highlighted below) and the experiment being conducted is not ideal. In this experiment, if the concentration levels of subjects vary between the two conditions, this may have been caused by the independent variable, *but it could also have been caused by one or more of the confounding variables.* For instance, if the subjects in the noisy environment have lower concentration levels, is it because it was louder, too hot or because they were tested in the afternoon? It is not possible to tell and, therefore, this is less than ideal.

Variables	Quiet Condition	Noisy Condition
Noise Level (IV)	Low	High
IQ (EV)	Average	Average
Room temperature (EV)	68 degrees	82 degrees
Sex of subjects (EV)	60 per cent F	60 per cent F
Task difficulty (EV)	Moderate	Moderate
Time of day (EV)	Morning	Afternoon
Etc. (EV)	Same as noisy environment	Same as quiet environ.
Etc. (EV)	Same as noisy environment	Same as quiet environ.

A Non-Ideal Experiment

Controlling the Confounding Variables

There are ways by which the extraneous variables may be controlled to ensure that they do not become confounding variables. All people-related variables can be controlled through the process of random assignment which will most likely ensure that the subjects will be equally intelligent, outgoing, committed, and so on. Random assignment does not necessarily ensure that this is the case for every extraneous variable in every experiment. However, when a sample is large, it works very well and the researcher's motives for using this method will never be questioned.

One of the way in which situation variables or task variables can be controlled is basically by keeping them constant. For instance, in the noise-concentration experiment above, we could adjust the thermostat and, thereby, keep the room temperature constant and test all the subjects in the same room. We would, of course, hold the difficulty of the tasks constant by giving all subjects in both environments the same task. It is common practice for instructions to be written or recorded and presented to each subject in exactly the same way.

At times, the researcher cannot hold a situation or task variable constant. In these situations too, random assignment can be of great help. Consider a situation where the same room is not available for testing the two groups and, in fact, one group is tested on a Monday in Room 1 and the other group on a Tuesday in Room 2. In this situation, we can use random assignment which can result in half the Monday subjects in Condition A and the rest in Condition B, and the same for the Tuesday subjects. Hence, both conditions will have roughly the same percentage of subjects tested in Room 1 and 2. On the other hand, consider what would happen if we did not use random assignment and instead tested the Monday subjects in Condition A and the Tuesday subjects in Condition B. In this situation, we have two confounding variables. Subjects in Condition A were tested on different days of the week and in different rooms from those in Condition B.

Any difference in the results could have been caused by one or more of the independent variable, the day of the week, or the room.

In other words, confounding variables are those aspects of a study or sample that might influence the dependent variable and whose effect may be confused with the effects of the independent variable. Confounding variables are of two types:

(a) ***Intervening Variables***: In many types of behavioural research, the relationship between independent and dependent variables is not a simple one of stimulus to response. Certain variables that cannot be controlled or measured directly may have an important effect on the outcome. These modifying variables intervene between the cause and the effect. For example, in a classroom language experiment, a researcher is interested in determining the effect of immediate reinforcement on learning the parts of speech. He suspects that certain factors or variables other than the one being studied may be influencing the result, even though they cannot be observed directly. These factors may be anxiety, fatigue or motivation. These factors cannot be ignored. Rather, they must be controlled as much as possible through the use of appropriate design. For example, a variable (as memory) whose effect occurs between the treatment in a psychological experiment (as the presentation of a stimulus) and the outcome (as a response) is difficult to anticipate or is unanticipated, and may confuse the results.

(b) ***Extraneous Variables***: These are variables that are not the subject of an experiment but may have an impact on the results. Hence, extraneous variables are uncontrolled and could significantly influence the results of a study. Often we find that research conclusions need to be questioned further because of the influence of extraneous variables. For instance, a popular study was conducted to compare the effectiveness of three methods of social science teaching. Ongoing, regular classes were used, and the researchers were not able to randomize or control the key variables as teacher quality, enthusiasm or experience. Hence, the influence of these variables could be mistaken for that of an independent variable.

For instance, in a study which attempts to measure the effect of temperature in a classroom on students' concentration levels, noise coming into the class through doors or windows can influence the results and is, therefore, an extraneous variable. This may be controlled by soundproofing the room, which illustrates how the extraneous variable may be controlled in order to eliminate its influence on the results of the test.

The following are the types of extraneous variables:

- Subject variables pertain specifically to the people being studied. These people's characteristics, such as age, gender, health status, mood, background, and so on, are likely to affect their actions.
- Experimental variables pertain to the persons conducting the experiment. Factors, such as gender, racial bias or language influence how a person behaves.
- Situational variables represent the environmental factors which were prevalent at the time when the study or research was conducted. These include the temperature, humidity, lighting and the time of day, and could have a bearing on the outcome of the experiment.
- Continuous variable is one wherein any value is possible within the range of the limits of the variable. For instance, the variable 'time taken to run the marathon' is continuous since it could take 2 hours 30 minutes or 3 hours 15 minutes to run the marathon. On the other hand, the variable 'number of

NOTES

days in a month that a worker came to office' is not a continuous variable since it is not possible to come to office on 14.32 days.

- Discrete variable is one that does not take on all values within the limits of the variable. For instance, the response to a five-point rating scale must only have the specific values of 1, 2, 3, 4 or 5. It cannot have a decimal value such as 3.6. Similarly, this variable cannot be in the form of 1.3 persons.
- Quantitative variable is any variable that can be measured numerically or on a quantitative scale, at an ordinal, interval or ratio scale. For example, a person's wages, the speed of a car or the person's waist size are all quantitative variables.
- Qualitative variables are also known as categorical variables. These variables vary with no natural sense of ordering. They are, therefore, measured on the quality or characteristic. For example, eye colour (black, brown, or blue) is a qualitative variable, as are a person's looks (pretty, handsome, ugly, and so on). Qualitative variables may be converted to appear numeric, but this conversion is meaningless and of no real value (as in male = 1, female = 2).

Experimental Designs

The various experimental designs have been discussed in this section.

(a) **Single Group Design:** In this design study is carried out on a single group. Experiments can be conducted in the following ways:

(i) ***One-Shot Case Study*:** This is a single group studied only once. A group is introduced to a treatment or condition and then observed for changes which are attributed to the treatment. This is like an expost facto method in which on the basis of a dependent variable, an independent variable is looked for.

(ii) ***One Group before after Design*:** This design entails the inclusion of a pre-test in order to establish base level scores. For instance, to use this design in a study of college performance, we could compare college grades prior to gaining the experience to the grades after completing a semester of work experience. In this design, we subtract the score of pre-test from post-test and see the differences. This difference is seen using a 't' test.

(iii) ***Time Series Designs*:** Time series designs refer to the pre-testing and post-testing of one group of subjects at different intervals. In this design, continuous observation is carried out till a clear result is not seen. The purpose is to establish the long-term effects of treatment and can often lead to the number of pre- and post-tests varying from just one each to many. At times, there is a period of interruption between the tests so as to assess the strength of the treatment over a long time frame.

(iv) ***Counterbalanced Design*:** Experiments that use counterbalanced design are effective ways to avoid the pitfalls of repeated measures, where the subjects are exposed to treatments one after the other.

Typically in an experiment, the order in which the treatments are administered can affect the behaviour of the subjects. It may also elicit a false response due to fatigue or any other external factors which may have a bearing on the behaviour of the subjects. To control or neutralize this, researchers use a counterbalanced design, which helps to reduce the adverse effects of the order of treatment or other factors on the results.

Counterbalancing helps to avoid confounding among variables. Take for example an experiment in which subjects are tested on both auditory reaction time task and visual reaction time task. If each and every subject were first tested on the auditory reaction

NOTES

time task and then on the visual reaction time task, the type of task and the order of presentation would be confounded. If the visual reaction time was lower, we would not be sure whether reaction time to a visual stimulus is 'really' faster to an auditory stimulus, as it is quite likely that the subjects would have learned something while performing the auditory task, which led to an improvement of their performance on the visual task.

(b) **Two Equivalent Group Design**

(i) ***Static Group Comparison Study:*** This design attempts to make up for the lack of a control group but falls short in relation to showing if a change has occurred. In this group, no treatment is given but only observation is carried out in a natural way of two groups, e.g., observation of the monkeys living in a city and observation of other monkeys living in the jungle. It is fair to mention here that in these groups nothing is manipulated as this design does not include any pre-testing and, therefore, any difference between the two groups prior to the study is unknown.

(ii) ***Post-Test Equivalent Groups Design*:** Randomization as well as the comparison of both the control and experimental group are used in studies of this nature. Each group is chosen and assigned randomly and presented with either the treatment or a type of control. Post-tests are subsequently administered to each subject to establish whether or not a difference exists between the two groups. While this is close to being the best possible method, it falls short on account of its lack of a pre-test measure. It is not possible to establish if the difference that seems to exist at the end of the study actually represents a change from the difference at the beginning of the study. Hence, while randomization mixes the subjects well, it does not necessarily create an equivalency between the two groups.

(iii) ***Pre-Test and Post-Test Equivalent Groups Design*:** This is the most effective as well as the most difficult method in terms of demonstrating cause and effect. The pre-test and post-test equivalent groups design ensures the presence of a control group as well as a measure of change. Importantly, it also adds a pre-test, thereby, assessing any differences that existed between the groups prior to the study taking place. In order to apply this method, we select students at random and then segregate them into one of two groups. We would subsequently evaluate the previous semester's grades for each group in order to arrive at a mean grade point average. The treatment (work experience) would be applied to one group, whereas a control would be applied to the other.

It is critical that the two groups should be treated similarly in order to control for variables, such as socialization, so that the control group may participate in an activity such as a softball league, while the other group participates in the work experience programme. The experiment ends at the end of the semester and the semester's grades are compared. If it is found that the grade change for the experimental group was significantly different from the grade change of the control group, one could conclude that a semester of work experience results in a significant difference in grades when compared to a semester of non-work related activity programme.

(iv) ***Counterbalanced Randomized Two Groups Design:*** In this design, the group is divided in two parts on a random basis. This design is also called 'rotation design'.

The simplest type of counterbalanced measure design is used when there are two possible conditions, A and B. As with the standard repeated measures design, the researchers want to test every subject for both conditions. They divide the subjects into two groups—one group is treated with condition A, followed by condition B, and the other is tested with condition B, followed by condition A, as shown in Figure 2.1.

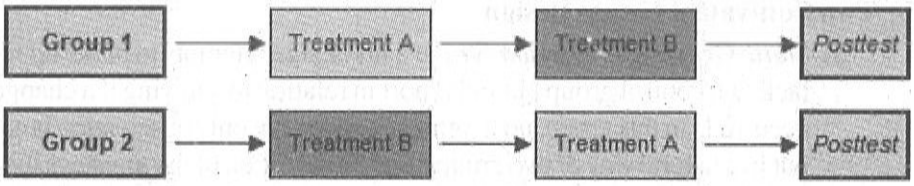


Fig. 2.1 Experiment to Show Counterbalanced Measure Design

(c) **Solomon Four Group Design:** The sample is randomly divided into four groups. Two of the groups are experimental samples, whereas the other two groups experience no experimental manipulation of variables. Two groups receive a pre-test and a post-test. Two groups receive only a post-test. Table 2.2 shows the effect of a particular teaching method on the following groups.

Table 2.2 Solomon Four Group Design

Group		Pre-test	Treatment	Post-test
(a)	R	No	No	No
(b)	R	No	Yes	No
(c)	R	Yes	No	No
(d)	R	Yes	Yes	No

Table 2.3 shows a teaching experiment using the Solomon design where testing before and without treatment have similar results, whilst results after teaching are significantly improved. This indicates that the treatment is effective and not subject to priming or learning effects.

Table 2.3 Pre-and Post-Testing

Group		Pre-test	Treatment	Post-test	Pre-result	Post-result
(a)	R	No	No	No	3	10
(b)	R	No	Yes	No	4	5
(c)	R	Yes	No	No		9
(d)	R	Yes	Yes	No		3

Internal and External Validity in Experimental Research

Internal validity is considered as a property of scientific studies which indicates the extent to which an underlying conclusion based on a study is warranted. This type of warrant is constituted by the extent to which a study minimizes a systematic error or a 'bias'. If a causal relation between two variables,is properly demonstrated, then the inferences are said to possess internal validity. A fundamental inference may be based on a relation when the following three criteria are satisfied:

1. The 'cause' precedes the 'effect' in time (temporal precedence).
2. The 'cause' and the 'effect' are related (covariation).
3. There are no plausible alternative explanations for the observed covariation (non-spuriousness).

NOTES

Internal validity refers to the ability of a research design for providing an adequate test of an hypothesis and the ability to rule out all plausible explanations for the results but the explanation being tested. For example, let us consider that a researcher decides that a particular medication prevents the development of heart disease because he found that research participants who took the medication developed lower rates of heart disease than those who never took the medication. This interpretation of the study's results is likely to be correct, however, only if the study has high internal validity. In order to have high internal validity, the research design must have controlled the directionality and third-variable problems, as well as for the effects of other extraneous variables. In short, the researcher would have needed to perform an experimental study in which:

- Participants were randomly assigned to the experimental and control groups.
- Participants did not know whether they were taking the medication.

The most internally valid studies are experimental studies because they are better than correlational and case studies at controlling for the directionality and third-variable problems, as well as for the effects of other extraneous variables.

Threats to Internal Validity

The following are the various threats to internal validity:

Ambiguous Temporal Precedence: Lack of precision about the occurrence of variable, i.e., which variable occurred first, may yield confusion that which variable is the cause and which is the effect.

Confounding: Confounding is a major threat to the validity of fundamental inferences. Changes in the dependent variable may rather be attributed to the existence or variations in the degree of a third variable which is related to the manipulated variable. Rival hypotheses to the original fundamental inference hypothesis of the researcher may be developed where spurious relationships cannot be ruled out.

Selection Bias: It refers to the problem that, at pre-test, differences between the existing groups may interact with the independent variable and, thus, be 'responsible' for the observed outcome. Researchers and participants bring to the experiment a myriad of characteristics, some learned and others inherent; for example, sex, weight, hair, eye and skin color, personality, mental capabilities and physical abilities, and so on. Attitudes like motivation or willingness to participate can also be involved. If an unequal number of test subjects have similar subject-related variables during the selection step of the research study, then there is a threat to the internal validity.

Repeated Testing: It is also referred to as testing effects. Repeatedly measuring or testing the participants may lead to bias. Participants of the testing may remember the correct answers or may be conditioned to know that they are being tested. Repeatedly performing the same or similar intelligence tests usually leads to score gains instead of concluding that the underlying skills have changed for good. This type of threat to internal validity provides good rival hypotheses.

Regression towards the Mean: When subjects are selected on the basis of extreme scores (one far away from the mean) during a test, then this type of threat occurs. For example, in a testing when children with the bad reading scores are selected for participating in a reading course, improvements in the reading at the end of the course might be due to regression toward the mean and not the course's effectiveness actually. If the children had been tested again before the course started, they would likely have obtained better scores anyway.

NOTES

External Validity

External Validity is considered as the validity of generalized (causal or fundamental) inferences in scientific studies. It is typically based on experiments as experimental validity. In other words, it is the degree to which the outcomes of a study can be generalized to other situations and people.

If inferences about cause and effect relationships which are based on a particular scientific study may be generalized from the unique and characteristics settings, procedures and participants to other populations and conditions, then they are said to possess external validity. Causal inferences possessing high degrees of external validity can reasonably be expected to apply:

- To the target population of the study, i.e., from which the sample was drawn. It is also referred to as population validity.
- To the universe of other populations, i.e., across time and space.

An experiment using human participants often employ small samples which are obtained from a single geographic location or with characteristic features is considered as the most common threat to external validity. Due to this reason, one cannot be certain that the conclusions drawn about cause and effect relationships do actually apply to people in other geographic locations or without these particular features.

External validity refers to the ability of a research design for providing outcomes that can be generalized to other situations, especially to real-life situations. For instance, if the researcher in the hypothetical heart disease medication study found that the medication, under controlled conditions, prevented the development of heart disease in research participants, he would want to generalize these findings to state that the medication will prevent heart disease in the general population. However, let us consider that the research design required the elimination of many potential participants, such as people who abuse alcohol or other drugs, suffer from diabetes, weigh more than average for their height, and have never suffered from a mood or anxiety disorder. These are common risk factors for heart disease and, by eliminating these factors, the outcomes of the study would provide little evidence that the medication will be effective for people with these risk factors. In other words, the study would have low external validity and, hence, its outcomes to the general population could not be generalized.

This commonly happens in tests of antidepressant medications. Because researchers want to make sure that the antidepressant effects of the medications being tested are not hidden by the effects of extraneous variables, they often have excluded potential participants with one or more of the following characteristics:

- People who are addicted to alcohol or illicit drugs
- People who take various medications
- People who have anxiety disorders (such as phobic disorders)
- People who suffer from depression with psychosis
- People with mild depression (because they would show only a small response to the medication)

If a study excluded people with these characteristic features, then most of the participants suffering from depression would be excluded from the final pool of participants. The outcomes of the study, therefore, would provide little information about how most depressed people will respond to the medication.

Threats to External Validity

A threat to external validity is an explanation of how you might be wrong in making a generalization. Usually, generalization is limited when the cause, i.e., independent variable depends on other factors; therefore, all threats to external validity interact with the independent variable.

- **Aptitude-Treatment Interaction:** The sample may have specific characteristic features that may interact with the independent variable, limiting generalization. For example, inferences based on comparative psychotherapy studies often employ specific samples (e.g., volunteers, highly depressed, no comorbidity). If psychotherapy is found effective for these sample patients, will it also be effective for non-volunteers or the mildly depressed or patients with concurrent other disorders?
- **Situation:** All situational features, such as treatment conditions, time, location, lighting, noise, treatment administration, investigator, timing, scope and extent of measurement, and so on, of a study potentially limit generalization.
- **Pre-Test Effects:** If cause and effect relationships can only be found when pre-tests are carried out, then this also limits the generality of the findings.
- **Post-Test Effects:** If cause and effect relationships can only be found when post-tests are carried out, then this also limits the generality of the findings.
- **Reactivity (Placebo, Novelty and Hawthorne Effects):** If cause and effect relationships are found, they might not be generalized to other situations if the effects found only occurred as an effect of studying the situation.
- **Rosenthal Effects:** Inferences about cause-consequence relationships may not be able to generalize to other investigators or researchers.

2.2.5 Focused Groups

Focus group as a method was developed in the 1940s in Columbia University by sociologist Robert Merton and his colleagues as part of a sociological technique. This was used as a method for measuring audience reaction to radio programmes. In fact, the method was uniquely adapted and modified in different branches of social sciences namely anthropology, sociology, psychology, education and advertising. It essentially emerged as an alternative method which was more cost effective and less time consuming, and could generate a large amount of information in a short time span. Another argument given in its favour was that group dynamics play a positive role in generating data that the individual would be hesitant about sharing when he was spoken to individually.

A focus group is a highly versatile and dynamic method of collecting information from a representative group of respondents. The process generally involves a moderator who manoeuvres the discussion on the topic under study. There are a group of carefully-selected respondents who are specifically invited and gathered at a neutral setting. The moderator initiates the discussion, and then the group carries it forward by holding a focused and an interactive discussion. The technique is extensively used and at the same time also criticized. While one school of thought places group dynamics at an important position, another negates its contribution as detrimental. We will examine these as we go along.

Key elements of a focus group: There are certain typical requirements for a conducive discussion. These need to be ensured in order to get meaningful and usable outputs from the technique.

Size: The size of the group is extremely critical and should not be too large or too small. Edward F. Fern, Professor Emeritus in the Pamplin College of Business at Virginia Tech, stated that every member is assumed to contribute meaningfully to the discussion; however, if the size of the group is too large, then contribution by the members might not be premium. Ideal recommended size, thus, for a group discussion is 8 to 12 members. Less than eight would not generate all the possible perspectives on the topic and the group dynamics required for a meaningful session.

Nature: Individuals who are from a similar background—in terms of demographic and psychographic traits—must be included; otherwise the disagreement might emerge as a result of other factors rather than the one under study. For example, a group of homemakers and working women discussing packaged food might not have a similar perspective towards the product because they have different roles to manage and balance; thus, what is perceived as convenience by one is viewed as indifferent and careless attitude towards one's family by the other. The other requirement is that the respondents must be similar in terms of the subject/policy/product knowledge and experience with the product under study. Moreover, the participants should be carefully screened to meet a certain criteria.

Acquaintance: It has been found that knowing each other in a group discussion is disruptive and hampers the free flow of the discussion, and it is believed that people reveal their perspectives more freely amongst strangers rather than friends. Terry Bristol, University of Arkansas at Little Rock, found that men revealed more about themselves amongst strangers, while females were more comfortable amongst acquaintances. Thus, it is recommended that the group should consist of strangers rather than subjects who know each other. There are exceptions, however, in certain cases; this would be further discussed in a subsequent section.

Setting: As far as possible, the external factors which might affect the nature of the discussion are to be minimized. One of these could be the space or setting in which the discussion takes place. Thus, it should be as neutral, informal and comfortable as possible. Even the ones that have one-way mirrors or cameras installed need to ensure that these gadgets are as unobtrusively placed as possible.

Time period: The conduction of the discussion should be held in a single setting unless there is a before and after design which requires group perceptions, initially before the study variable is introduced; and later in order to gauge the group's reactions. The ideal duration of conduction should not exceed one and a half hour. This is usually preceded by a short rapport formation session between the moderator and the group members.

The recording: Earlier there were human recorders, either sitting behind one-way mirrors or in the discussion room. Today, these have been replaced by cameras that video-record the entire discussion. This can, then, be replayed for analysis and interpretation. The advantage over human recording is that one is able to observe the non-verbal cues and body language as well. This technology has been further enhanced and one can evaluate the discussion happening at one location, being observed and transmitted at another.

The moderator: He is the key conductor of the whole session. The nature, content and validity of the data collected are dependent to a large extent on the skills of the moderator. His role might be that of a participant where he might be a part of the group discussion or he might be a non-participant and has the task of rapport formation, initiating the discussion and steering the discussion forward. Well-known researchers Helen Morgan and Kerry Thomas have stated that any group task has two clear agendas. One is the conscious agenda to complete the overt task and the second, more important, plan is related to the unconscious. This is concerned with the emotional needs of the group and has been described differently as 'group mind', 'group as a whole' and 'group as a group'. The moderator is clearly responsible for this as he needs to work with the group as a group in order to maximize the group performance. Thus, he needs to possess some critical moderating skills like:

- Ability to listen attentively and have a positive demeanour that encourages others to discuss. At the same time, he must be detached and give no indication about his personal opinion in order to skew the discussion. He should be dressed in a manner that is informal and similar to the group.
- He needs to make others feel comfortable; thus, the language used should be in the subjects' lingo, with no use of technical words at all.
- He needs to be flexible in approach, so that the discussion flows naturally rather than becoming compartmentalized into a question and answer session. At the same time, he also needs to act as a translator in case someone's point is not understood or interpreted correctly.
- He must also discreetly handle the overbearing and dominating participants and encourage all the members to contribute by drawing out the hesitant ones as well. Thus, sensitivity to the respondents' feelings must be present at all times.
- There is no external signal, so he needs to be sufficiently trained and acquainted with the topic to understand the specific interval when all the possible viewpoints get exhausted and the discussion needs to move on.

In conducting the discussions, he might use the *summary and closure* approach where he might pick up a similar point made by a participant to another and summarize it and ask for his opinion. Another tactic that can be used is to bring in the *extreme opinions* on the topic, in case no counter points are coming through; this, then, is able to generate more arguments into the discussion. Sometimes, rather than the moderator introducing another viewpoint, he might ask 'is that all?' This might sometimes trigger a fresh stance.

Steps in planning and conducting focus groups: The focus group conduction has to be handled in a structured and stepwise manner as stated below:

(i) Clearly define and enlist the research objectives of the research study that require qualitative research.

(ii) Then these objectives have to be split into information needs to be answered by the group. These may be bulleted as topics of interest or as broad questions to be answered by the group.

(iii) Next, a list of characteristics needs to be prepared, which would be used to select the respondent group. Based on this screening, a questionnaire is prepared to measure the demographic, psychographics, topic-related familiarity and knowledge. In case of a product or policy, one also needs to find out the experience and attitude towards it. Next, a comprehensive moderator's outline for conducting the whole process needs to be charted out. Here, it is critical to involve the decision-maker (if any), the business researcher as well as the moderator. This is done so that there is complete clarity for the moderator in terms of the intention and potential applicability of the discussion output. This involves extensive discussions-among the researcher, client and the moderator. Another advantage of having a structured guideline is that in case of multiple moderators, who might need to conduct focus group discussions at different locales, collection of similar information and reliability of the method can be maintained.

(iv) After this, the actual focus group discussion is carried out. Different sociologists have enlisted various stages that take place in a focus group. The most famous and comprehensive is the linear model of group development formulated by Research theorist Tuckman. This has been adapted by well-known qualitative researcher Joanna Chrzanowska to explain the stages in the focus group discussions (Table 2.4).

(v) The focus group reveals rich and varied data; thus, the analysis cannot be quantitative or even in frequencies. The summary of the findings are clubbed under different heads as indicated in the focus group objectives and reported in a narrative form. This may include expressions like 'majority of the participants were of the view' or 'there was a considerable disagreement on this issue'. A summary report on the focus group discussion held in the organic food study is presented below along with the moderator guide.

Focus group study: Potential consumers

Two separate focus group discussions were conducted—one in Noida (UP) and the other in Hi-Tech City, Hyderabad. The group at Noida was predominantly of housewives and the one in Hi-Tech had professionals from different walks of life. Their opinion on a variety of subjects was sought. A summary of the discussions is presented below:

Adulteration in food

All the participants were unanimously concerned about adulterated food that they and their families were consuming. The discussion went from pesticides to chemicals and spurious food products. The ladies felt that they experienced a lot of health problems, specifically acidity, because of adulteration in the food. Some stated that they tried to grind all masalas at home as they felt that most of the problem was with masalas. However, some felt that this was meaningless as the whole masala was adulterated and contaminated by chemical residues. Thus, even though it was a matter of concern for them, they felt helpless to verbalize the possible solution.

There was one lady (Noida group), however, who felt that some of the problems were exaggerated and were basically created by the media and were plain hype. Another lady (HT group) felt that the problem of pollution was too deep- rooted and just adulterated food or food grown with chemical fertilizers and pesticides was too elementary and small to comprehend the problem of health hazards of the general population.

Changes in lifestyle

The consumers observed major changes in the recent years. The groups were unanimously of the opinion that they were more health conscious and concerned than their mothers and grandmothers. The younger generation (post-teens especially) are extremely conscious about the nutritional content of their food. They actively avoid excess sugar and fats in their diet. As a regime, people said that they exercise in some form or the other. Some said they drink more water and include healthy supplements like sprouts and olive oil in their diets.

Awareness of organic food products

Almost all the consumers, with the exception of one, had read or heard of organic food. One respondent had tried the product and found it very tasty. Three of the group members, as stated earlier, were sceptical about the benefits of organic food.

Willingness to try

The product was formally introduced to the groups and their reactions were noted to the same. Most of them, with the exception of two, were extremely enthusiastic about the products and wanted to know more about them, and had a number of queries about the availability, price, brands and benefits of the products.

Suggestions for marketing the product

- Divided opinion on who should sell the product. Some felt that a government-approved outlet like Mother Dairy/Trinetra should sell the products, whereas others felt that there should be exclusive organic food outlets. There were two or three people who felt that there should be no distinction and the products should be available everywhere. Some were also of the opinion that the products could be sold at high-end grocery stores or departmental stores since this was an expensive product. One consumer suggested the vegetable *mandi* also as a possible outlet; however, most of the others felt that the products would not be purchased by the masses.
- All the group members were unanimously of the opinion that they would buy a product only if it was certified as organic from an authentic and reputed body.
- The product should be vaccum packed, preferably in a brown paper packet with the label having the certification information and the source of the product clearly displayed.
- All felt that the price difference should not be too steep. At the same time, the Indian consumer who is buying a quality product accepts a price difference, so the product should be slightly expensive than the non-organic option.
- All the respondents felt that television was the best medium for promoting the product. All opined that there was a dire need for creating awareness. They felt that there was absolutely no visibility for the products, and more availability and awareness would mean more sales and more organically converted consumers. Some suggested popular soap operas and others were in favour of educational programmes.
- Some respondents felt that product promotions should be effectively and widely-conducted by tying up with environment-related organizations that would be willing to promote a healthy cause.
- In terms of endorsement, they wanted sports personalities, film stars like Hema Malini, Simi Grewal, among others, and politicians like Menaka Gandhi and Sushma Swaraj endorsing the product; some even suggested common people who eat organic products and the farmer who produces.
- The groups were generally of the opinion that the campaigns should be targeted at housewives and school children who would be wonderful and effective change agents.

NOTES

- Comparative advertising demonstrating the benefits of organic versus non-organic was another valuable suggestion discussed in the group. Some, however, argued for simply enlisting the benefits and resolving the myths about the products.
- Price and availability and the reputation of the organization or brand would be important issues in marketing the product effectively.
- Some punch lines suggested for the product were as follows:
  - It is the future
  - The healthy alternative
  - *Shudh* and *swachh*
  - Shuddhaahaar
  - Healthorganic
  - Organic is healthy
  - Go organic

Table 2.4 Stages in a Focus Group Discussion

Stage	Affective reactions	Behaviour patterns	Moderator role
Forming	The group members are uncomfortable, insecure, a little lost and apprehensive.	Silence or general talk, greetings and introductions. Mundane activity.	Tries to bring clarity by explaining the purpose of gathering together, and the expected behaviour during the discussion.
Storming	There is chaos, as emotions start flying with members questioning others and voicing their own opinion.	Arguments directed at each other or trying to seek support from the moderator. Generally there is rigidity in terms of sticking to ones position. The leaders and the followers emerge.	Do not take side. Play poker face and say that all opinions are welcome. Steer the direction to the topic rather than arguments which might go off the tangent. Try to draw out the passive participants.
Norming	Cliques and sides start forming based on the stand that people have taken. More supportive and positive signals, especially non-verbal.	People have got the hang of the process and do not really need any steering by the moderator.	Takes it easy, is more bothered about sequencing of information and managing time now.
Performing	Individuals are subservient to the group, roles are flexible and task-oriented.	Sense of concentration and flow, everything seems easy, high energy, group works without being asked.	Time to introduce difficult issues, stimulus material, projective techniques.
Re-adjustment: There might be role reversals. People may have another perspective with which the loosely-defined cliques might not agree, so one of the earlier stages might emerge.			
Mourning	Group task nearing completion, so there might be a sense of loss as the energy generated with the discussion might be sapped.	If members do not feel that any clear stand is emerging, they might want to continue and not disband the group.	Signal conclusion. If you want to summarize, ask if any one has something to add. Thank everyone and disperse for refreshments or closure.

(***Source:*** Chrzanowska, 2002)

Types of focus groups: As stated earlier, there could be several variations to the standard procedure. Some such innovations and alternative approaches are presented below:

Two-way focus group: Here, one respondent group sits and listens to the other and after learning from them or understanding the needs of the group, carry out a discussion amongst themselves.

For example, in a management school, the faculty group could listen to the opinions and needs of the student group. Subsequently, a focus group of the faculty could be held to study the solutions or changes that they perceive which need to be carried out in the dissemination of the programme.

NOTES

Dual-moderator group: Here, there are two different moderators: one responsible for the overt task of managing the group discussion and the other for the second objective of managing the 'group mind' in order to maximize the group performance.

Fencing-moderator group: The two moderators take opposite sides on the topic being discussed and, thus, in the short time available, ensure that all possible perspectives are thoroughly explored.

Friendship groups: There are situations where the comfort level of the members needs to be high so that they elicit meaningful responses. This is especially the case when a supportive peer group encourages admission about the related organizations or people/ issues. Eminent research thinker Stevens used the technique successfully when studying women groups for their experiential consumption of women magazines.

Mini-groups: These groups might be of a smaller size (usually four to six) and are usually expert groups/committees that on account of their composition are able to decisively contribute to the topic under study.

Creativity group: These are usually of longer than one and a half hour duration and might take the workshop mode. Here, the entire group is instructed which then brainstorms into smaller sub-groups and then reassembles to present their sub-groups opinion. They might also stretch across a day or two. A variation of the technique uses projective methods to extract alternative thinking.

Brand-obsessive group: These are special respondent sub-strata who are passionately involved with a brand or product category (say cars). They are selected as they can provide valuable insights that can be successfully incorporated into the brand's marketing strategy.

Online focus group: This is a recent addition to the methodology and is extensively used today. Thus, it will be elaborated in detail. Like in the case of regular group process, the respondents are selected from an online list of people who have volunteered to participate in the discussion. They are then administered the screening questionnaire to measure their suitability. Once they qualify, they are given a time, a participating id and password and the venue where they need to be so that they can be connected with the others. The group size here varies from four to six, as otherwise there might be technical problems and lack of clarity in the voices received. To ensure a standardized way of responding, the respondents are mailed details of how to use specific symbols to express emotions while typing the responses. Usual ways of denoting happiness and displeasure are used. These could also be coloured differently; also to show a higher degree of the emotion, additional faces may be used. There are other symbols as well. Besides, a brief about the purpose of the discussion and clarity on specific or technical terms is provided before the conduction. At the designated time, the group assembles in a web-based chat room and enters their id and password to log on. Here the chatting between the moderator and the participant is real time. Once the discussion is initiated, the group is on its own and chats amongst themselves, with the moderator playing the typical role. The session lasts for one to one and a half hour, and the process is much faster than a normal focus group.

The advantage of the method is that geographic locations arc not a constraint and persons from varied locations can participate meaningfully in the discussion. Also, since it does not require a commitment to be physically assembled at a particular place and time, people who are busy and otherwise are not able to participate, can also be tapped. Since the addresses of the members are available to the moderators, it is also possible subsequently to probe deeper at a later date or seek çlarifications. The interaction is faceless so the person interacting is completely assured of his/her anonymity and is thus, less inhibited. The method also has a cost advantage as compared to a traditional focus group. People are generally less inhibited in their responses and are more likely to fully express their thoughts. A lot of online focus groups go well past their allotted time since so many responses are expressed. Finally, as there is no travel, videotaping or facilities to arrange, the cost is much lower than for traditional focus groups. Firms are able to keep costs between one-fifth and one-half the cost of traditional focus groups.

However, the method can be actively and constructively used only with those who are computer savvy. Another disadvantage is that since anonymity is assured, actual authentication of the respondent being a part of the population under study might be a little difficult to establish. Thus, to verify the details, one may use the traditional telephone method and cross check the information. Since the person is typing his/her response, other sensory cues of tone, body language and facial expressions are not available. Thus, while the apparent emotions or attitudes can be tapped, however, the unconscious or sub-conscious cannot be judged.

These techniques have extensive use for companies that are into e-commerce. Most companies today have started using this technique to get employee reactions to various organizational issues, in what is termed as a 'virtual town hall meeting'. Thus, cyber dialogues can be carried out and meaningful feedback as well as population reaction can be measured with considerable ease and accuracy.

Evaluating focus group as a method

Focus groups are extensively criticized and yet have widespread usage in all areas of business research, to the extent that the technique is considered by some as synonymous with qualitative research. Before concluding the discussion on focus groups, let us examine the benefits and drawbacks of using the method.

Idea generation: As discussed earlier, the collective group mind creates an atmosphere where ideas and suggestions are churned out which are more holistic and significant than those that would be generated in an individual interview. The other advantage is that the group process works towards vetting each idea as it is presented. The dialogue between the members helps to refine and rephrase the perspective into a usable solution at the end of the discussion.

Group dynamics: Once the moderator has initiated the debate and some members have expressed their opinion, the atmosphere becomes charged and the respondents' involvement with the topic increases with most members presenting reactions and counter reactions. The expressiveness becomes contagious and the contrived discussion slowly becomes a free-flowing discussion. As the comfort level of individuals with the other members increases, they start feeling at ease with the setting and expression becomes more open.

Process advantage: The discussion situation permits considerable flexibility in extracting the relevant information as the flow of topics and the extent to which the topic can be debated is dependent upon the group members and the emerging dynamics. Also, the situation permits a simultaneous conduction and collection of information from a number of individuals at a single point of time.

Reliability and validity: Since the objectives of the study have been listed out and the structure of the moderator outline is predetermined, the reliability of the information obtained is high. The mechanical recording of the data removes the element of human bias and error in the information collected.

However, the technique is not without shortcomings.

NOTES

Group dynamics: The advantage could also be a disadvantage. On account of the group setting, the members might present a perspective not necessarily their own, but one that is along the lines of the group expression. This is the 'nodding dog syndrome', which is often a result of group conformity.

Scientific process: The group discussion must be treated as indicative and, thus, generalizing must be avoided. The answers obtained are varied and in a narrative form. Thus, coding and analysing this data is quite cumbersome.

Moderator/investigator bias: As discussed in earlier sections, the success or failure of the process depends, to a large extent, on the skills of the moderator. An unbiased and sensitive moderator who is able to generate meaningful and unbiased discussions is quite a rarity.

2.3 METHOD OF DATA COLLECTION

This section discusses the various methods of data collection in detail.

2.3.1 Observation

Observation can be defined as viewing or seeing. Observation means specific viewing with the purpose of gathering the data for a specific research study. Observation is a classical method of scientific study. It is very important in any research study as it is an effective method for data collection.

Characteristics of Observation Method

The following are the characteristics of observation method of data collection:

- **Physical and mental activity:** Eyes observe so many things in our surroundings but our focus or attention is only on data which is relevant to research study.
- **Observation is selective:** It is very difficult for a researcher to observe everything in his surroundings. He only observes the data which is purposive for his research study and meets with the scope of his study. The researcher ignores all the data which is not relevant to the study.
- **Observation is purposive and not casual:** Observation is purposive as it is relevant to a particular study. The purpose of observation is to collect data for the research study. It focuses on human behaviour which occurs in a social phenomenon. It analyses the relationship of different variables in a specific context.
- **Accuracy and standardization:** Observation of pertinent data should be accurate and standardized for its applications.

Types of Observation

Different concepts define the classification of observations.

With respect to an investigator's role, observation may be:

- Participant observation
- Non-participant observation

Check Your Progress

1. Define experimental research.
2. What are the threats to internal validity?
3. List the different types of focus groups.

NOTES

With respect to the method of observation, it can be classified into the following:

- Direct observation
- Indirect observation

With reference to the control on the system to be observed, observation can be classified into the following:

- Controlled observation
- Uncontrolled observation

(i) Participant observation

In participant type of observation, the observer is an active participant of the group or process. He participates as well as observes as a part of phenomenon; for example, to study the behaviour of management students towards studying and understanding marketing management, the observer or researcher has to participate in the discussion with students without telling them about the observation or purpose. When respondents are unaware of observations, then only their natural interest can be studied.

Advantages

The following are the main advantages of participant observation:

- In-depth understanding of the respondent group.
- The context which is meaningful to observed behaviour can be recorded or documented by the researcher.

Disadvantages

The following are the disadvantages of participant observation:

- If a participant is at lower level in hierarchy of group, his participation may be less.
- Emotions of the observer may result in loss of objectivity.

(ii) Non-participant observation

In non-participant observation, the observer does not participate in the group process. He acknowledges the behaviour of the group without telling the respondents. It requires a lot of skills to record observations in an unnoticeable manner.

(iii) Direct observation

In direct observation, the observer and researcher personally observe all the happenings of a process or an event when the event is happening. In this method, the observer records all the relevant aspects of an event which are necessary for study. He is free to change the locations and focus of the observation. One major limitation of the method is that the observer may not be able to cover all relevant events when they are happening.

(iv) Indirect observation

Physical presence of an observer is not required and recording is done with the help of mechanical, photographic or electronic devices; for example, close circuit TV (CCTV) cameras are used in many showrooms to observe the behaviour of customers. It provides a permanent record for an analysis of different aspects of the event.

(v) Controlled observation

All observations are done under pre-specified conditions over extrinsic and intrinsic variables by adopting experimental design and systematically recording observations. Controlled observations are carried out either in the laboratory or the field.

(vi) Uncontrolled observation

There is no control over extrinsic and intrinsic variables. It is mainly used for descriptive research. Participant observation is a typical uncontrolled one.

Prerequisites of Observation

The following are the prerequisites of observation:

- The conditions of observation must provide accurate results. An observer should be in a position to observe the object clearly.
- The right number of respondents should be selected as the sample size for the observation to produce the desired results.
- Accurate and complete recording of an event.
- If it is possible, two separate observers and sets of instruments can be used in all or some observations. Then the result can be compared to measure accuracy and completeness.

Advantages and Limitations of Observation

The following are the advantages of observations:

- It ensures the study of behaviour in accordance with the occurrence of events. The observer does not ask anything from the representatives; he just watches the doing and saying of the sample.
- The data collected by observation defines the observed phenomenon as they occur in their natural settings.
- When an object is not able to define the meaning of its behaviour, observation is best method for analysis; for example, animals, birds and children.
- Observation covers the entire happenings of an event.
- Observation is less biased as compared to questioning.
- It is easier to conduct disguised observation studies as opposed to disguised questioning.
- The use of mechanical devices can generate accurate results.

The following are the limitations of observation:

- Past studies and events are of no use to observation. For these events and study, one has to go through narrations, people and documents.
- It is difficult to understand attitudes with the help of observation.
- Observations cannot be performed by the choice of the observer. He has to wait for an event to occur.
- It is difficult to predict when and where the event will occur. Thus, it may not be possible for an observer to reach in every event.
- Observation requires more time and money.

NOTES

Uses of observation in business research

Observation is very useful in the following business research purposes:

- Buying behaviour of customer, lifestyles, customs, interpersonal relations, group dynamics, leadership styles, managerial style and actions
- Physical characteristics of inanimate things such as houses, factories, stores, and so on
- Movements in a production plant
- Flow of traffic, crowd and parking on road

2.3.2 Interview

Interviewing is a very effective method of data collection. It is a systematic and objective conversation between an investigator and respondent for collecting relevant data for a specific research study. Along with conversation, learning about the gestures, facial expressions and environmental conditions of a respondent are also very important. Generally, interview collects a wide range of data from factual demographic data to highly personal and intimate information relating to a person's opinions, attitudes, values and beliefs, past experience and future intentions. The interview method is very important in the collection of data from the respondent who is less educated or illiterate. Personal interview is more feasible when the area covered for survey is compact. Probing is a very important part of an interview.

Types of Interview

The following are the various types of interviews:

(i) Structured or directive interview

In this type of interview, the investigator goes to the respondent with a detailed schedule. Some questions in same sequence are asked from all respondents.

(ii) Unstructured or non-directive interview

In this type of interview, the respondent is encouraged to give his honest opinion on a given topic without or with minimum help of others.

(iii) Focused interview

This is a semi-structured interview where the respondent shares the effect of the experience to the given conditions with the researcher or investigator. It is conducted with those respondents only who have prior experience of conditions given by the investigator. Analysis of the attitude and emotional feelings for the situations under study is main purpose behind conducting these interviews. A set of fix questions may not be required in this interview but a relevant topic is required which is known to the respondent.

(iv) Clinical interview

While a focused interview is concerned with effects of specific experience, clinical interviews are concerned with broad underlying feelings or motivations or the course of the individual's life experiences with reference to the research study. It encourages the interviewee to share his experience freely.

NOTES

(v) Depth interview

To analyse or study the respondent's emotions, opinions, and so on, depth interviews are conducted. This kind of interview aims to collect intensive data about individuals, especially their opinions. It is a lengthy process to get unbiased data from the respondent. Interviewers should avoid advising or showing this agreement. Instead, the investigator has to motivate the respondent to answer the questions.

Features of Interviews

The following are some of the features of interviews:

- The interviewer and the respondent are the participants in any interview. They both are unknown to each other and so it is important for an interviewer to introduce himself first to the respondent.
- An interview has a beginning and a termination point in the relationship between the participants.
- Interview is not a mere casual conversational exchange. It has a specific purpose of collecting data which is relevant to the study.
- Interview is a mode of obtaining a verbal response to questions to put verbally. It is not always face to face.
- Success of interview depends on the interviewer and respondent, and how they perceive each other.
- It is not a standardized process.

Essentials for an Effective Interview

The following are the requirements for a successful interview:

- **Data availability:** The respondent should have complete knowledge of the information required for specific study.
- **Role perception:** The interviewer and the respondent should be aware of their roles in the interview process. The respondent should be clear about the topic or questions which have to be answered by him. Similarly, it is the responsibility of the interviewer to make the respondent comfortable by introducing himself first. The investigator should not affect the interview situation through subjective attitude and argumentation.
- **Respondent's motivation:** The respondent can hesitate to answer the questions. In this case, the approach and skills of the interviewer are very important as he has to motivate the respondent to answer or express himself.

Advantages and Disadvantages of Interviews

The following are the advantages of the interview method:

- In-depth and detailed information is collected.
- The interviewer tries to improve the responses and quality of data received. He can control the conditions in favour of the research study.
- Interviews help in gathering supplementary information which can be helpful to the study.
- Interviews use special scoring devices, visuals and materials to improve the quality of data or information collected.

NOTES

- Interviews use observation and probing by the interviewer to see the accuracy and dependability of given data by the respondent.
- Interviews are flexible in nature.

The following are the disadvantages of interviews:

- Interviews consume more time and cost.
- The respondent's responses can be affected by the way the interviewer asks the questions.
- The respondent may refuse to answer some personal questions which are relevant to the study.
- Recording and coding of data during the interview process may sometimes be difficult for the interviewer.
- The interviewer may not have good communication or interactive skills.

Interview Process

The following are the stages in an interview process:

(i) Preparation

The interviewer needs to make certain preparations to make an interview successful. The interviewer should keep all the copies of the schedule or guide ready. He should prepare the lists of respondents with their addresses, contact number and meeting time. He should prepare himself with all the approaches and skills required to conduct an interview. He should prepare himself to face all adverse situations during the interview. If the interviewer is not doing such planning, he can fail to collect the right information from the respondent.

(ii) Introduction

The interviewer is not known to the respondent. Therefore, the interviewer must introduce himself first to every respondent. In the introduction, the interviewer should tell about himself, his organization details and the purpose of his visit. If the interviewer knows someone who the respondent is familiar with, then he can use that person's reference to make the respondent more comfortable. The following are some steps which help in motivating the respondent:

- The interviewer should introduce himself with a smiling face and always greet the respondent.
- He should identify and call the respondent by name.
- He must describe how the respondent is selected.
- He should explain the purpose and usefulness of the study.
- He should focus on the value of the respondent's cooperation.

(iii) Developing rapport

It is important for an interviewer to develop a rapport with the respondent before starting the interview. By doing this, a cordial relationship is established between them. It helps the interviewer understand the inherent nature of the respondent which helps in building a rapport and the discussion can be started with some general topic or with the help of a person who is commonly known to both of them.

NOTES

(iv) Carrying the interview forward

After establishing a rapport, the skills of the interviewer are required to carry the interview forward. The following are some guidelines which should be followed:

- Start the interview in an informal and natural manner.
- Ask all the questions in the same sequence as in the schedule.
- Do not take an answer for granted. It is not necessary that an interviewee will know all answers or will give all answers. The interviewer has to create interest for answering questions.
- The objective of the question should be known to the interviewer to ensure that the correct information is collected for research study.
- Explain the question if it has not been understood properly by the respondent.
- Listen to the respondent carefully with patience.
- Never argue with the respondent.
- Show your concern and interest in the information given by the respondent.
- Do not express your own opinion for answers of any question in the schedule.
- Continue to motivate the respondent.
- If the respondent is unable to frame the right answer, the interviewer should help him by providing alternate questions.
- Ensure that the conversation does not go off the track.
- If the respondent is unable to answer a particular question due to some reasons, drop the question at that moment. This question can be asked indirectly later on.

(v) Recording the interview

Responses should be recorded in the same sequence as they are given by the respondent. The response should be recorded at the same time as it is generated. It may be very difficult to remember all the responses later for recording them. Recording can be done in writing but there may be some problems if the writing skills of an interviewer are not good. Hence, use of electronic devices like tape recorders can help in this purpose. The interviewer should also record all his probes and other comments on the schedule, but they should be in brackets to ensure that they are set off from response.

(vi) Closing the interview

After the interview is over, the interviewer must thank the respondent for his cooperation. He must collect all the papers before leaving the respondent. If the respondent wants to know the result of the survey, the interviewer must ensure that the results are mailed to him when they are ready.

(vii) Editing

At the end, the interviewer must edit the schedule to check that all the questions have been asked and recorded. Also, abbreviations in recording should be replaced by full words.

NOTES

Problems Faced in an Interview

The following are some of the main problems faced in an interview:

(i) Inadequate response

Kahn and Cannel laid down five principal symptoms of inadequate response. They are given as follows:

- **Partial response** in which the respondent gives a relevant but incomplete answer.
- **Non-response** in which the respondent remains silent or refuses to answer the questions.
- **Irrelevant response** in which the respondent's answer is not relevant to the question asked.
- **Inaccurate response** in which the reply is biased.
- **Verbalized response problem** which arises because of the respondent's failure to understand the question.

(ii) Interviewer's biasness, refusal, incapability to understand questions, and so on.

An interviewer can affect the performance of an interview with his own responses and suggestions. Such biasing factors can never be overcome fully, but their effect can be reduced by training and development techniques.

(iii) Non-response

Some respondents out of the total respondents fail to respond to the schedule. The reasons for this non response can be non availability, refusal, incapability to understand questions, and so on.

(iv) Non-availability

Some respondents are not available at their places at the time of call. This could be because of odd timings or working hours.

(v) Refusal

Some respondents refuse to answer the questions. There can be many reasons for this, such as language, odd hours, sickness, no interest in such studies, and so on.

(vi) Inaccessibility

Some respondents can be inaccessible because of various reasons such as migration, touring job, and so on.

Methods and aims of controlling non-response

Researcher Kish suggests the following methods to reduce either the percentage of non-response or its effects:

1. Improved procedure for collecting data is the most obvious remedy for non-response. The improvements advocated are as follows:
   - Guarantee of anonymity
   - Motivation of the respondent to cooperate

- Arousing the respondent's interest by clever opening remarks and questions
- Advance notice to the respondent

2. Call backs are the most effective way of reducing not-at-home responses in personal interviews, as are repeated mailings in no-returns in mail survey..
3. Substitution for non-response is often suggested as a remedy. Usually, this is a mistake because the substitutes resemble the responses rather than the non-responses. Nevertheless, beneficial substitution methods can sometimes be designed with references to important characteristics of population.

Attempts to reduce the percentage or effect of non-response is aimed at reducing the bias caused by vast differences between non-respondents and respondents. The response bias should not be confused with the reduction of sample size due to non-response. The latter effect can be easily overcome either by anticipating the size of non-response in designing the sample size or by compensating for it with a supplement. These adjustments increase the size of the response and the sampling precision, but they do not reduce the non-response percentage or bias.

Telephonic Interview

Telephonic interview is a non-personal method of data collection. It may be used as a major method or supplementary method of data collection. It is useful in the following conditions:

- When the population is composed of those people who are listed in telephone directories
- When less number of questions have to be answered by the respondents
- When the time available for the survey is less
- When the subject is of the interest to the respondent
- When the respondents are widely scattered

Advantages

The following are the advantages of telephonic interviews:

- Less time and low cost
- Good quality of response
- Less demanding on interviewer
- No field work is required
- Easy to contact those respondents who cannot be reached

Disadvantages

Telephonic interviews have the following limitations:

- Restricted to persons who are listed in telephone or other relevant directories
- Not feasible to conduct long interviews
- Limitation of information collected
- No answer to personal questions by respondents
- Respondent's emotions, facial expressions and other environmental factors cannot be recorded
- Difficult to develop rapport

NOTES

Group Interview

Group interview is the method of collecting primary data from a number of individuals with common interests. In group interviews, the interviewer performs the role of a discussion leader. Free discussion is encouraged on the same aspects of the subject under the study. Information is collected either through a self-administered questionnaire or through an interview. Samples for the group can be selected from schools, colleges, clubs and other associations.

Advantages

The following are the advantages of this technique:

- Respondent gets freedom to express his views.
- Flexible method.
- Use of visual aids.
- Less time consuming as group can be interviewed in the time required for one respondent's interview.
- Respondents are more confident in groups.
- Eliminates the limitation of individual interviews.

Disadvantages

The following are the main disadvantages of group interviews:

- Difficulty in selecting the desired sample group.
- Dominance of one individual in a group.
- Respondents can be biased or they can try to please the interviewer or others.

2.3.3 Questionnaire

Primary data can be collected with the help of mails and surveys. The respondents receive the questionnaires from the researcher, and are asked to fill them completely and return them to the researcher. It can be performed only when the respondents are educated. The mail questionnaire should be simple and easy to understand, so that the respondents can answer all questions easily. In mail questionnaires, all the answers have to be given and recorded by the respondents and not by the researcher or investigator, as in the case of personal interview method. There is no face-to-face interaction between the investigator and respondent, and so the respondent is free to give answers of his own choice.

Importance of questionnaires

A questionnaire is a very effective method as well as research tool in any research study. It ensures the collection of a diversified and wide range of scientific data to complete the research objectives. The questionnaire provides all the inputs in the form of relevant data to all statistical methods used in a research study.

Types of Questionnaire

The following are the various categories of questionnaires:

(i) Structured or standard questionnaire

Structured or standard questionnaires contain predefined questions in order to collect the required data for research study. These questions are the same for all the respondents. Questions are in the same language and in the same order for all the respondents.

(ii) Unstructured questionnaire

In unstructured questionnaires, the respondent has the freedom to answer all the questions in his own frame of reference and in his own terms.

Process of Data Collection

The researcher prepares the mailing list by collecting the addresses of all the respondents with the help of primary and secondary sources of data. A covering letter must accompany every questionnaire, indicating the purpose and importance of the research, and importance of cooperation of the respondent for the success of the research study.

Alternate modes of sending questionnaire

The following are the alternate modes of distributing questionnaire to respondents:

(i) Personal delivery

The researcher or investigator himself delivers the questionnaire to the respondents and requests them to fill it within a specific duration, i.e., one day or two days, as per the convenience of the researcher. After the given duration, they collect the questionnaire from the respondents. This adds the advantage of personal interview and mail survey. Alternatively, the questionnaire can be delivered personally to the respondents and the respondents return the questionnaire by mail to the researcher.

(ii) Attaching questionnaire to a product

When a firm is launching a new product or wants to collect the feedback on old products, the firm attaches a questionnaire with its product and requests the customers to fill the questionnaire. The company can give some discount or gift to the respondent of every return questionnaire.

(iii) Advertising the questionnaire

The questionnaire is advertised in magazines and newspapers with instructions to complete it. After filling the questionnaire from the magazine or newspaper, the respondents mail it to the advertiser.

(iv) News-stand inserts

In this method of sending questionnaires to the respondents, the questionnaire, along with covering letter and a self addressed reply-paid envelope, is inserted into a random sample of news-stand copies of a newspaper or magazine.

NOTES

Improving the Response in a Mail Survey

Generally, the response rate in mail surveys in countries like India is very low. The following techniques can be adopted to increase the rate of response:

- **Covering letter:** The covering letter should be in a language which generates the interest of the respondent. It should address the respondent by name.
- **Quality printing:** Sometimes the quality of the printed questionnaire is so bad that the respondent faces a lot of problems in reading it. This results in loss of interest, and so the quality of printing should be excellent and attractive.
- **Prior information:** Prior information can be given to the concerned respondent by telephone, e-mail, newsletters, and so on. Such steps bring more success than follow-ups.
- **Incentives:** Monetary and non-monetary incentives can be given to respondents who are filling the questionnaire. This generates a higher response.
- **Follow-ups:** The respondent can be approached with the help of an investigator to collect the questionnaire or to solve the problems faced by the respondent in filling the questionnaire.
- **Larger sample size:** We should always select a sample size which is larger than what is actually required. This will help the researcher in getting answers from the effective sample size.

Advantages and Disadvantages of Questionnaires

The following are the advantages of questionnaires:

- Low cost
- Wide reach and extensive coverage
- Easy to contact the person who is busy
- Respondent's convenience in completion of the questionnaire
- More impersonal; provides more anonymity
- No interviewer's bias
- Accuracy

The following are the disadvantages of questionnaires:

- Low response by the respondent
- Low scope in many societies where literary level is low
- More time requirement

Preparation of an Effective Questionnaire

While preparing a questionnaire, the researcher must focus on some key parameters to prepare it. These key parameters are as follows:

- Proper use of open and close probe
- Proper sequence of questions
- Use of simple language
- Asking no personal question in which the respondent is hesitating to answer

- Should not be time consuming
- Use of control questions indicating reliability of the respondent

Collecting Data through Schedule

This method is very similar to the collection of data through questionnaires. The only difference is that in schedule, enumerators are appointed. These enumerators go to the respondents, ask the stated questions in the same sequence as the schedule and record the reply of the respondents. Schedules may be given to the respondents and the enumerators should help them solve the problems faced while answering the questions in the given schedule. Thus, enumerator selection is very important in data collection through schedules.

Distinction between schedule and questionnaire

Both questionnaire and schedule are popular methods of data collection. The following are the main differences between a questionnaire and a schedule:

- A questionnaire is generally sent to the respondents through mail, but in case of schedule, it is sent through enumerators.
- Questionnaires are relatively cheaper mediums of data collection as compared to schedules. In the case of questionnaires, the cost is incurred in preparing it and mailing it to the respondent, while in schedule, more money is required for hiring enumerators, training them and incurring their field expenses.
- The response rate in questionnaires is low as many people return it without filling. On the other hand, the response rate in schedules is high because they are filled by the enumerators.
- In collecting data through questionnaires, the identity of the respondent may not be known, but this is not the case when it comes to schedules.
- Data collection through questionnaires requires a lot of time, which is comparatively very less in case of schedules.
- Generally, there is no personal contact in case of questionnaires, but in schedules, personal contact is always there.
- The literacy level of the respondent is very important while filling questionnaires, but in schedules, the literacy level of the respondent is not a major concern as the responses have to be recorded by the enumerators.
- Wider distribution of questionnaires is possible but this is difficult with schedules.
- There is less accuracy and completeness of responses in questionnaires as compared to schedules.
- The success of questionnaires depends on the quality of questions but the success of a schedule depends on the enumerators.
- The physical appearance of questionnaire matters a lot, which is less important in case of schedules.
- Observation method cannot be used along with questionnaires but it can be used along with schedules.

NOTES

Check Your Progress

4. List the advantages of participant observation.
5. What are the uses of observation in business research?
6. State the advantages of questionnaires.

2.4 SAMPLING: INTRODUCTION

NOTES

A part of the population is called sample. Selecting a part of the 'universe' with a view to draw conclusions about the 'universe' or 'population' for a study is known as sampling. A researcher uses sampling for saving time and costs as a selected sample is a replica of the population.

Sampling design: Census and sample survey

All items in any field of inquiry constitute a 'universe' or 'population'. A complete enumeration of all the items in the 'population' is known as a census inquiry. It can be presumed that in such an inquiry, when all the items are covered, no element of chance is left and highest accuracy is obtained. In practice, this may not be true. Even the slightest element of bias in such an enquiry will get larger and larger as the number of observations increase. Moreover, there is no way of checking the element of bias or its extent, except through a resurvey or use of sample checks. Besides, this type of inquiry involves a great deal of time, money and energy. Therefore, when the field of inquiry is large, this method becomes difficult to adopt because of the resources involved. At times, this method is practically beyond the reach of ordinary researchers. Perhaps, the government is the only institution which can get the complete enumeration carried out. Even the government adopts this method in very rare cases, such as population census conducted once in a decade. Further, many a times, it is not possible to examine every item in the population and sometimes it is possible to obtain sufficiently accurate results by studying only a part of the total population. In such cases, there is no utility of census surveys.

However, it needs to be emphasized that when the universe is a small one, it is no use resorting to a simple survey. When field studies are undertaken in practical life, consideration of time and cost invariably leads to a selection of respondents, i.e., selection of only a few items. The respondents selected should act as representatives of the total population in order to produce a miniature cross-section. The selected respondents constitute what is technically called a 'sample' and the selection process is called the 'sampling technique'. The survey so conducted is known as a 'sample survey'.

Algebraically, let the population size be N and if a part of size n (which is < N) of this population is selected according to some rule for studying some characteristics of the population, the group consisting of these n units is known as a 'sample'. The researcher must prepare a sample design for his study, i.e., he must decide how a sample should be selected and of what size such a sample would be.

Implications of a sample design

A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample. Sample design may as well lay down the number of items to be included in the sample, i.e. the size of the sample. Sample design is determined before data are collected. There are many sample designs from which a researcher can choose. Some designs are relatively more precise and easier to apply than others. The researcher must select/prepare a sample design which should be reliable and appropriate for his research study.

Advantages of sampling

The following are the advantages of sampling:

- **Size of population:** It is very difficult to study a large population for a research study; hence, a sample from the population is selected for the study and it represents all characteristics of population.
- **Funds requirement for the study:** When the funds availability is lesser than the anticipated cost of census survey, sampling is an effective method.
- **Facilities:** When facilities like technology and staff members are limited, sampling is preferable.
- **Time:** As the time required for the sampling procedure is less, a researcher prefers this method.

Sampling procedure

Sampling is a complicated process. A researcher has to identify all the factors which can affect the sample. The various criteria related to choice of sampling procedure are as follows:

- **Purpose of survey:** Defining the purpose of a survey helps the researcher in the selection of a particular method of sampling. A particular method of sampling choice depends on the geographical area of the survey, and the size and nature of the study.
- **Measurability:** The application of statistical inference theory requires computation of the sampling error from the sample itself. Probability samples only allow such computation. Hence, where the research objectives require statistical inference, the sample should be drawn by applying simple random sampling method or stratified random sampling method, depending whether the population is homogeneous or heterogeneous.
- **Degree of precision:** A desired level of precision of the result of the survey decides the method adopted for sampling.
- **Information about population:** Details of information available about the population to be studied help in deciding the method of sampling. If no data is available about population, it is difficult to apply probability random sampling. In this condition, the non-probability sampling method can be used for getting an idea of the population.
- **Nature of population:** Whether the population is homogeneous or heterogeneous decides the variables to be studied. Simple random sampling can be used for a homogeneous population. If the population is heterogeneous, stratified random sampling is a better option.
- **Geographical area of study and size of population:** Multi-stage, cluster sampling is used for the study of wide geographical area and large size of population.
- **Financial resources:** Availability of finance decides the need of sampling method.
- **Time limitation:** The time limit to complete a study decides the method of sampling.

NOTES

Characteristics of a good sample

The following are the characteristics of a good sample:

- Representative
- Accuracy
- Precision
- Size

2.5 SAMPLING TECHNIQUES

Sampling methods have been classified in detail in the following section.

2.5.1 Probability and Non-probability

Probability and non-probability sampling is discussed as follows:

1. Probability or random sampling

Probability sampling is based on the theory of probability. It is also known as random sampling. It provides the known non-zero chance of selection to each population element. When generalization is the objective of study and high accuracy of estimation of population parameter is required, random sampling is used. The following are the types of random sampling:

(i) Simple random sampling

It provides each element an equal and independent chance of being selected. Equal chance means equal probability of selection. Independent chance means that draw of one element will not affect the chances of other elements being selected. It is used for small homogeneous population.

Advantages

The following are the advantages of simple random sampling:

- Easy to use
- Equal and independent chance of selection of every element
- No need of prior information of population

Disadvantages

The following are the disadvantages of simple random sampling:

- Impractical because of non-availability of population details
- Does not represent proportionate representation because observations are selected randomly, so there is a chance of selecting same type of observation in one sample, whereas, population may consist of different type of observations
- May be expensive and time consuming.

(ii) Stratified random sampling

When the population to be studied is heterogeneous in nature, it is divided into a homogeneous group or strata, and from each homogeneous group, a random sample is drawn. Stratified random sampling can be classified as follows:

Check Your Progress

7. Define census enquiry.
8. What are the advantages of sampling?
9. List the characteristics of a good sample.

(a) Proportionate stratified random sampling

Proportionate stratified random sampling involves drawing a sample from each stratum in proportion to the latter's share in total population. In this, proper representation is given to each stratum. For example, let us assume that the management faculty of a university consists of the following specialization groups:

Specialization Stream	No. of Students	Proportion of Each Stream
Marketing	40	0.40
HR	20	0.20
Finance	30	0.30
IT	10	0.10
Total	100	1.00

If the researcher wants to draw an overall sample of 40, then the strata sample sizes would be:

Strata	Sample Size
Marketing	$40 \times 0.4 = 16$
HR	$40 \times 0.20 = 8$
Finance	$40 \times 0.30 = 12$
IT	$40 \times 0.10 = 4$
	Total 40

Advantages

The following are the advantages of proportionate stratified sampling:

- Enhancement of representativeness to each sample as proportionate stratified sample consists of observations in proportion of various strata.
- Higher statistical efficiency.
- Easy to carry out.
- Giving self weighing sample.

Disadvantages

The following are the limitations of proportionate stratified sampling:

- Prior knowledge of composition and of distribution of population
- Time consuming and expensive
- Classification error

(b) Disproportionate stratified random sampling

Proportionate representation is not given to strata. It necessarily involves giving over-representation to some strata and under representation to others. The desirability of disproportionate sampling is usually determined by the following three factors:

- The sizes of strata
- Internal variances among strata
- Sampling costs

This method is suitable when the population has some small but important subgroups, when certain groups are homogeneous and it is expected that there will be significant differences in the response of the subgroups in the population.

Advantages

The following are the advantages of disproportionate stratified random sampling:

- Less time consuming
- Appropriate weightage to a particular group which is small but more important

Disadvantages

The following are the limitations of disproportionate stratified random sampling:

- No proportionate representation to each stratum
- Prior knowledge of composition of population required
- Doubtful practical feasibility
- Classification errors

(iii) Systematic random sampling

Systematic random sampling is an alternate to random selection. It consists of taking n item in population after a random start with an item from 1 to n. It is also known as fixed interval method; for example, 1st, 11th, 21st, and so on. It possesses the characteristics of randomness and some non-probability traits.

Systematic selection can be applied to various populations such as students in class, houses in a street, yellow pages, telephone directory, and so on.

Advantages

The following are the advantages of systematic random sampling:

- Simpler than random sampling
- Easy to use
- Easy to instruct
- Less time consuming
- Cost effective
- Statically more efficient

Disadvantages

The following are the disadvantages of systematic random sampling:

- Ignorance of all other elements between two n elements
- Each element does not get an equal chance
- Method gives biased sample

(iv) Cluster sampling

Each sample chapter is a cluster of the population elements in this method. There is random selection of the sampling chapter, which consists of the population elements. Then from each selected sampling chapter, a sample of the population element is drawn.

NOTES

Cluster sampling is used in socio-economic surveys, public opinions polls, ecological studies, farm management services, rural credit services, demographic studies and large scale surveys of political and social behaviour, attitude surveys, and so on.

Advantages

The following are the advantages of cluster sampling:

- Easier and more convenient
- Cost effective
- Convenience of field work as it would be done in compact places
- Less time consumption
- Substitution of units for other units
- More flexible

Disadvantages

The following are the limitations of cluster sampling:

- Variation in cluster size
- Increased bias of resulting sample because of variation
- Sampling error

(v) Area sampling

Area sampling is also a form of cluster sampling. In a large field survey, a cluster consisting of specific geographical areas like districts, tallukas blocks, and villages, in a city are randomly drawn. When geographical areas are selected as sampling units, their sampling is known as area sampling.

(vi) Multi-stage sampling

Sampling is carried out in two or more stages. Firstly, a sample of the first stage sampling chapter is drawn, then from each of the selected first stage sampling units, a sample of the second stage sampling chapter is drawn. The procedure continues up to final sampling units or population elements. An appropriate random sampling method is adopted at each stage.

The population is scattered over a wider geographical area and no details are available for sampling.

Advantages

The following is the advantage of multi-stage sampling:

- Time and cost effective

Disadvantages

The following are the limitations of multi-stage sampling:

- Procedure of estimating sampling error
- Cost advantage is complex

In multi-stage sampling, sampling at the second stage is called sub-sampling.

(vii) Random sampling with probability proportional to size

The procedure of selecting clusters with probability proportional to size (PPS) is used widely. If one primary cluster has twice as large a population as another, it is given twice the chance of being selected. If the same number of people are then selected from each of the selected clusters, the overall probability of any person will be the same. Thus, PPS is a better method of securing a representative sample of population elements in multi-stage cluster sampling. .

Advantages: The advantage of a PPS cluster is that it is more accurate than a simple random sample of clusters.

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

(viii) Double sampling and multi-phase sampling

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

(ix) Replicated or interpenetrating sampling

Replicated sampling involves selection of a certain number of sub-samples rather than one full sample from a population. All the sub-samples should be drawn using the same sampling technique, and each is a self contained and adequate sample of the population. Replicated sampling technique can be used with any basic sampling technique: simple or stratified, single or multi-stage or single or multi-phase sampling. It provides a simple means of calculating the sampling error. It is practical. The replicated samples can throw light on variables and non-sampling errors. The only disadvantage is that it limits the amount of stratification that can be employed.

2. Non-probability or Non-random sampling

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

Advantages: The only merit of this type of sampling is simplicity, convenience and low cost.

Disadvantages: The demerit of this type of sampling is that it does not ensure a selection chance to each population chapter. The selection probability sample may not be a representative one. The selection probability is unknown. It suffers from sampling bias which will distort results.

This sampling method is used when there is no other feasible alternative due to non-availability of a list of population, when the study does not aim at generalizing the findings to the population, when the costs required for probability sampling may be too large and when probability sampling require more time, but the time constraints and the time limit for completing the study do not permit it. It may be classified as follows:

Convenience or accidental sampling

Convenience sampling means selecting sample units in a just 'hit and miss' fashion; for example, interviewing people who we happen to meet. This sampling also means selecting whatever sampling units are conveniently available; for example, a teacher may select students in his class. This method is also known as accidental sampling because the respondents who the researcher meets accidently are included in the sample.

This type of sampling may be used for simple purposes, such as testing ideas or rough impressions about a subject of interest.

Advantage: The main advantage of this method is that it is cheap and simple. Also, it does not require a list of population or statistical expertise.

Disadvantage: The only disadvantage of this method is that it is highly biased because of the researcher's subjectivity. This method is used the least and its findings cannot be generalized.

2.6 SUMMARY

- Survey is an important tool in research. No research can be performed without them.
- There are basically two types of surveys: descriptive and analytic.
- The purpose of case study method is to identify the factors and reasons that account for particular behaviour patterns of a sample chapter and its association with other social or environmental factors.
- In-depth analysis of selected cases is of particular value to business research when a complex set of variables may be at work in generating observed results and intensive study is needed to unravel the complexities.
- The strength of the census survey is associated with generalized characteristics of data. Description of population data acts as a major source of identifying several pertinent issues and questions for research.
- Experimental research refers to the research activity wherein the manipulation of variables takes place, and the resultant effect on other variables is studied.
- A variable is any feature or aspect of an event, function or process that, with its presence and nature, affects some other event or process which is being studied.
- Confounding variables are those aspects of a study or sample that might influence the dependent variable and whose effect may be confused with the effects of the independent variable.
- Internal validity is considered as a property of scientific studies which indicates the extent to which an underlying conclusion based on a study is warranted. This type of warrant is constituted by the extent to which a study minimizes a systematic error or a 'bias'.
- External Validity is considered as the validity of generalized (causal or fundamental) inferences in scientific studies. It is typically based on experiments as experimental validity.
- Focus group as a method was developed in the 1940s in Columbia University by sociologist Robert Merton and his colleagues as part of a sociological technique.

NOTES

Check Your Progress

10. What do you mean by probability sampling?
11. Define double sampling and multi-phase sampling.
12. State the meaning of convenience sampling with a suitable example of a good sample.

NOTES

- A focus group is a highly versatile and dynamic method of collecting information from a representative group of respondents. The process generally involves a moderator who manoeuvres the discussion on the topic under study.
- Observation means specific viewing with the purpose of gathering the data for a specific research study. Observation is a classical method of scientific study.
- Interviewing is a very effective method of data collection. It is a systematic and objective conversation between an investigator and respondent for collecting relevant data for a specific research study.
- Telephonic interview is a non-personal method of data collection. It may be used as a major method or supplementary method of data collection.
- A questionnaire is a very effective method as well as research tool in any research study. It ensures the collection of a diversified and wide range of scientific data to complete the research objectives.
- A part of the population is called sample. Selecting a part of the 'universe' with a view to draw conclusions about the 'universe' or 'population' for a study is known as sampling.
- A sample design is a definite plan for obtaining a sample from a given population. It refers to the technique or the procedure the researcher would adopt in selecting items for the sample.
- Non-probability sampling or non-random sampling is not based on the theory of probability. This sampling does not provide a chance of selection to each population element.

2.7 KEY TERMS

- **Descriptive surveys:** These surveys generally collect information on what people think and do.
- **Analytic surveys:** These surveys are generally used to either test hypotheses or to answer particular research questions.
- **Group interview:** It is a method of collecting primary data from a number of individuals with common interests.
- **Observation:** It means specific viewing with the purpose of gathering the data for a specific research study.
- **Sampling:** It means selecting a part of the 'universe' with a view to draw conclusions about the 'universe' or 'population' for a study.
- **Sample design:** It is a definite plan for obtaining a sample from a given population.

2.8 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Experimental research refers to the research activity wherein the manipulation of variables takes place and the resultant effect on other variables is studied. It provides a logical and structured basis for answering questions.
2. The threats to internal validity are as follows:
   (a) Ambiguous temporal precedence
   (b) Confounding

(c) Selection bias

(d) Repeated testing

(e) Regression towards the mean

3. The different types of focus groups are as follows:
   (a) Two-way focus group
   (b) Dual-moderator group
   (c) Fencing-moderator group
   (d) Friendship groups
   (e) Mini-groups
   (f) Creativity group
   (g) Brand-obsessive group
   (h) Online focus group
4. The following are the main advantages of participant observation:
   (a) In-depth understanding of the respondent group.
   (b) The context which is meaningful to observed behaviour can be recorded or documented by the researcher.
5. Observation is very useful in the following business research purposes:
   (a) Buying behaviour of customer, lifestyles, customs, interpersonal relations, group dynamics, leadership styles, managerial style and actions
   (b) Physical characteristics of inanimate things such as houses, factories, stores, and so on
   (c) Movements in a production plant
   (d) Flow of traffic, crowd and parking on road
6. The following are the advantages of questionnaires:
   (a) Low cost
   (b) Wide reach and extensive coverage
   (c) Easy to contact the person who is busy
   (d) Respondent's convenience in completion of questionnaire
   (e) More impersonal, provides more anonymity
   (f) No interviewer's biasness
   (g) Accuracy
7. A complete enumeration of all the items in the 'population' is known as a census inquiry. It can be presumed that in such an inquiry, when all the items are covered, no element of chance is left and highest accuracy is obtained.
8. The following are the advantages of sampling:
   (a) **Size of population:** It is very difficult to study a large population for a research study; hence, a sample from the population is selected for the study and it represents all characteristics of population.
   (b) **Funds requirement for the study:** When the funds availability is lesser than the anticipated cost of census survey, sampling is an effective method.
   (c) **Facilities:** When facilities like technology and staff members are limited, sampling is preferable.
   (d) **Time:** As the time required for the sampling procedure is less, a researcher prefers this method.

9. The following are the characteristics of a good sample:
   (a) Representative
   (b) Accuracy
   (c) Precision
   (d) Size
10. Probability sampling is based on the theory of probability. It is also known as random sampling. It provides the known non-zero chance of selection to each population element.
11. Double sampling refers to the subsection of the final sample form a pre-selected larger sample that provided information for improving the final selection. When the procedure is extended to more than two phases of selection, it is then called multi-phase sampling.
12. Convenience sampling means selecting sample units in a just 'hit and miss' fashion; for example, interviewing people who we happen to meet. This sampling also means selecting whatever sampling units are conveniently available; for example, a teacher may select students in his class.

2.9 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the various types of surveys?
2. Identify the strengths and limitations of census survey.
3. State the various types of variables.
4. Differentiate between internal and external validity in experimental research.
5. Briefly describe the various stages in an interview process.
6. Tabulate the advantages and limitations of observation.
7. List the various features of interviews.
8. What are the alternate modes of distributing questionnaire to respondents?
9. Distinguish between schedule and questionnaire.
10. List the merits of sampling.

Long-Answer Questions

1. Explain the various stages in the survey method.
2. Describe the key elements of a focus group.
3. What are the various types of interviews? Explain.
4. Identify the problems faced in an interview. Also, distinguish between telephonic interview and group interview.
5. What is the observation method of data collection? Explain its various types.
6. Discuss the various methods of sampling.

2.10 FURTHER READING

Booth, Wayne. 2008. *The Craft of Researcl,*, Third edition. Illinois: University of Chicago Press.

Creswall, John W. 2008. *Research Designs: Quantitative, Qualitative and Mixed Methods Approaches*. London: Sage Publications.

Christenson, Larry B. *et al.* 2010. *Research Methods, Design and Analysis*, Eleventh edition. New Jersey: Allyn and Bacon.

Kothari, C. R. 2008. *Research Methodology: Methods and Techniques*. New Delhi: New Age International.

NOTES

UNIT 3 DESIGN OF RESEARCH

NOTES

Structure

3.0 INTRODUCTION

A research design is a conceptual framework for conducting research. Before embarking on a research, it is imperative that a researcher prepares a design for his research. This is the basic blueprint on which will rest all his future course of actions. Research design as well as sample strategy or sample design form very crucial components of a research process. A research design can be defined as a plan and a systematic procedure for collecting the data and performing analysis on that data for the purpose of research. In social research too, a social scientist needs to prepare a design that provides a direction for analysing the problem, preparing a sample, collecting data from the sample, analysing this data and finally gathering inference from the process. A research design depends to a large extent on the type of research study that is being conducted. If the research study is exploratory, then major emphasis is on the discovery of ideas. The formation of two similar groups that are equivalent to each other is ensured by randomly assigning people or participants into two groups from a common pool of people or participants.

A conclusive research design is more structured and formal than an exploratory research design because it is based on large representative samples and the data obtained is subjected to quantitative analysis. Research design provides the details for how a research study is supposed to be performed. A good research design will ensure that the information gathered is relevant to the research questions, and also that it was collected economically and objectively. The basic requirement of a good research is to provide a framework and guidelines to the researchers in the most accurate and efficient manner.

Research design plays a pivotal ro!e in the entire research process and can be classified into exploratory research design and conclusive research design. In this unit, you will study about the types and significance of research design. Apart from this, you will also learn about the various principles of experimental designs. You will also be introduced to formal experimental designs.

Finally, you will learn about content analysis. This technique involves studying a previously recorded or reported communication, and systematically and objectively breaking it up into more manageable units that are related to the topic under study.

NOTES

3.1 UNIT OBJECTIVES

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

- Define research design and explain its concepts
- Assess the need for formulating a research design
- Discuss the importance of sociometric techniques
- Explain the concept of constructive typology
- Examine the significance of statistical survey
- Differentiate between external criticism and internal criticism in evaluation studies

3.2 RESEARCH DESIGN: DEFINITION AND IMPORTANCE

It is not possible for any researcher to remember all the decisions he has taken. Even if he does remember these, he would have difficulty in understanding how these are interrelated. Therefore, he records all his decisions on a paper or record disc by using relevant concepts or symbols. Such symbolic construction can be called the research design. A research design is a systematic, objective and scientific plan developed for directing a research study. It constitutes the overview for data collection, measurement and analysis of data. Research design is the road map for the functioning of a researcher.

Need for research design

There is a need for research design as it ensures a smooth flow of many research operations, thereby making research as efficient as possible, producing maximum information with minimum effort, time and cost. The ideal design is concerned with specifying the optimum research procedure that could be followed where there are no practical restrictions. To manage with the future changes, a researcher must have a flexible research design. This flexibility ensures the desired achievements in a research. A research design tells the researcher about the methodologies adopted for research work.

Features of a good research design

The following are the features of a good research design:

- Ensuring research progress in the right direction
- Minimizing time and cost of research
- Encouraging coordination and effective organization
- Minimizing bias and maximizing the reliability of the data collected and analysed

3.2.1 Types of Research Designs

Research designs can be categorized as follows:

- Research design in case of exploratory research studies
- Research design in case of descriptive and diagnostic research studies
- Research design in case of hypothesis-testing research studies

Each of these have been explained in detail in the following sections.

1. Research Design in Case of Exploratory Research Design

Formulative research is another term used for exploratory research. The main objective of such studies is problem formation with more precision for research and developing research hypothesis to get the results for operations. The key concern in such type of studies is to generate ideas and finding the insights. Thus, relevant research designs for this type of studies must be flexible to provide an opportunity for various dimensions of the issues under the study. In such studies:

- The sample size is small.
- Non-probability sampling designs are used.
- Data requirements are vague.
- The objective is general rather than specific.
- No definite recommendations are made as a result of the analysis.

The following are the methods of research design for such studies:

- **Survey concerning literature:** This is one of the most uncomplicated and easy methods to formulate the problem with more precision for research and developing research hypothesis to get the results for operations. Hypothesis formulated by previous researchers can be assessed and an evaluation of their importance is done for further research. Many a times, the work of intellectual researchers provides the framework for formulating hypothesis for operations.
- **Experience survey:** It refers to a survey of the respondents who are familiar with the research problem (to be studied). This means that they have already experienced similar problems in past. The main objective of such a survey is to know the relationship between the variable and new ideas related to research problems. In this survey, it is important to select competent people to share their new ideas about the same problem with the researcher.
- **Researcher's interpretation:** These are fruitful methods for selecting the hypothesis for research. This method is suitable in areas where small experience serves as a guide to research study. The detailed study of choice phenomenon in which the researcher wants to research is required. Investigator's attitude, the concentration of the study and the availability of the investigator to draw together diverse information into a united interpretation are the main features of this method.

2. Research Design in Case of Descriptive and Diagnostic Research Studies

Studies describing the individuality of a particular person or group are called descriptive research, whereas research studies defining the occurrence of any happening

NOTES

or association of one happening with others are called diagnostic research. In such studies:

- The study describes the phenomenon under study.
- The collected data may relate to the demographic or the behaviour variables of the respondents under study.
- The research has got very specific objective, clear cut data requirements and uses a large sample which is drawn through probability sampling designs.
- The recommendation/findings in descriptive research are definite.

3. Research Design in Case of Hypothesis Testing Research Studies

Hypothesis testing research studies (also known as experimental studies) are the research studies where the hypothesis is tested to define the causal relationship between variables in an operation.

Principles of Experimental Designs

The three principles enumerated by English statistician Professor R.A. Fisher for experimental design are as follows:

(i) Principle of replication

According to the principle of replication, the same experiment is repeated more than once. Every time, the same experiment is repeated in different experimental units instead of one. By doing so, the numerical precision of the experiments is improved. For instance, let us consider that we have to examine the two ranges of pulse. For this rationale, we divide the entire field into two parts and cultivate one range in one part and the other range in the other part. By comparing the yield of two parts, we can get results for comparative analysis. To apply the principle of replication to this trial, first we divide the entire field into several parts; then cultivate one range in half of these parts and other range in remaining parts. By collecting the statistics yield of two ranges, we can draw a conclusion by comparing the same. Therefore, results are more reliable when we are applying the principle of replication in comparison to the results attained without applying the principle. The more we repeat the experiment, the better the results that we get.

(ii) Principle of randomization

The principle of randomization provides protection against the effects of extraneous factors by randomization when we conduct an experiment. The principle of randomization indicates the need for a design or plans the experiment in such a way that variations caused by extraneous factors can be united under the universal course of chance.

(iii) Principle of local control

Under this principle, the extraneous factors, the identified basis of variability, is made to vary intentionally over as wide a range as necessary, and this needs to be done in such a way that the variability it causes can be measured and, hence, eliminated from the experimental error. This means that we should plan the trial in a manner that we can perform a two-way analysis of variance in which the total variability of data is divided into three components attributed to treatments, the extraneous factors and experimental error.

Formal Experimental Designs

The following are the various formal experimental designs:

- **Completely Randomized Design (CR design):** It involves only two principles, viz., the principle of replication and randomization. The CR design is used when the experimental areas are homogenous for study.
- **Randomized Block Design (RB design):** It is an improvisation of the CR design. Along with the other two principles, local control can be applied in the RB design.
- **Latin Square Design (LS design):** For agriculture-related researches, the LS design is used. This design is used where the researcher desires to control the variation in an experiment that is related to rows and columns in the field.
- **Factorial Design:** Factorial design is used for studies where more factors show more than one effect.

3.3 CONTENT ANALYSIS

Content analysis is a technique which involves studying a previously recorded or reported communication and systematically and objectively breaking it up into more manageable units that are related to the topic under study. It is peculiar in its nature that it is classified as a primary data collection technique and yet makes use of previously produced or secondary data. However, since the analysis is original, first hand and problem specific, it is categorized under primary methods. Some researchers classify it under observation methods, the reason being that in this, one is also analysing the communication in order to measure or infer about variables. The only difference being that one analyses communication that is ex-post facto rather than live. One can content-analyse letters, diaries, minutes of meetings, articles, audio and video recordings, and so on. The method is structured and systematic and, thus, of considerable credibility.

The first step involves defining U, or the universe of content. For example, in the case of Ritu, who wants to know what makes the young Indian tick, she could make use of the blogs written by youngsters, essays and reality shows featuring the age group. She decides that she wants to assess value systems, attitudes towards others/ elders, clarity of life goal and peer influences. This step is extremely critical as this indicates the assumptions or hypotheses the researcher might have formulated.

This universe can be reported in any of five different formats. The smallest reported unit could be a word. This is especially useful as it can be easily subjected to a computer analysis. In Ritu's case, the values that she wants to evaluate are individualistic or collectivistic, aggressive or compliant. Thus, she can shift the communication and place words such as 'I' or 'we' under the respective heads. Words like 'hate' and 'dislike' go under aggression, and 'alright', 'fine' and 'may be not so good' for complacency. Then counts and frequencies are calculated to arrive at certain conclusions.

The next level is a *theme*. This is very useful but a little difficult to quantify as this involves reporting the propositions and sentences or events as representing a theme. For example, disrespect towards elders is the theme and one picks out the following as a representative: a young teenager's blog which says 'My old man (father) has gone senile and needs to be sent to the looney bin for expecting me to become a space scientist, just because he could not become one…'

NOTES

Check Your Progress

1. Define research design.
2. List the features of a good research design.
3. State the main objective of formulative research.
4. Differentiate between descriptive research and diagnostic research.
5. List the various formal experimental designs.

NOTES

This categorization becomes more complex as the element of the observer's bias comes into play. Thus, this kind of analysis could be extremely useful when carried out by an expert. However, in the case of an untrained analyst, the reliability and validity of the findings would be questionable.

The other units are characters, and space and time measures. The character refers to the person producing the communication, for example, the young teenager writing the blog. Space and time are more related to the physical format, i.e., the number of pages used, the length of the communication and the duration of the communication.

The last unit is the item, which is more Gestaltian in nature and refers to categorizing the entire communication as say 'responsible and respectful' or 'aggressive and amoral'. As in the case of theme, this categorization is equally complex as the observer's bias is likely to be high. Thus, to ensure the reliability of the findings, one may ask another coder to evaluate the same data. Well-known researcher J. Cohen states the measuring of the percentage of agreement between the two analyses by the following formula:

$$K = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

Here, Pr(a) is the relative observed agreement between the two raters. Pr(e) is the probability that this is due to chance. If the two raters are in complete agreement, then Kappa is 1. If there is no agreement, then Kappa = 0, 0.21–0.40 is fair, 0.41–0.80 is good and 0.81–1.00 is considered excellent.

Content analysis of large volumes becomes tedious and prone to error if handled by humans. Thus, there are various computer programs available that can assist in the process. For computers running on Windows, one can use TEXTPACK. This is a dictionary word approach, where it can tag defined words for word frequency by sorting them alphabetically or by frequencies. Open-ended questions can be sorted by a program called Verbastat (generally used by corporate users) or Statpac, which has an automatic coding module and is of considerable use to individual researchers.

Content analysis is a very useful technique when one has a large quantity of text as data and it needs to be structured in order to arrive at some definite conclusions about the variables under study. Computer assistance has greatly aided in the active usage of the technique. However, it can appear too simplistic when one reduces the whole data to counts or frequencies.

3.3.1 Sociometric Techniques

According to well-known thinker, U. Bronfenbrenner, sociometry is 'a method for discovering, describing and evaluating social status, structure, and development through measuring the extent of acceptance or rejection between individuals in groups'. Social thinker J. G. Franz defines sociometry as 'a method used for the discovery and manipulation of social configurations by measuring the attractions and repulsions between individuals in a group'.

It is a technique to study the choices a person makes, the way he communicates and interacts with other people in his group. It is concerned with the dynamics between individuals in a group. In this method, a person is asked to select one or more persons from the group, given certain criteria, and it is interesting to note who the person would choose.

NOTES

Eminent sociologist and theorist William J. Goode and others state: 'These and other variants of sociometric techniques offer rather simple methods of ranking individuals on a continuum of 'acceptability' or 'outgoingness' on the part of group members. When their use is justified they may be powerful research tools since they meet the general problems of scaling very well.'

Sociometry Test

The key method used in this technique is the 'sociometric test'. Here, a member of a group is asked to select amongst the other members who they would choose for certain situations. The situation must be a real one to the group under study, e.g., group study, play, classroom seating, class monitor for students of a school, and so on.

The person can be allowed to make two or three choices depending on the size of the group and each choice can be assigned a level of preference. For example, if asked to choose from a group of eight who they would like to work with for a group assignment, the person can choose three people stating his preference by numbering them 1, 2 and 3.

Another example would be, each member of a group consisting of 10 students is asked to write his first, second and sometimes third choices about some significant and pertinent type of social setting. He may be asked questions like:

- Whom would you choose to be the secretary of your debating society?
- Whom would you like to sit next to you in the class or in the bus while going for a picnic?
- With whom do you enjoy the most?
- With whom would you like to work in the science laboratory?
- With whom would you like to walk home?

All these questions are positive questions and, hence, show social acceptances. Negative questions may also be given to show social rejections.

In the above example, the individual has to name three persons in order of preference.

Data may be tabulated as under:

(i) Let the members of the group be numbered from A to J.

(ii) Write 'Choosers' in the vertical column and 'Chosen' in the horizontal column.

(iii) Total choices received by each member are shown at the bottom.

(iv) In the cells, check marks may be shown.

(v) Let 'f' stand for first, 's' for second and 't' for third choices, respectively.

(vi) Add the number of each choice.

A similar table can be prepared for social rejections. In the vertical column will be listed the 'rejecters' and in the horizontal column 'rejectees'.

NOTES

Sociometric Matrix Showing who Chooses Whom

Q.: Whom would you like to be the secretary of your club?

Chosen Choose	A	B	C	D	E	F	G	H	I	J
A			F		S	T				
B				F	S	T				
C	S	F			T					
D			T		S		F			
E			F						S	T
F							F	S		T
G	F			S	T					
H		F			S					T
I					T			S		F
J		F			S			T		
First Choice	1	3	2	1	-	-	2	-	-	1
Second Choice	1	-	-	1	5	-	-	2	1	-
Third Choice	-	-	1	-	3	2	-	1	-	3
Total	2	3	3	2	8	2	2	3	1	4

Guidelines for using this Technique and Interpretation of the Sociometric Matrix

- One person should be concentrated on at a time.
- A detailed study of the choices made and received should be made.
- The 'isolates' and the 'starts' may be looked for. An 'isolate' is one whom nobody chooses. Of course, he is not rejected. A 'star' is a member of the group who receives most of the choices. Here 'E' is the 'star' with maximum choices in his favour. 'I' is an 'isolate' with one choice only.
- Attempts should be made to discover the causes for such selections.

  An individual may be isolated because:

  (a) She is a new member of the group.

  (b) She has a shy and withdrawing nature.

  (c) She does not try to be friends with others.

  (d) She may belong to a lower or upper socio-economic level and, therefore, is not acceptable to the group.
- Look for individuals who select each other. This might be due to factors like:

  (a) Close relations

  (b) Neighbours

  (c) Common interests and the like
- A triangle shows three persons selecting each other. This may be an evidence of cliques or sharp divisions in the group.

Simplest Form of Sociometric Matrix

For finding out the social structure of a group of 10 students, a one line questionnaire is given to the students asking 'Who would you like to be the secretary of your club?

Students would be asked to mark against one of the roll members. Thereafter, the results are tabulated. Figure 3.1 illustrates the sociometric matrix which shows who chooses whom.

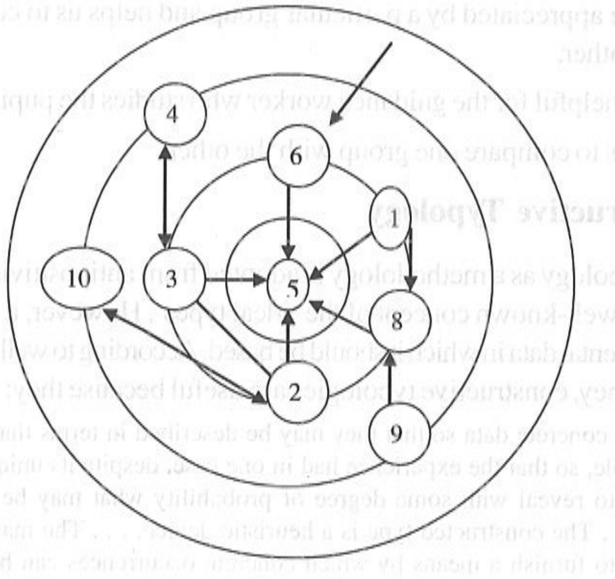


Fig. 3.1 Sociometric Matrix Showing Who Chooses Whom

NOTES

Interpretation

- Student No. 5 is the 'star' as she has been chosen by the maximum number of students.
- Student Nos. 7 and 9 did not get any choice. This indicates that they tended to be isolated, i.e., not being social.
- Student No. 2 and 8 came next to roll No. 1.
- Student No. 5, i.e., the star has preferred roll No. 2.
- Mutual choices were: 2 and 5; 3 and 4; and 2 and 10.

Role of the Research Worker

In general, the research worker can work on three points:

(i) Providing opportunities for developing friendly relations

(ii) Improving social skills of the group

(iii) Building up competency among the group members

Reliable results can be achieved only when all the members constituting a group are fully acquainted with each other. The worker of the counsellor must establish friendly relations with the members of the group so that they may give their frank opinion about an individual or individuals.

Advantages of Sociometry

(i) This technique helps us get an idea of the group at a glance.

(ii) This enables us to form appropriate groups of students for carrying out various activities and projects.

(iii) Such tests at different times enable us to find out the changes taking place in the group structure.

(iv) It is useful in enabling us to understand the characteristics of an individual who is liked or disliked by the group. It gives an insight into the qualities of leadership which are appreciated by a particular group and helps us to compare one group with the other.

(v) It is very helpful for the guidance worker who studies the pupil relationships.

(vi) It helps us to compare one group with the other.

3.3.2 Constructive Typology

Constructive typology as a methodology is adopted from antipositivist sociologist Max Weber Weber's well-known concept of the 'ideal types'. However, it is more integrated with the experimental data in which it should be based. According to well-known sociologist John C. McKinney, constructive typologies are useful because they:

> Order the concrete data so that they may be described in terms that make them comparable, so that the experience had in one case, despite its uniqueness, may be made to reveal with some degree of probability what may be expected in others. . . . The constructed type is a heuristic device. . . . The main purpose it serves is to furnish a means by which concrete occurrences can be compared, potentially measured, and comprehended within a system of general categories which may be developed to comprise the types.

As a methodology in the social sciences, constructive typology has experienced extensive criticism for its importance on description rather than explanation, and also for developing basic classifications rather than testable theories. Nonetheless, if an exploratory study is carried out in the present, precise description should precede any attempts at explanation. According to American political scientist Nelson Polsby:

> Case studies are a practical halfway house between arrant speculation and arid precision. . . . So long as our stock of ideas about policy initiation is relatively primitive, and so long as we are still learning and disagreeing about what a policy is and what an initiation is, the strategy of laying out case studies and searching for ideas about the experience they embody seems not only defensible but desirable.

American social scientist Robert K. Yin suggests that even the most exploratory of the case studies should clearly describe the intention of the of the researcher as in what he wants to investigate, the objective for such exploration and some primary conditions by which the investigation can be considered as either a success or a failure. As opposed to the grounded theory or ethnographic approaches to the case study, the approach adopted by Yin is established on existing theoretical constructs available to the researcher before designing the study. This is, certainly, a most practical approach to 'constructive typology', and is well suited to analysing the careers of theories, as there is an important but dispersed literature relevant to such efforts.

The subject of construct validity, particularly in a case study relating to constructive typology, is important. If the supposed theories are questionable and the so-called careers bear no resemblance to their metaphorical prototypes, the study's purpose is disproved. Yin instructs that the investigator should, in advance, select the specific types of phenomena that are required to be studied and accordingly develop an appropriate operational set of measures. Construct validity can be increased by the use of multiple sources of evidence, the establishment of a 'chain of evidence' and the review of analytic findings by knowledgeable informants.

Yin also proposes four methods for improving internal validity of case studies: These are as follows:

(i) Pattern matching

(ii) Explanation-building

(iii) Time-series analysis

(iv) Logic-modelling

As this was an exploratory, descriptive case study, pattern-matching was considered to be the most approachable method for analysing the resultant data. However, the proposed internal and external theory criteria were intended to provide the future foundations for an ultimately explanatory framework, Moreover, the intrinsically chronological nature of the study itself offers the potential for time-series analysis, and the ultimate use for the constructive typology of 'theory careers' would be a logic model for future studies.

3.3.3 Projective Techniques

There are some other methods that are also used for data collection. These are as follows:

- **Warranty Cards:** These are cards that dealers use for collecting information regarding their products. The information required is printed in the form of questions on the warranty cards, which are placed inside the package along with the product. The consumer is requested to fill the card and post it back to the dealer.
- **Distributor or Store Audits:** Distributors and manufacturers, through their salesmen, conduct distributor or store audits at regular intervals. Retailers also get their stores audited by salesmen and use the information to estimate the market size, market share, seasonal purchasing pattern, and so on. The data is not by questioning, but by observation; for example, while doing an audit for grocery, a sample of stores is visited periodically and data is recorded on inventories either by observation or by copying the data from store records. The advantage of this method is that it offers the most efficient way to evaluate the effect of different techniques on sales.
- **Pantry Audits:** Pantry audit is used to estimate the consumption of goods at the consumer level. In this type of audit, the researcher collects information, such as list of different products, quantities of each product and the prices of each product consumed. All this data is recorded by observing the consumer's pantry. The main objective of a pantry audit is to determine which brand and type of product is being used by which category of consumer, assuming that the contents of the pantry accurately signify their favourites. Pantry audits do not require a series of operations; only one visit is enough to determine the actual preferences of the consumers. An important drawback of pantry audits is that it is not possible to determine the actual preferences of consumers only from the audit data.
- **Consumer Panels:** A consumer panel is an extension of the pantry audit. In this technique, the daily record of a set of consumers is maintained to obtain information about consumer preferences. Later, these records are provided to the officers investigating the consumer preferences. Alternatively, you can say that a consumer panel includes a sample of consumers who are interviewed over a fixed interval of time. Consumer panels are of two types:

NOTES

- o **Transitory Consumer Panel:** A transitory consumer panel is set up in order to determine the influence of a particular phenomenon. A transitory consumer panel is performed on a before-and-after basis. This means that the panel examines the consumer response before and after implementing a particular phenomenon. In this technique, the initial interview of the consumers is conducted before implementing the phenomenon. A second round of interview is conducted after that phenomenon has occurred, to determine the changes in consumer attitude, if any. Such panels are mostly used in advertising and social research.
- o **Continuing Consumer Panel:** A continuing consumer panel is set up for an indefinite period of time to obtain data about certain aspects of the attitude of consumers over a particular period of time. This panel acts as a general-purpose panel to help investigators on different subjects. Such panels are mostly used in the areas of consumer expenditure, public opinion, radio and TV listenership.

- **Use of Mechanical Devices:** Mechanical devices are extensively used to obtain information related to consumers. The devices used for collecting information are as follows:
  - o **Eye Camera:** These are used to collect information about the focus of the respondents on a specific portion of a sketch or diagram or written material. The information collected with the help of eye cameras is used to design advertising material.
  - o **Pupilometric Camera:** These are used to record the dilation of pupils because of a visual stimulus. The extent of dilation of pupils helps determine the amount of interest produced by the stimulus.
  - o **Psychogalavanometer:** It is used to measure the degree of body excitement aroused by a visual stimulus.
  - o **Motion Picture Camera:** It is used to record the body movements language of a buyer when he/she decides whether to buy a particular product from a shop or a big store.
- **Projective Techniques** are also known as direct interviewing techniques. These techniques were developed by psychologists to collect data about the primary reason, desire or intention of respondents by using projections. In a projective technique, while providing information about a particular topic, the respondent automatically projects his/her own attitude or feelings on that subject. The projective technique is mostly used in inspirational research and in attitude surveys. Some of the important projective techniques are as follows:
  - o **Word Association Test:** It is a test that provides information regarding words that have maximum correlation.
  - o **Sentence Completion Test:** It is an extension of the word association test. In this technique, the informant is asked to complete a sentence in order to determine the perception of the informant about a topic.
  - o **Story Completion Test:** It is a technique where the informant is given a story to help focus on the given subject and then asked to give the conclusion for the story.
  - o **Verbal Projection Test:** It is a technique where the informant is asked to give a comment or an explanation on a particular topic.

NOTES

- o **Play Technique:** It is the technique where the informants are given a situation and are asked to perform for improving the situation. For this, the informants are given various roles.
- o **Quizzes, Tests and Examinations:** It is a technique that helps in extracting information regarding specific ability of the candidates on various subjects.
- o **Sociometry:** It is the technique that describes social relationships among individuals in a group.

3.3.4 Statistical Survey

Statistical techniques have contributed greatly in gathering, organizing, analysing and interpreting numerical data. The processing of numerical data through statistics calls for competence in the use of statistical methods and for understanding of concepts that underline their development and their application. The researcher must know the strengths and the weaknesses of the statistical methods which he uses so that he may not mislead or be misled by such methods.

A discussion of two major areas of statistics, descriptive statistics and inferential statistics, is presented in some detail. The main purpose of such discussion is to help the researcher develop an understanding of statistical terminology, and the concepts necessary to study with understanding of the literature dealing with educational research. It also serves to help the student develop competence and know-how to conduct investigations using simple types of statistical analysis.

Organization of Quantitative Data

Organization of data includes editing, classifying and tabulating quantitative information. Editing implies checking of the gathered raw data for accuracy, usefulness and completeness. Classification refers to dividing of the data into different categories, classes, groups or heads. For this, the researcher is guided by the nature of the problem, the hypotheses to be verified or by the responses or characteristics of the samples he has selected. If the problem or hypotheses, for example, involved the difference between attitudes of men and women teachers towards co-education at the secondary school stage, the categories male and female serving in government and private aided schools would be clearly indicated. In some situations when the group is sufficiently homogeneous, no breakdown into categories or subgroups is necessary and it is desirable to describe the group as a whole. However, in the situations where the group is sufficiently heterogeneous, it is desirable to divide the group into homogeneous sub-groups or categories that have in common some distinctive attributes significant for the purpose of analysis.

(i) **Tabulation** is the process of transferring classified data from data gathering tools to the tabular form in which they may be systematically examined. This process may be performed in a number of ways. In simple and less sophisticated types of research, hand-sorting and tabulating procedures are usually employed. More extensive and sophisticated investigations make use of the card-tabulating process.

- ***Hand-Sorting and Hand-Tabulation:*** Hand-sorting and hand-tabulation require careful planning. It includes the method of hand-sorting and recording on tabulation sheets in accurate mathematical terms by marking and counting frequency tallies for different items on which information is sought. The sorting of response sheets in case of psychological tests or scales in

NOTES

various categories must be done before the tabulation of responses. At times without proper planning, a researcher may waste his time and energy by tabulating the responses first before it occurs to him than it would be interesting to compare the responses of the various sub-groups comprising the sample under investigation. This process would require another handling of the response sheets, scales or opinionnaires and would involve reticulating the responses.

- ***Modern Computational Mechanical Aids:*** Modern computational mechanical aids are a boon to the modern researcher. They are used to save time and effort, and to minimize error during the organizing and analysis of research data. The increasing and popular use of these computational devices has advanced educational research in terms of both quality and quantity. The computational mechanical aids commonly used are 'calculators' and 'computers'.

(ii) Calculators

The most common computational mechanical device available to the researcher is the calculator. Its principal advantages are speed and accuracy in performing addition, subtraction, multiplication and division tasks. These operations are performed easily, merely by the pressing of the necessary keys to enter the data and another key to begin the desired operation. The calculations involving combinations of the fundamental operations can also be performed by setting their order as required in computational problem. The desk calculator provides reliable results. At times, improper input of the data or incorrect operations of the machine, or both, furnish erroneous result.

The electromechanical calculators perform the calculations by electrically operated mechanical devices. On the other hand, electronic calculators developed recently operate electronically and perform calculations without the use of mechanical counters and with greater speed. Some of these electronic calculators are capable of performing operations beyond the four basic operations of addition, subtraction, multiplication and division. These additional operations include interpolation, extraction of square roots and reciprocals.

The manufacturers of calculators usually provide instruction manuals with them for the use of their operators. These manuals provide directions even for simple operations. If they are studied carefully, the user may not face any difficulty in performing any operation.

(iii) Computers

A computer system operates in accordance with specific instructions. Each instruction defines an operation to be performed. It also specifies the data, device or mechanism needed to carry out the operation. These instructions are referred to as a program. A computer is useless until a programmer writes a detailed set of instructions to be loaded into its internal storage (memory) unit. There has been a revolution in the field of information technology in recent times. Simultaneously, programming of computers has made it easy to analyse data. Statistical Programming in Social Sciences (SPSS) is used by researchers to analyse and interpret the results. Another program is MS EXCEL which can analyse large volumes of data.

The researcher should keep the following factors in view while interpreting the results:

- **Influence of Unstudied Factors:** In any type of educational research, the researcher is generally guided by the factors or variables which he has studied during the research process. He totally ignores the influence or effect of unstudied factors while interpreting the results of his study. To totally ignore the unstudied factors and ascribe the findings of the research to the occurrence of studied factors alone may be misinterpreting the actual truth, for the findings in any research are conditioned not by one or two but innumerable variables. It is truer in the case of experimental or causal-comparative type of research in which the researcher studied a very limited number of variables. For example, a researcher, finds that a group of eighth class students following programmed instruction material in social studies has performed better compared with another group of students of the same class taught through lecture method. If he were to ascribe the better achievement of the first group to the method alone and ignore the other possible determining factors like high general mental ability, high achievement motivation, better study habits, interest in the subject and better socio-economic conditions found among the higher achieving group, he will be misinterpreting the truth.
- **Selective Factors:** A researcher may hideously misrepresent the truth if he ignores the selective factors. This is more evident in the studies where a selective group is made the subject of investigation or where a particular factor is operating in the situation studied. For example, if a researcher finds that the boys of a particular tribe are mostly low in intelligence and then concludes that the boys of all tribes have a low intelligence, then the researcher is ignoring the fact that there exist outside the particular tribe, many tribal boys with average or high intelligence. Similarly, to find that in a particular secondary school, the number of the tenth class students failing in mathematics is greater than the number of students failing in other subjects and to conclude from this that mathematics is comparatively more difficult than other subjects of study, then it is ignoring the fact that the students of mathematics did not receive good instruction in the subject.
- **Expected Results:** While interpreting the expected results, the researcher has to keep in mind that he does not go beyond his data support and that he does not forget the limitations of the study. The researcher has to be cautious in reporting all such factors which could account for the results.
- **Negative Results:** Researchers, often, on arriving at results contrary to what they had hypothesized, jump to develop a sort of defiance mechanism by exaggerating all the factors that could have possibly vitiated the results. They often list shortcomings in terms of the use of inadequate tools or sample fluctuations. These things may be true and there is no harm in reporting all such factors which come in the way of making the study precise. Nevertheless, it is not always correct to get results that confirm hypotheses. Hypotheses arise from guesswork and cannot be accepted as correct without being tested for confirmation. Only after the research is completed is the researcher in a position to declare his results with certainty. When the results contradict the original hypothesis of the study, the interpretation and discussion of results should include the researcher's reconsideration of the original

hypothesis in the light of his findings. At times, researchers are reluctant to discuss results that contradict the existing known facts. This attitude is not fair and is likely to impede the progress of research. It must be noted that hypotheses are tentative and results can differ from them.

- **Results when the Null Hypothesis is Retained:** A retained null hypothesis may occur when:

  (a) There is no relationship between the variables; or the experimental variable is not more effective than the control variable.

  (b) The null-hypothesis is false, but the internal validity problems of the data contaminated the investigation so badly that the actual relationship between variables could not be established.

3.3.5 Evaluation Studies

The main feature of historical research is the evaluation of historical data. The backbone of historiography is the authenticity of data collected through different sources. Even when the data are collected through different sources, doubts can be raised about their validity, reliability and relevance. The process of judging validity, reliability and relevance of data is carried out through two devices, viz., (a) External criticism and (b) Internal criticism.

(a) External Criticism

External criticism is also known as lower criticism. It involves testing the sources of data for integrity, i.e., every researcher must test the information received to ensure that any source of data is in fact what it seems to be. External criticism helps to determine whether it is what appears or claims to be and whether it reads true to the original so as to save the researcher from being the victim of fraud. On the whole, the general criteria followed for such criticism depends on:

- A good chronological sense, a versatile intellect, common sense, an intelligent understanding of human behaviour, and plenty of patience and persistence on the part of the researcher
- Recent validation of the quality of the source
- A good track record of the source

This information may be found in relevant literature. Thereafter, these literary sources can be verified for genuineness of content by verifying signatures, handwriting, writing styles, language, and so on. Further, material sources of information can be verified through physical and chemical tests on the ink, paint, paper, cloth, metal, wood, and so on.

(b) Internal Criticism

After the integrity of the data sources are established, the actual data content is subject to verification. This process is known as the internal criticism of the data. It is also called higher criticism which is concerned with the validity, truthfulness or worth of the content of document.

At the outset, the information obtained through a particular source is examined for internal consistency. The higher the internal consistency, the greater the accuracy. The researcher should establish the literal as well as the real meaning of the content within its historical context.

This is followed by an evaluation of the external consistency of the data. This is important because, although the authorship of a report is established, the report may comprise distorted pictures of the past. For verifying that the content is accurate, the researcher should first compare the information received through two independent sources and then match new information obtained with the information already on hand which has been tested for reliability. Three major principles need to be followed in order to establish external consistency of the data: (i) data from two independent sources to be matched for consistency, (ii) data must have been obtained from at least one independent primary source, and (iii) data should not be gathered from a source that has a track record of providing contradictory information. It is recommended that the researcher should apply his professional knowledge and judgment to make a final evaluation in case it is not possible to find matching information from two comparable sources.

NOTES

The following series of questions have been listed by well-known research thinkers Cater V. Good, A. S. Barr and Douglas E. Sates to guide a researcher in the process of external and internal criticism of historical data:

- Who was the author, not merely what his name was but what his personality, character and position were like, and so on?
- What were his general qualifications as a reporter—alertness, character and bias?
- What were his special qualifications as a reporter of the matters here treated?
- How was he interested in the events related?
- Under what circumstances was he observing the events?
- Had he the necessary general and technical knowledge for learning and reporting the events?
- How soon after the events was the document written?
- How was the document written, from memory, after consultation with others, after checking the facts or by combining earlier trial drafts?
- How is the document related to other documents?
- Is the document an original source—wholly or in part? If the latter, what parts are original and what borrowed? How credible are the borrowed materials? How accurately is the borrowing done? How is the borrowed material changed and used?

Perpetually, the researcher needs answers for all these questions and, therefore, he has to depend, somewhat, upon evidence he can no longer verify. At times, he will have to rely on the inferences based upon logical deductions in order to bridge the gaps in the information.

3.4 SUMMARY

- A research design is a systematic, objective and scientific plan developed for directing a research study. It constitutes the overview for data collection, measurement and analysis of data.
- To manage with the future changes, a researcher must have a flexible research design. This flexibility ensures the desired achievements in a research.
- Formulative research is another term used for exploratory research. The main objective of such studies is problem formation with more precision for research and developing research hypothesis to get the results for operations.

Check Your Progress

6. What do you mean by sociometry?
7. What is the role of a research worker?
8. State the advantages of sociometry.
9. What is pantry audit? State its objective.
10. Define tabulation.

NOTES

- Studies describing the individuality of a particular person or group are called descriptive research, whereas research studies defining the occurrence of any happening or association of one happening with others are called diagnostic research.
- Hypothesis testing research studies (also known as experimental studies) are the research studies where the hypothesis is tested to define the causal relationship between variables in an operation.
- Content analysis is a technique which involves studying a previously recorded or reported communication and systematically and objectively breaking it up into more manageable units that are related to the topic under study.
- Content analysis of large volumes becomes tedious and prone to error if handled by humans. Thus, there are various computer program available that can assist in the process.
- Content analysis is a very useful technique when one has a large quantity of text as data and it needs to be structured in order to arrive at some definite conclusions about the variables under study.
- Sociometry is a technique to study the choices a person makes, the way he communicates and interacts with other people in his group. It is concerned with the dynamics between individuals in a group.
- Constructive typology as a methodology is adopted from antipositivist sociologist Max Weber's well-known concept of the 'ideal types'.
- As a methodology in the social sciences, constructive typology has experienced extensive criticism for its importance on description rather than explanation, and also for developing basic classifications rather than testable theories.
- Distributors and manufacturers, through their salesmen, conduct distributor or store audits at regular intervals. Retailers also get their stores audited by salesmen and use the information to estimate the market size, market share, seasonal purchasing pattern, and so on.
- A consumer panel is an extension of the pantry audit. In this technique, the daily record of a set of consumers is maintained to obtain information about consumer preferences.
- Mechanical devices are extensively used to obtain information related to consumers. The devices used for collecting information are eye camera, pupilometric camera, psychogalavanometer and motion picture camera.
- Statistical techniques have contributed greatly in gathering, organizing, analysing and interpreting numerical data.
- Organization of data includes editing, classifying and tabulating quantitative information. Editing implies checking of the gathered raw data for accuracy, usefulness and completeness.
- Tabulation is the process of transferring classified data from data gathering tools to the tabular form in which they may be systematically examined.
- The most common computational mechanical device available to the researcher is the calculator. Its principal advantages are speed and accuracy in performing addition, subtraction, multiplication and division tasks.

- Programming of computers has made it easy to analyse data. Statistical Programming in Social Sciences (SPSS) is used by researchers to analyse and interpreting the results. Another program is MS EXCEL which can analyse large volumes of data.
- The main feature of historical research is the evaluation of historical data. The backbone of historiography is the authenticity of data collected through different sources.

3.5 KEY TERMS

- **Hypothesis testing:** It is a process by which an analyst tests a statistical hypothesis. The methodology employed by the analyst depends on the nature of the data used and the goals of the analysis.
- **Probability sampling:** It is a sampling technique wherein the samples are gathered in a process that gives all the individuals in the population equal chances of being selected.
- **Factorial design:** It involves having more than one independent variable, or factor, in a study.
- **Sociometry:** It is a quantitative method for measuring social relationships. It was developed by psychotherapist Jacob L. Moreno in his studies of the relationship between social structures and psychological well-being.
- **Grounded theory:** It is a research method that involves forming a theory based on the gathered data as opposed to gathering data after forming a theory.
- **Psychogalvanometer:** It is a device used in determining the changes in the electrical resistance of the skin in response to emotional stimuli.
- **Educational research:** It is the scientific field of study that examines education and learning processes and the human attributes, interactions, organizations, and institutions that shape educational outcomes.
- **Interpolation:** It is an estimation of a value within two known values in a sequence of values.

3.6 ANSWERS TO 'CHECK YOUR PROGRESS'

1. A research design is a systematic, objective and scientific plan developed for directing a research study. It constitutes the overview for data collection, measurement and analysis of data. Research design is the road map for the functioning of a researcher.
2. The following are the features of a good research design:
   (a) Ensuring research progress in the right direction
   (b) Minimizing time and cost of research
   (c) Encouraging coordination and effective organization
   (d) Minimizing bias and maximizing the reliability of the data collected and analysed
3. The main objective of formulative research is problem formation with more precision for research and developing research hypothesis to get the results for operations.

NOTES

The key concern in such type of studies is to generate ideas and finding the insights.

4. The following are the various formal experimental designs:
   (a) Completely Randomized Design (CR design)
   (b) Randomized Block Design (RB design)
   (c) Latin Square Design (LS design)
   (d) Factorial Design

5. Studies describing the individuality of a particular person or group are called descriptive research, whereas research studies defining the occurrence of any happening or association of one happening with others are called diagnostic research.

6. Sociometry is a technique to study the choices a person makes, the way he communicates and interacts with other people in his group. It is concerned with the dynamics between individuals in a group. In this method, a person is asked to select one or more persons from the group, given certain criteria, and it is interesting to note who the person would choose.

7. In general, a research worker can work on three points:
   (a) Providing opportunities for developing friendly relations
   (b) Improving social skills of the group
   (c) Building up competency among the group members

8. The advantages of sociometry are as follows:
   (a) This technique helps us get an idea of the group at a glance.
   (b) This enables us to form appropriate groups of students for carrying out various activities and projects.
   (c) Such tests at different times enable us to find out the changes taking place in the group structure.
   (d) It is useful in enabling us to understand the characteristics of an individual who is liked or disliked by the group. It gives an insight into the qualities of leadership which are appreciated by a particular group and helps us to compare one group with the other.
   (e) It is very helpful for the guidance worker who studies the pupil relationships.
   (f) It helps us to compare one group with the other.

9. Pantry audit is used to estimate the consumption of goods at the consumer level. In this type of audit, the researcher collects information, such as list of different products, quantities of each product and the prices of each product consumed. All this data is recorded by observing the consumer's pantry. The main objective of a pantry audit is to determine which brand and type of product is being used by which category of consumer, assuming that the contents of the pantry accurately signify their favourites.

10. Tabulation is the process of transferring classified data from data gathering tools to the tabular form in which they may be systematically examined. This process may be performed in a number of ways. In simple and less sophisticated types of research, hand-sorting and tabulating procedures are usually employed. More extensive and sophisticated investigations make use of the card-tabulating process.

3.7 QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. What are the various methods of research design?
2. State the principles of experimental design.
3. Write a short note on sociometric techniques.
4. List some of the important projective techniques.
5. When can a retained hypothesis occur?
6. Briefly describe the importance of statistical survey.
7. What are hand-sorting and tabulating procedures?
8. State the factors that a researcher should keep in view while interpreting research results.

Long-Answer Questions

1. Explain the different types of research designs.
2. Describe the concept of content analysis in research design.
3. Critically analyse the importance of constructive typology in research design.
4. Discuss the various methods used for data collection.
5. What role does evaluation play in research? How is the process of evaluation carried out?

3.8 FURTHER READING

Kothari, C. R. 1995. *Research Methodology–Methods and Techniques*. New Delhi: Wiley Eastern Ltd.

Creswall, John W. 2008. *Research Designs: Quantitative, Qualitative and Mixed Methods Approaches*. London: Sage Publications.

Christenson, Larry B. *et al.* 2010. *Research Methods, Design and Analyśis*, Eleventh edition. New Jersey: Allyn and Bacon.

Wilkinson, T. S. and P. L. Bhandarkar. 2003. *Methodology and Techniques of Social Research.* Mumbai: Himalaya Publishing House.

Chaudhary, C. M. 1991. *Research Methodology.* Rajasthan: RBSA Publishers.

Gupta, S. C. and V. K. Kapoor. 1996. *Fundamentals of Applied Statistics.* New Delhi: Sultan Chand & Sons.

[illegible]

[illegible]

1. [illegible]
2. [illegible]
3. [illegible]
4. [illegible]
5. [illegible]
6. [illegible]
7. [illegible]
8. [illegible]

[illegible]

1. [illegible]
2. [illegible]
3. [illegible]
4. [illegible]
5. [illegible]

FURTHER READING

[illegible] Wiley Eastern Ltd.

[illegible]

[illegible]

[illegible]

[illegible]

Carter S.C. and [illegible]

UNIT 4 DATA ANALYSIS AND PRESENTATION

NOTES

Structure

4.0 INTRODUCTION

When data is transformed to extract useful information, it is known as analysis of data. The process of analysis facilitates the discovery of some useful conclusions. Finding conclusions from the analysed data is known as interpretation of data. The search for knowledge is referred to as research. Research can also be defined as an art of scientific

investigation. Within the academic scenario, research comprises defining and redefining problems, formulating hypothesis or suggested solutions; collecting, organizing and evaluating data; making deductions and reaching conclusions; and in the end, carefully testing the conclusions to determine whether they fit the formulating hypothesis. Research generally begins with a question or a problem. The purpose of research is to find solutions through the application of systematic and scientific methods. The unit explains the meaning, process and purpose of research.

This unit will also discuss the measures of central tendency, which are of various types, such as arithmetic mean, mode and median. This is also commonly known as simply the mean. Even though average, in general, means any measure of central location, when we use the word average in our daily routine, we always mean the arithmetic average. The term is widely used by almost everyone in daily communication. A measure of dispersion, or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency. Some common measures of dispersion are the range, the semi-interquartile range or the quartile deviation, the mean deviation and the standard deviation. Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable.

Analysis of data is a process of inspecting, cleaning, transforming and modelling data with the goal of highlighting useful information, suggesting conclusions and supporting decision-making. The term 'data analysis' is sometimes used as a synonym for data modelling. Data analysis is still necessary to demonstrate the experimental hypotheses. A hypothesis is an assumption that is tested to find its logical or empirical consequence. It refers to a provisional idea whose merit needs evaluation, but which has no specific meaning. A hypothesis should be clear and accurate. Various concepts such as null and alternative hypotheses help verify the testability of an assumption. During the course of hypothesis testing, some inference about population, like the mean and proportion are made. Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected.

In this unit, you will also learn about the importance of research reports and the procedure followed for report writing. A report should not only be prepared in a structured, presentable and understandable way but the data provided should also be reliable and statistically analysed. The report should be written in a proper style. This unit discusses the various types of research report. It also deals with the importance of interpretation, the techniques of interpretation and the precautions that need to be taken while preparing research reports. The significance of scientific writing as well as the steps involved in report writing will be suggested, followed by the layout plan of a research report.

Finally, the unit will discuss the concepts of tables and charts in data analysis and presentation. Classification of data is usually followed by tabulation, which is considered as the mechanical part of classification. In a graph, the independent variable should always be placed on the horizontal or x-axis and the dependent variable on the vertical or y-axis. A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations. Two-dimensional diagrams take two components of data for representation. These are also called area diagrams as it considers two dimensions. The types are rectangles, squares and pie. Three-dimensional diagrams are also termed as volume diagram and consist of cubes, cylinders, spheres, and so on. In these diagrams, three dimensions namely length, width and height are taken into account.

4.1 UNIT OBJECTIVES

NOTES

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

- Define the identification process and formulation of research problem
- Explain the measures of central tendency
- Discuss the measures of dispersion
- Describe about the simple and multiple regression analysis
- Assess the importance of statistical hypotheses in data analysis
- Explain the proper layout of a report and list the types of reports

4.2 PROBLEM MEASUREMENT

Research problems are questions that indicate gaps in the scope or the certainty of our knowledge. They point either to the problematic phenomena, observed events that are puzzling in terms of the accepted ideas or to problematic theories and current ideas that are challenged by new hypothesis.

Defining the Research Problem

Problem discovery puts the research process into action and identification of the problem is the first step towards its solution. Properly and completely defining a business problem is easier said than done. Actually, the research task may be to define or evaluate an opportunity or to clarify a problem. The definition and discovery of the research problem is viewed under this broader context. In research, often, only symptoms are apparent to begin with. The adage 'A problem well defined is a problem half solved' is worth remembering. The investigation gets a sense of direction with an orderly definition of the research problem. A careful attention to the problem definition allows a researcher to set the proper research objectives. When the purpose of research is clear, the chances of collecting the relevant and necessary information are greater. However, just because a problem has been discovered or an opportunity has been recognized does not mean that the problem has been defined. A problem definition indicates a specific managerial decision area to be clarified or a particular problem to be solved. It specifies research questions to be answered and the objectives of the research.

Problem Identification and Formulation

The first and the most important step of the research process is to identify the path of enquiry in the form of a research problem. It is like the onset of a journey, in this instance the research journey, and the identification of the problem gives an indication of the expected result being sought. A research problem can be defined as a gap or uncertainty in the decision-makers' existing body of knowledge which inhibits efficient decision-making. Sometimes it may so happen that there might be multiple reasons for these gaps, and identifying one of these and pursuing its solution might be the problem. According to educational researcher Fred Kerlinger *'If one wants to solve a problem, one must generally know what the problem is. It can be said that a large part of the problem*

NOTES

lies in knowing what one is trying to do'. The defined research problem might be classified as simple or complex. Simple problems are those that are easy to comprehend, and their components, and identified relationships are linear and easy to understand, e.g., the relation between cigarette smoking and lung cancer. Complex problems, on the other hand, talks about interrelationship between antecedents and subsequently with the consequential component. Sometimes the relation might be further impacted by the moderating effect of external variables as well, e.g., the effect of job autonomy and organizational commitment on work exhaustion, at the same time considering the interacting (combined) effect of autonomy and commitment. This might be further different for males and females. These kinds of problems require a model or framework to be developed to define the research approach.

Thus, the significance of a clear and well-defined research problem cannot be overemphasized, as an ambiguous and general issue does not lend itself to scientific enquiry. Even though different researchers have their own methodology and perspective in formulating the research topic, a general framework which might assist in problem formulation are discussed below.

Problem Identification Process

The problem recognition process invariably starts with the decision-maker and some difficulty or decision dilemma that he/she might be facing. This is an action oriented problem that addresses the question of what the decision-maker should do. Sometimes, this might be related to actual and immediate difficulties faced by the manager (applied research) or gaps experienced in the existing body of knowledge (basic research). The broad decision problem has to be narrowed down to information oriented problem which focuses on the data or information required to arrive at any meaningful conclusion. Given in Figure 4.1 is a set of decision problems and the subsequent research problems that might address them.

Management Decision Problem

The entire process explained above begins with the acknowledgement and identification of the difficulty encountered by the business manager/researcher. If the manager is skilled enough and the nature of the problem requires to be resolved by him or her alone, the problem identification process is handled by him or her, else he or she outsources it to a researcher or a research agency. This step requires the author to carry out a problem appraisal, which would involve a comprehensive audit of the origin and symptoms of the diagnosed business problem. For illustration, let us take the first problem listed in the Figure 4.1. An organic farmer and trader in Uttarakhand, Nirmal farms, wants to sell his organic food products in the domestic Indian market. However, he is not aware if this is a viable business opportunity, and since he does not have the expertise or time to undertake any research to aid in the formulation of the marketing strategy, he decides to outsource the study.

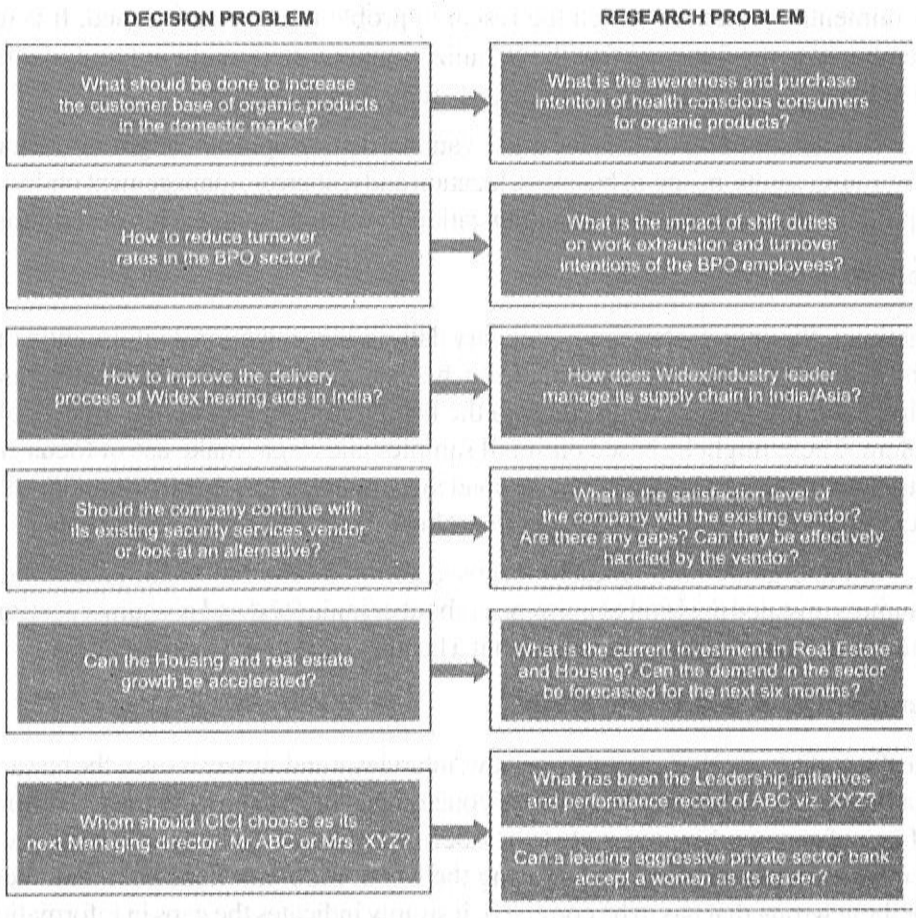

DECISION PROBLEM
RESEARCH PROBLEM
What should be done to increase the customer base of organic products in the domestic market?
What is the awareness and purchase intention of health conscious consumers for organic products?
How to reduce turnover rates in the BPO sector?
What is the impact of shift duties on work exhaustion and turnover intentions of the BPO employees?
How to improve the delivery process of Widex hearing aids in India?
How does Widex/industry leader manage its supply chain in India/Asia?
Should the company continue with its existing security services vendor or look at an alternative?
What is the satisfaction level of the company with the existing vendor? Are there any gaps? Can they be effectively handled by the vendor?
Can the Housing and real estate growth be accelerated?
What is the current investment in Real Estate and Housing? Can the demand in the sector be forecasted for the next six months?
Whom should ICICI choose as its next Managing director- Mr ABC or Mrs. XYZ?
What has been the Leadership initiatives and performance record of ABC viz. XYZ?
Can a leading aggressive private sector bank accept a woman as its leader?

Fig. 4.1 Converting Management Decision Problem into Research Problem

Discussion with Subject Experts

The next step involves getting the problem in the right perspective through discussions with industry and subject experts. These individuals are knowledgeable about the industry as well as the organization. They could be found both within and outside the company. The information on the current and probable scenario required is obtained with the assistance of a semi-structured interview. Thus, the researcher must have a predetermined set of questions related to the doubts experienced in problem formulation. It should be remembered that the purpose of the interview is simply to gain clarity on the problem area and not to arrive at any kind of conclusions or solutions to the problem. For example, for the organic food study, the researcher might decide to go to food experts in the Ministry for Food and Agriculture or agricultural economists or retailers stocking health food as well as doctors and dieticians. This data, however, is not sufficient in most cases, while in other cases, accessibility to subject experts might be an extremely difficult task as they might not be available. The information should, in practice, be supplemented with secondary data in the form of theoretical as well as organizational facts.

Organizational Analysis

Another significant source for deriving the research problem is the industry and organizational data. In case, the researcher/investigator is the manager himself/herself, the data might be easily available. However, in case the study is outsourced, the detailed background information of the organization must be compiled, as it serves as the

NOTES

environmental context in which the research problem has to be defined. It is to be remembered at this juncture that the organizational context might not be essential in case of basic research, where the nature of study is more generic.

This data needs to include the organizational demographics—origin and history of the firm; size, assets, nature of business, location and resources; management philosophy and policies as well as the detailed organizational structure, with the job descriptions.

Qualitative Survey

Sometimes, the expert interview, secondary data and organizational information might not be enough to define the problem. In such a case, an exploratory qualitative survey might be required to get an insight into the behavioural or perceptual aspects of the problem. These might be based on small samples and might make use of focus group discussions or pilot surveys with the respondent population to help uncover relevant and topical issues which might have a significant bearing on the problem definition.

In the organic food research, focused group discussions with young and old consumers revealed the level of awareness about organic food and consumer sentiments related to purchase of more expensive but a healthy alternative food product.

Management Research Problem

Once the audit process of secondary review, interviews and survey is over, the researcher is ready to focus and define the issues of concern that need to be investigated further, in the form of an unambiguous and clearly-defined research problem. Once again, it is essential to remember that simply using the word 'problem' does not mean there is something wrong that has to be corrected; it simply indicates the gaps in information or knowledge base available to the researcher. These might be the reason for his inability to take the correct decision. Second, identifying all possible dimensions of the problem might be a monumental and impossible task for the researcher. For example, the lack of sales of a new product launch could be due to consumer perceptions about the product, ineffective supply chain, gaps in the distribution network, competitor offerings or advertising ineffectiveness. It is the researcher who has to identify and then refine the most probable cause of the problem and formalize it as the research problem. This would be achieved through the four preliminary investigative steps indicated above.

Statement of Research Objectives

Next, the research question(s) that were formulated need to be broken down and spelt out as tasks or objectives which are need to be met in order to answer the research question.

Based on the framework of the study, the researcher has to numerically list the thrust areas of research. This section makes active use of verbs such as 'to find out', 'to determine', 'to establish' and 'to measure' so as to spell out the objectives of the study. In certain cases, the main objectives of the study might need to be broken down into sub-objectives which clearly state the tasks to be accomplished.

In the organic food research, the objectives and sub-objectives of the study were as follows:

1. ***To Study the Existing Organic Market*:** This would involve:
   - To categorize the organic products available in Delhi into grain, snacks, herbs, pickles, squashes, fruits and vegetables

NOTES

- To estimate the demand pattern of various products for each of the above categories
- To understand the marketing strategies adopted by different players for promoting and propagating organic products

2. ***Consumer Diagnostic Research:*** This would entail:
   - To study the existing consumer profile, i.e., perception and attitudes towards organic products and purchase and consumption patterns
   - To study the potential customers in terms of consumer segments, level of awareness, perception and attitude towards health and organic products
3. ***Opinion Survey:*** To assess the awareness and opinions of experts such as doctors, dieticians and chefs in order to understand organic consumption and propagation
4. ***Retail Market:*** This would involve:
   - To find the gap between demand and supply for existing retailers
   - To forecast demand estimates by considering the existing as well as potential retailers

Thus, the research problem formulation involves the following interrelated steps:

- Ascertaining the objectives of the decision-maker
- Understanding the problem's background
- Identifying and isolating the problem, rather than its symptoms
- Determining the unit of analysis
- Determining the relevant variables
- Stating the research objectives and research questions (hypotheses)

The above-mentioned process ensures that the real research objectives/questions are identified for the proposed research.

Statement of Research Problem

Both the decision-makers and researchers expect that the problem definition efforts should result in a statement of the research problem or research objectives. On completion of the exercise of formulating the research problem, the researcher must prepare a written statement(s) that clarifies any ambiguity about what s/he hopes the research will accomplish. Writing a series of research questions and hypotheses can add clarity to the statement of the business problem. These research questions are the researcher's translation of the business problem into a specific need for inquiry; hypothesis is an unproven proposition that tentatively explains certain facts or phenomena, a proposition that is empirically testable. In other words, research objectives/hypotheses explain the purpose of research in measurable terms and define standards what the research should accomplish.

Values and Cost of Information

The value and cost of information play an essential role in estimating the importance of information as well as the total expenditure for buying the information.

Value of Information

Human beings have evolved in a way that they can appreciate the role of information in their life without much effort. The initial phase in human civilization has taught us to appreciate instinctively the importance of information and communication. We are instinctively alert to information. However, the human brain has evolved to understand that information has different degrees of 'value' (which the brain unconsciously rates). Information can be defined as processed data, which helps in decision-making and/or facilitates communication within an organization. More often, information provides answers to 'who', 'what', 'where' and 'when' type of questions. The human brain prioritizes information, according to its perceived value (most often, this unconscious valuation mechanism in our brain is correct, more so in the case of instinct-based information).

For example, let us assume that a driver notices a child suddenly crossing the road and calculates that he will hit the child unless he stops, and at the same time, he feels an itching sensation on his forehead. In this case, the brain of the driver prioritizes two different information received from different sensory inputs. It reacts by sending a signal to the driver's right foot to press the brake pedal to stop the car and only after the car stops will the brain react to the itch. We unconsciously do this every day. Evolution has taught us that information has a context and, hence, different degrees of value.

The principal objective of research is to find solutions to problems systematically. In general, the objectives of value of information with respect to research can be specified as follows:

- To extend the knowledge of human beings, environment and natural phenomenon to others
- To bring the information which is not developed fully during ordinary course of life
- To verify existing facts and identify the changes into these existing facts
- To develop facts for critical evaluation
- To analyse interrelationships between variables and deriving casual explanations
- To develop new tools and techniques that study unknown phenomenon
- To help in planning and development
- To acquire familiarity with a phenomenon
- To study the frequency of connection or independence of any activity or occurrence
- To determine the characteristics of an individual or a group of activities and the frequency of occurrence of these activities
- To test a hypothesis about a casual relationship that exists between variables

The value of information is determined based on the benefits that are derived from the information. Consider an example where two products A and B are developed. The benefits derived from product A evaluates to 20 and the benefits derived from product B evaluates to 30. The difference between the benefits of the two products is 10 units.

If you add some information, the benefits derived from product A increases by 20 points from 20 to 40. The actual value of information needs to be calculated from simple mathematics. The cost of information increases by 20 units. You need to subtract the cost involved in obtaining the information to determine the actual value of the information.

NOTES

Cost of Information

The cost of information determines the cost involved in obtaining the information, which includes:

- Cost of acquiring the data
- Cost of maintaining the data
- Cost of generating the information
- Cost of communicating the information

The cost is estimated from the point the information is generated, to the point the information is retrieved. The cost of obtaining accurate and complete information is more as compared to the cost generally retrieved from the system.

4.3 RELIABILITY AND VALIDITY

The required data for management research can be classified into two categories: primary data and secondary data. Secondary data is the data collected by others for their own use and which can be collected from the original source for another research. It needs to be searched and obtained from many reliable sources, which is increasingly becoming a specialized and skilled task in the present context of information explosion and the advent of complex computer search systems. In general, acquiring secondary data is less expensive and less time consuming than collecting primary data, and this is the biggest advantage with secondary data. But often, accuracy, availability, suitability and reliability are some of the major concerns for secondary data.

Primary data is the data specially collected in a research by the researcher. These are products of experiments, surveys, interviews or observations conducted in the research. Primary data is generated and collected through specific tools of data collection, like questionnaires, by the researcher.

4.3.1 Levels of Measurement and Questions of Validity and Reliabilty

In the field of social sciences, 'scaling' is applied to the procedures that attempt to determine quantitative measures of subjective, abstract concepts. The term 'scaling' is defined as a procedure for the assignment of numbers (or other symbols) to a property of objects in order to impart some the characteristics of numbers to the properties in question. In other words, if we have to measure some event, object, property, activity, characteristics or behaviour quantitatively, then any one of the following scales can be used.

1. Nominal Scale

Nominal scales are least restrictive and are widely used in social sciences and business research. This is the lowest level of quantitative measurement which is used for classification of objects, events and individuals into categories. Each category with its given name is assigned a number and the numbers are used only as labels without any relation, like order distance, or origin between the numbered categories. This classification scheme is referred to as a nominal scale.

Check Your Progress

1. Define research problem.
2. List the steps involved in the research problem formulation.

2. Ordinal Scales

As the name suggest, this scale possesses the characteristics of ordering which defines the relative position of objects or individuals according to some single attribute or property (for example if x > y and y > z, then x > z). This scale provides the task of ordering or ranking. While using this scale, investigation is limited to determination of 'greater than', 'equal to' or 'less than' without being able to explain how much greater or lesser (the difference).

3. Interval Scale

This measurement scale possesses the characteristics of the nominal and ordinal scales, and in addition, the units of measure, i.e., intervals between successive positions, are equal. It has a constant unit of measurement but an arbitrary zero because of which this can be changed from one to another by linear transformation only.

4. Ratio Scale

This scale possesses all the characteristics of the number system, and determines equality, rank-order, equality of intervals and equality of ratios because it has an absolute or true zero. This scales is very commonly used in the physical sciences than in the social sciences. Measures of weight, length, area, velocity, and so on, all conform to ratio scales. In the social sciences, the properties of concern that can be ratio scaled are money, age, years of education, and so on. However, successful ratio scaling of behavioural attributes is rare. All types of statistical analyses can be used on ratio scale measurements.

5. Attitude Measurement Scales

Attitude is a psychological construct, a way of conceptualizing the intangible. It is usually described as a mental state that is used by individuals to structure the way they perceive the environment around them and guides the way they respond to their environment. It is very difficult to observe or measure attitude directly because its existence is inferred from their consequences. People's values and beliefs may affect or dictate their attitude, and values and beliefs in retrospect are influenced by a person's attitude. For measurement of attitude, there are many ways a continuum of numbered categories can be presented before respondents. Figure 4.2 below is a diagrammatic representation of the various attitudinal scales in business research. You will learn more about attitude measurement in the subsequent section.

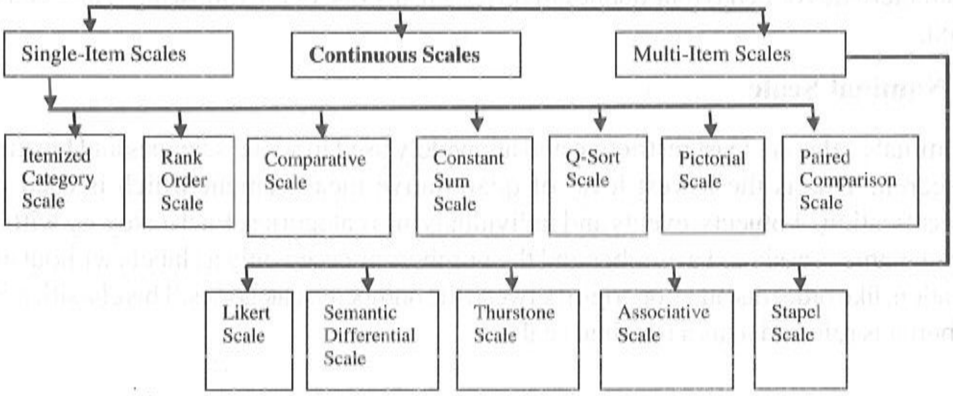


Fig. 4.2 Attitude Scales in Business Branch

4.3.2 Validity in Measurement

It is very obvious that in practice, errors will creep into measurements, making it necessary to evaluate the accuracy and dependability of the measuring instrument. For such an evaluation, the criteria are validity, practicality and reliability. The extent to which instrument a test measures what we have actually wished to measure is referred to as *validity*. *Reliability* is connected with the precision and accuracy of the procedure for measurement, while *practicality* is concerned with a wide range of factors, such as: convenience, economy and interpretability.

Though the literature provides many forms of validitation, the two major forms that need to be discussed are external and internal validity. The external validity of research findings is the data's ability to be generalized across persons, settings and times. The widely accepted classification of internal validity consists of three major forms: content, criterion-related and construct.

Content validity of a measuring instrument is the extent to which it provides adequate coverage of the topic under study. Content validity has been defined as the representativeness of the content of a measuring instrument. The content validity is good if the instrument contains a representative sample of the universe of the subject matter of interest. To evaluate the content validity of an instrument, one must first agree on what elements constitute adequate coverage of the problem and then determine what forms of these opinions constitute relevant positions on these topics. If the questionnaire adequately covers the topics that have been defined as the relevant dimensions, it is possible to conclude that the instrument has good content validity.

Criterion-related validity reflects the success of measures used for some empirical estimating purpose. One may want to predict some outcome or estimate the existence of some current behaviour or condition. These cases involve predictive and concurrent validity, respectively. They differ only in a time perspective. An opinionaire or opinion questionnaire that correctly forecasts the outcome of a union election has predictive validity. An observational method that correctly categorizes families by current income class has concurrent validity. While these examples appear to have rather simple and unambiguous validity criteria, there are difficulties in estimating validity.

Construct validity justifies how well the results obtained from the use of the measure fits the theory around which the test is designed. It is concerned with knowing more than just that a measuring instrument works. It is involved with the factors that lie behind the measurement scores obtained; with what factors or characteristics (that is, constructs) account for or explains the variance in measurement scores. One may also wish to measure or infer the presence of abstract characteristics for which no empirical validation seems possible. A[illegible] scales, and aptitude and personality tests are generally concerned concepts that fall [illegible] category.

Construct validity is as[illegible] ough convergent validity and discriminant validity. When the scores obtained by [illegible] ent instruments measuring the same concept are highly correlated, the conve[illegible] lity is established. Based on theory, when two variables are predicted to be [illegible] d and the scores obtained by measuring them are indeed empirically found t[illegible] discriminant validity is established.

4.3.3 Reliability in Mea[illegible]

The reliability of a measure is i[illegible] e consistency and stability with which the instrument measures the conce[illegible] n assessing the 'goodness' of a measure.

NOTES

A measure is reliable to the degree that it supplies consistent results. Reliabilit partial contributor to validity. A reliable instrument need not be valid, but a valid instrume is reliable. Reliability is not as valuable as validity, but is easier to assess. Reliable instruments can at least be used with confidence as a number of transient and situational factors are not interfering. They are robust instruments in that they work well under different conditions and at different times. This distinction of time and condition is the basis for identification of two aspects of reliability: stability and equivalence:

Stability: The ability of a measure to maintain stability over time, despite uncontrollable testing conditions and the state of the respondents themselves, indicates its stability and low vulnerability to changes in the situation. With a stable measure, we can secure consistent results with repeated measurements of the same person with the same instrument. It is often a simple matter to secure a number of repeat readings in observational studies but not so with questionnaires. Two tests of stability are *test-retest reliability*, which is the reliability coefficient obtained by the repetition of an identical measure on a second occasion, and *parallel-form reliability*, which is when responses on two comparable sets of measures tapping the same construct are highly correlated.

Equivalence: A second aspect of reliability considers how much error may be introduced by different investigators or different samples of the items being studied. Thus, while stability is concerned with personal and situational fluctuations from one time to another, equivalence is concerned with variations at one point in time among investigators and sample of items or with the internal consistency. A good way to test for the equivalence of measurements by two investigators is to compare their observations of the same events.

Practicality: The scientific requirements of a project call for the measurement process to be reliable and valid, while the operational requirements call for it to be practical. Well-known research thinkers Robert L. Thorndke and Elizabeth P. Hagen define practicality in terms of economy, convenience, adinterpretability.

4.4 MEASURES OF CENTRAL TEDENCY

There are several commonly used measures of central tency, such as arithmetic mean, mode and median. These values are very useful not in presenting the overall picture of the entire data but also for the purpose of makinnparisons among two or more sets of data.

As an example, questions like 'How hot is the mf June in Delhi?' can be answered generally by a single figure of the average for onth. Similarly, suppose we want to find out if boys and girls at age 10 years dieight for the purpose of making comparisons. Then, by taking the average heigs of that age and average height of girls of the same age, we can compare and e differences.

While arithmetic mean is the most commonly asure of central location, mode and median are more suitable measures und set of conditions and for certain types of data. However, each measure of ndency should meet the following requisites:

1. It should be easy to calculate and und nly one interpretation so
2. It should be rigidly defined. It sho tigator does not affect its that the personal prejudice or bias usefulness.

Check Your Progress

3. What do you mean by primary data?
4. How is the reliability of a measure indicated?

3. It should be representative of the data. If it is calculated from a sample, then the sample should be random enough to be accurately representing the population.
4. It should have sampling stability. It should not be affected by sampling fluctuations. This means that if we pick 10 different groups of college students at random and compute the average of each group, then we should expect to get approximately the same value from each of these groups.
5. It should not be affected much by extreme values. If few very small or very large items are present in the data, they will unduly influence the value of the average by shifting it to one side or other, so that the average would not be really typical of the entire series. Hence, the average chosen should be such that it is not unduly affected by such extreme values.

Let us consider the three measures of central tendency:

(a) **Arithmetic Mean:** This is also commonly known as simply the mean. Even though average, in general, means any measure of central location, when we use the word average in our daily routine, we always mean the arithmetic average. The term is widely used by almost every one in daily communication. We speak of an individual being an average student or of average intelligence. We always talk about average family size or average family income or Grade Point Average (GPA) for students, and so on.

Combined Mean: If the arithmetic averages and the number of items in two or more related groups are known, the combined (or composite) mean of the entire group can be obtained by the following formula:

$$\overline{\overline{X}} = \left[\frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}\right]$$

The advantage of combined arithmetic mean is that, one can determine the overall mean of the combined data without going back to the original data.

Example 4.1: Find the combined mean for the data given below:

$n_1 = 10, x_1 = 2, n_2 = 15, x_2 = 3$

Solution: The combined mean is obtained as follows:

$$\overline{\overline{X}} = \left[\frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}\right]$$

$$= \left[\frac{10\times 2 + 15\times 3}{10+15}\right]$$

$$= \frac{20+45}{25}$$

$$= 2.6$$

For discussion purposes, let us assume a variable X which stands for some score, such as the ages of students. Let the ages of 5 students be 19, 20, 22, 22 and 17 years, respectively. Then variable X would represent these ages as follows:

X: 19, 20, 22, 22, 17

NOTES

Placing the Greek symbol Σ(Sigma) before X would indicate a command that all values of X are to be added together. Thus:

$$\Sigma X = 19 + 20 + 22 + 22 + 17$$

The mean is computed by adding all the data values and dividing it by the number of such values. The symbol used for sample average is $\overline{X}$ so that:

$$\overline{X} = \frac{19 + 20 + 22 + 22 + 17}{5}$$

In general, if there are n values in the sample, then:

$$\overline{X} = \frac{X_1 + X_2 + \ldots\ldots\ldots + X_n}{n}$$

In other words,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \qquad i = 1,\ 2 \ldots n.$$

The above formula states to, add up all the values of X_i where the value of i starts at 1 and ends at n with unit increments so that $i = 1, 2, 3, \ldots n$.

If instead of taking a sample, we take the entire population in our calculations of the mean, then the symbol for the mean of the population is μ (mu) and the size of the population is N, so that:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}, \qquad i = 1,\ 2 \ldots N.$$

If we have the data in grouped discrete form with frequencies, then the sample mean is given by:

$$\overline{X} = \frac{\Sigma f(X)}{\Sigma f}$$

Where Σf = Summation of all frequencies
$\Sigma f(X)$ = Summation of each value of X multiplied by its corresponding frequency (f)

Example 4.2: The ages of 10 students are as follows:

19, 20, 22, 22, 17, 22, 20, 23, 17, 18

Calculate the sample average.

Solution: This data can be arranged in a frequency distribution as follows:

(X)	(f)	$f(X)$
17	2	34
18	1	18
19	1	19
20	2	40
22	3	66
23	1	23
	Total = 10	200

In the above case, we have $\Sigma f = 10$ and $\Sigma f(X) = 200$, so that:

$$\overline{X} = \frac{\Sigma f(X)}{\Sigma f}$$

$$= 200/10 = 20$$

Characteristics of the Mean

The arithmetic mean has three interesting properties. These are as follows:

1. The sum of the deviations of individual values of X from the mean will always add up to zero. This means that if we subtract all the individual values from their mean, then some values will be negative and some will be positive, but if all these differences are added together, then the total sum will be zero. In other words, the positive deviations must balance the negative deviations. Or symbolically:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0, i = 1, 2, \ldots n.$$

2. The second important characteristic of the mean is that it is very sensitive to extreme values. Since the computation of the mean is based upon inclusion of all values in the data, an extreme value in the data would shift the mean towards it, thus, making the mean unrepresentative of the data.
3. The third property of the mean is that the sum of squares of the deviations about the mean is minimum. This means that if we take differences between individual values and the mean, and square these differences individually and then add these squared differences, then the final figure will be less than the sum of the squared deviations around any other number other than the mean. Symbolically, it means that:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \text{Minimum}, i = 1, 2, \ldots n.$$

Weighted Arithmetic Mean

In the computation of arithmetic mean, we had given equal importance to each observation in the series. This equal importance may be misleading if the individual values constituting the series have different importance, as in the following example:

The Raja Toy shop sells

Toy Cars at	₹ 3 each
Toy Locomotives at	₹ 5 each
Toy Aeroplanes at	₹ 7 each
Toy Double Decker at	₹ 9 each

What shall be the average price of the toys sold, if the shop sells 4 toys, one of each kind?

Mean Price, i.e., $\bar{x} = \frac{\Sigma x}{4} = ₹\frac{24}{4} = ₹\,6$

In this case, the importance of each observation (Price quotation) is equal in as much as one toy of each variety has been sold. In the above computation of the arithmetic mean, this fact has been taken care of by including 'once only' the price of each toy.

But if the shop sells 100 toys: 50 cars, 25 locomotives, 15 aeroplanes and 10 double deckers, the importance of the four price quotations to the dealer is **not equal** as a

NOTES

source of earning revenue. In fact, their respective importance is equal to the number of units of each toy sold, i.e.,

The importance of Toy Car	50
The importance of Locomotive	25
The importance of Aeroplane	15
The importance of Double Decker	10

It may be noted that 50, 25, 15, 10 are the quantities of the various classes of toys sold. It is for these quantities that the term 'weights' is used in statistical language. Weight is represented by symbol 'w' and Σw represents the sum of weights.

While determining the 'average price of toy sold', these weights are of great importance and are taken into account in the manner illustrated below:

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4}{w_1 + w_2 + w_3 + w_4} = \frac{\sum wx}{\sum w}$$

When w_1, w_2, w_3, w_4 are the respective weights of x_1, x_2, x_3, x_4, which, in turn, represent the price of four varieties of toys, viz., car, locomotive, aeroplane and double decker, respectively.

$$\bar{x} = \frac{(50\times 3)+(25\times 5)+(15\times 7)+(10\times 9)}{50+25+15+10}$$

$$= \frac{(150)+(125)+(105)+(90)}{100} = \frac{470}{100} = ₹\,4.70$$

The table below summarizes the steps taken in the computation of the weighted arithmetic mean.

Weighted Arithmetic Mean of Toys Sold by the Raja Toy Shop

Toys	*Price per Toy* *₹x*	*Number Sold* *w*	*Price × Weight* *xw*
Car	3	50	150
Locomotive	5	25	125
Aeroplane	7	15	105
Double Decker	9	10	90
		$\Sigma w = 100$	$\Sigma xw = 470$

$$\Sigma w = 100; \quad \Sigma wx = 470$$

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{470}{100} = 4.70$$

The weighted arithmetic mean is particularly useful where we have to compute the *mean of means.* If we are given two arithmetic means, one for each of two different series, in respect of the *same variable*, and are required to find the arithmetic mean of the combined series, the weighted arithmetic mean is the only suitable method of its determination.

Example 4.3: The arithmetic mean of daily wages of two manufacturing concerns A Ltd. and B Ltd. is ₹ 5 and ₹ 7, respectively. Determine the average daily wages of both concerns if the number of workers employed were 2,000 and 4,000, respectively.

Solution: (*i*) Multiply each average (viz., 5 and 7) by the number of workers in the concern it represents.

(*ii*) Add up the two products obtained in Step (*i*) above.

(*iii*) Divide the total obtained in Step (*ii*) by the total number of workers.

Weighted Mean of Mean Wages of A Ltd. and B Ltd.

Manufacturing Concern	*Mean Wages* x	*Workers Employed* w	*Mean Wages × Workers Employed* wx .
A Ltd.	5	2,000	10,000
B Ltd.	7	4,000	28,000
		$\Sigma w = 6{,}000$	$\Sigma wx = 38{,}000$

$$\bar{x} = \frac{\Sigma wx}{\Sigma w}$$

$$= \frac{38{,}000}{6{,}000}$$

$$= ₹6.33$$

The above-mentioned examples explain that 'Arithmetic Means and Percentage' are not original data. They are derived figures and their importance is relative to the original data from which they are obtained. This relative importance must be taken into account by weighting while averaging them (means and percentage).

(b) Mode: The mode is another form of average and can be defined as the most frequently occurring value in the data. The mode is not affected by extreme values in the data and can easily be obtained from an ordered set of data. It can be useful and more representative of the data under certain conditions and is the only measure of central tendency that can be used for qualitative data. For instance, when a researcher quotes the opinion of an average person, he is probably referring to the most frequently expressed opinion which is the modal opinion. In our example of ages of 10 students as:

19, 20, 22, 22, 17, 22, 20, 23, 17 and 18

The mode is 22, since it occurs more often than any other value in this data.

(c) Median: The median is a measure of central tendency and it appears in the centre of an ordered data. It divides the list of ordered values in the data into two equal parts so that half of the data will have values less than the median and half will have values greater than the median.

If the total number of values is odd, then we simply take the middle value as the median. For instance, if there are 5 numbers arranged in order, such as 2, 3, 3, 5, 7, then 3 is the middle number and this will be the median. However, if the total number of values in the data is even, then we take the average of the middle two values. For instance, let there be 6 numbers in the ordered data, such as 2, 3, 3, 5, 7, 8, then the average of middle two numbers which are 3 and 5 would be the median, which is:

$$\text{Median} = \frac{(3+5)}{2} = 4$$

In general, the median is $\frac{n+1}{2}$th observation in the ordered data.

The median is a useful measure in the sense that it is not unduly affected by extreme values and is specially useful in open-ended frequencies.

NOTES

Check Your Progress

5. List the requisites that should be met by each measure of central tendency.
6. Define median.

NOTES

4.5 AVERAGE AND TYPES OF AVERAGES

Seasonal variation has been defined as predictable and repetitive movement around the trend line in a period of one year or less. For the measurement of seasonal variation, the time interval involved may be in terms of days, weeks, months or quarters. Because of the predictability of seasonal trends, we can plan in advance to meet these variations. For example, studying the seasonal variations in the production data makes it possible to plan for hiring of additional personnel for peak periods of production or to accumulate an inventory of raw materials or to allocate vacation time to personnel, and so on.

In order to isolate and identify seasonal variations, we first eliminate, as far as possible, the effects of trend, cyclical variations and irregular fluctuations on the time series. Some of the methods used for the measurement of seasonal variations are described as follows.

Simple Averages

This is the simplest method of isolating seasonal fluctuations in time series. It is based on the assumption that the series contain only the seasonal and irregular fluctuations. Assume that the time series involve monthly data over a time period of, say, 5 years. Assume further that we want to find the seasonal index for the month of March. (The seasonal variation will be the same for March in every year. Seasonal index describes the degree of seasonal variation.)

Then, the seasonal index for the month of March will be calculated as follows:

$$\text{Seasonal Index for March} = \left(\frac{\text{Monthly average for March}}{\text{Average of monthly averages}}\right) \times 10$$

The following steps can be used in the calculation of seasonal index (variation) for the month of March (or any month), over the five years period, regarding the sale of cars by one distributor:

1. Calculate the average sale of cars for the month of March over the last 5 years.

2. Calculate the average sale of cars for each month over the 5 years and then calculate the average of these monthly averages.

3. Use the formula to calculate seasonal index for March.

Let us say that the average sale of cars for the month of March over the period of 5 years is 360, and the average of all monthly average is 316. Then the seasonal index for March = (360/316) × 100 = 113.92.

Moving Averages

This is the most widely used method of measuring seasonal variations. The seasonal index is based upon a mean of 100 with the degree of seasonal variation (seasonal index) measured by variations away from this base value. For example, if we look at the seasonality of rental of row boats at the lake during the three summer months (a quarter) and we find that the seasonal index is 135, and we also know that the total boat rentals for the entire last year was 1680, then we can estimate the number of summer rentals for the row boats.

The average number of quarterly boats rented = 1680/4 = 420

The seasonal index 135 for the summer quarter means that the summer rentals are 135 per cent of the average quarterly rentals.

Hence, summer rentals = 420 × (135/100) = 567

The steps required to compute the seasonal index can be enumerated by illustrating an example.

NOTES

Example 4.4: Assume that a record of rental of row boats for the previous 3 years on a quarterly basis is given as follows:

Year	*Rentals per quarter*				*Total*
	I	II	III	IV	
1991	350	300	450	400	1500
1992	330	360	500	410	1600
1993	370	350	520	440	1680

Solution:

Step 1. The first step is to calculate the four-quarter moving total for time series. This total is associated with the middle data point in the set of values for the four quarters, which is shown as follows:

Year	*Quarters*	*Rentals*	*Moving Total*
1991	I	350	
	II	300	
			1500
	III	450	
	IV	400	

The moving total for the given values of four quarters is 1500 which is simply the addition of the four quarter values. This value of 1500 is placed in the middle of values 300 and 450, and recorded in the next column. For the next moving total of the four quarters, we will drop the value of the first quarter, which is 350, from the total and add the value of the fifth quarter (in other words, first quarter of the next year), and this total will be placed in the middle of the next two values, which are 450 and 400, and so on. These values of the moving totals are shown in column 4 of the following table.

Step 2. The next step is to calculate the quarter moving average. This can be done by dividing the four quarter moving total, as calculated in Step 1 earlier, by 4, since there are 4 quarters. The quarter moving average is recorded in column 5 in the table. The entire table of calculations is shown as follows:

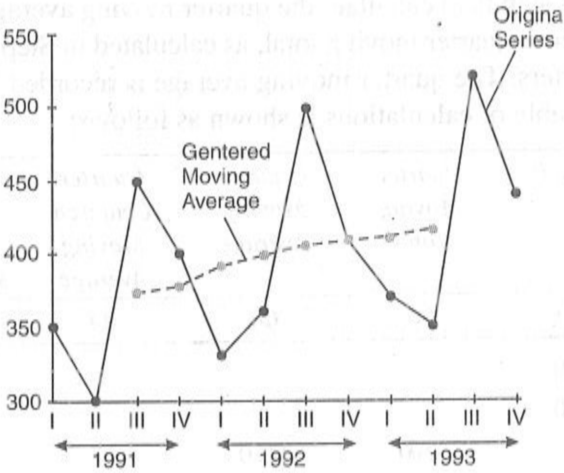
Year	*Quarters*	*Rentals*	*Quarter Moving Total*	*Quarter Moving Average*	*Quarter Centered Moving Average*	*Percentage of Actual to Centered Moving Average*
(1)	(2)	(3)	(4)	(5)	(6)	(7)
	I	350				
	II	300				
			1500	375.0		

	III	450			372.50	120.80
			1480	370.0		
	IV	400			377.50	105.96
			1540	385.0		
1992	I	330			391.25	84.35
			1590	397.5		
	II	360			398.75	90.28
			1600	400.0		
	III	500			405.00	123.45
			1640	410.0		
	IV	410			408,75	100.30
			1630	407.5		
1993	I	370			410.00	90.24
			1650	412.5		
	II	350			416.25	84.08
			1680	420.0		
	III	520				
	IV	440				

Step 3. After the moving averages for each of the consecutive four quarters have been taken, we centre these moving averages. As we see from the table, the quarterly moving average falls between the quarters. This is because the number of quarters is even which is 4. If we had odd number of time periods, such as 7 days of the week, then the moving average would already be centred and the third step here would not be necessary. Accordingly, we centre our averages in order to associate each average with the corresponding quarter, rather than between the quarters. This is shown in column 6, where the centred moving average is calculated as the average of the two consecutive moving averages.

The moving average (or the centred moving average) aims to eliminate seasonal and irregular fluctuations (S and I) from the original time series, so that this average represents the cyclical and trend components of the series.

As the following graph shows, the centred moving average has smoothed the peaks and troughs of the original time series.



Step 4. Column 7 in the table contains calculated entries which are percentages of the actual values to the corresponding centred moving average values. For example, the first four quarters centred moving average of 372.50 in the table has the corresponding actual value of 450, so that the percentage of actual value to centred moving average would be:

$$\frac{\text{Actual Value}}{\text{Centred Moving Average Value}}\times 100$$

$$=\frac{450}{372.5}\times 100$$

$$= 120.80$$

Step 5. The purpose of this step is to eliminate the remaining cyclical and irregular fluctuations still present in the values in Column 7 of the table. This can be done by calculating the 'modified mean' for each quarter. The modified mean for each quarter of the three years time period under consideration is calculated as follows:

(a) Make a table of values in column 7 of the previous table (percentage of actual to moving average values) for each quarter of the three years as shown in the following table:

Year	*Quarter I*	*Quarter II*	*Quarter (III)*	*Quarter (IV)*
1991	–	–	120.80	105.96
1992	84.35	90.28	123.45	100.30
1993	90.24	84.08	–	–

(b) We take the average of these values for each quarter. It should be noted that if there are many years and quarters taken into consideration instead of 3 years as we have taken, then the highest and lowest values from each quarterly data would be discarded and the average of the remaining data would be considered. By discarding the highest and lowest values from each quarter data, we tend to reduce the extreme cyclical and irregular fluctuations, which are further smoothed when we average the remaining values. Thus, the modified mean can be considered as an index of seasonal component. This modified mean for each quarter data is shown as follows:

$$\text{Quarter I} = \frac{84.35 + 90.24}{2} = 87.295$$

$$\text{Quarter II} = \frac{90.28 + 84.08}{2} = 87.180$$

$$\text{Quarter III} = \frac{120.80 + 123.45}{2} = 122.125$$

$$\text{Quarter IV} = \frac{105.96 + 100.30}{2} = 103.13$$

$$\text{Total} = 399.73$$

The modified means as calculated here are preliminary seasonal indices. These should average 100 per cent or a total of 400 for the 4 quarters. However, our total is 399.73. This can be corrected by the following step.

NOTES

Step 6. First, we calculate an adjustment factor. This is done by dividing the desired or the expected total of 400 by the actual total obtained of 399.73, so that:

$$\text{Adjustment} = \frac{400}{399.73} = 1.0007$$

By multiplying the modified mean for each quarter by the adjustment factor, we get the seasonal index for each quarter, so that:

$$\text{Quarter I} = 87.295 \times 1.0007 = 87.356$$
$$\text{Quarter II} = 87.180 \times 1.0007 = 87.241$$
$$\text{Quarter III} = 122.125 \times 1.0007 = 122.201$$
$$\text{Quarter IV} = 103.13 \times 1.0007 = 103.202$$
$$\text{Total} = 400.000$$

$$\text{Average seasonal index} = \frac{400}{4} = 100$$

(This average seasonal index is approximated to 100 because of rounding-off errors.)

The logical meaning behind this method is based on the fact that the centred moving average part of this process eliminates the influence of secular trend and cyclical fluctuations ($T \times C$). This may be represented by the following expression:

$$\frac{T \times S \times C \times I}{T \times C} = S \times I$$

where ($T \times S \times C \times I$) is the influence of trend, seasonal variations, cyclic fluctuations and irregular or chance variations.

Thus, the ratio to moving average represents the influence of seasonal and irregular components. However, if these ratios for each quarter over a period of years are averaged, then the most random or irregular fluctuations would be eliminated so that,

$$\frac{S \times I}{I} = S$$

and this would give us the value of seasonal influences.

4.6 MEASURES OF DISPERSION

A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.

A measure of dispersion may be expressed in an 'absolute form', or in a 'relative form'. It is said to be in an absolute form when it states the actual amount by which the value of an item on an average deviates from a measure of central tendency. Absolute measures are expressed in concrete units, i.e., units in terms of which the data have been expressed, e.g., rupees, centimetres, kilograms, and so on, and are used to describe frequency distribution.

A relative measure of dispersion computed is a quotient obtained by dividing the absolute measures by a quantity in respect to which absolute deviation has been computed.

Check Your Progress

7. Define simple averages.
8. State the aim of moving averages.

It is as such a pure number and is usually expressed in a percentage form. Relative measures are used for making comparisons between two or more distributions.

A measure of dispersion should possess all those characteristics which are considered essential for a measure of central tendency, which are as follows:

(i) It should be based on all observations.
(ii) It should be readily comprehensible.
(iii) It should be fairly easily calculated.
(iv) It should be affected as little as possible by fluctuations of sampling.
(v) It should be amenable to algebraic treatment.

Some common measures of dispersion are (i) the range, (ii) the semi-interquartile range or the quartile deviation, (iii) the mean deviation, and (iv) the standard deviation. Of these, the standard deviation is the best measure. We describe these measures in the following sections.

4.6.1 Quartile Deviation

There are many types of measures of dispersion. One of this is the semi-interquartile range, usually termed as 'quartile deviation' (Q.D.). Quartiles are the points which divide the array into four equal parts. More precisely, Q_1 gives the value of the item 1/4th the way up the distribution and Q_3 the value of the item 3/4th the way up the distribution. Between Q_1 and Q_3 are included half the total number of items. The difference between Q_1 and Q_3 includes only the central items but excludes the extremes. Since under most circumstances, the central half of the series tends to be fairly typical of all the items, the interquartile range ($Q_3 - Q_1$) affords a convenient and often a good indicator of the absolute variability. The larger the interquartile range, the larger the variability.

Usually, one-half of the difference between Q_3 and Q_1 is used and it is given the name of quartile deviation or semi-interquartile range. The interquartile range is divided by 2 for the reason that half of the interquartile range will, in a normal distribution, be equal to the difference between the median and any quartile. This means that 50 per cent items of a normal distribution will lie within the interval defined by the median plus and minus the semi-interquartile range.

Symbolically,

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} \qquad \text{...(4.1)}$$

Let us find quartile deviations for the weekly earnings of labour in the four workshops whose data is given in Table 4.1. The computations are as shown in Table 4.2.

Table 4.1 Weekly Earnings of Labourers in Four Workshops of the Same Type

Weekly earnings ₹	No. of workers			
	Workshop A	Workshop B	Workshop C	Workshop D
15–16	...	...	2	...
17–18	...	2	4	...
19–20	...	4	4	4
21–22	10	10	10	14
23–24	22	14	16	16

NOTES

25–26	20	18	14	16
27–28	14	16	12	12
29–30	14	10	6	12
31–32	...	6	6	4
33–34	...	...	2	2
35–36	...	...	...	...
37–38	...	...	4	...
Total	80	80	80	80
Mean	25.5	25.5	25.5	25.5

Workshop	*Range*
A	9
B	15
C	23
D	15

As shown in Table 4.2, Q.D. of workshop *A* is ₹ 2.12 and median value in 25.3. This means that if the distribution is symmetrical, the number of workers whose wages vary between (25.3–2.1) = ₹ 23.2 and (25.3 + 2.1) = ₹ 27.4, shall be just half of the total cases. The other half of the workers will be more than ₹ 2.1 removed from the median wage. As this distribution is not symmetrical, the distance between Q_1 and the median Q_2 is not the same as between Q_3 and the median. Hence, the interval defined by median plus and minus semi inter-quartile range v.ill not be exactly the same as given by the value of the two quartiles. Under such conditions, the range between ₹ 23.2 and ₹ 27.4 will not include precisely 50 per cent of the workers.

If quartile deviation is to be used for comparing the variability of any two series, it is necessary to convert the absolute measure to a coefficient of quartile deviation. To do this, the absolute measure is divided by the average size of the two quartiles.

Symbolically,

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1} \qquad ...(4.2)$$

Applying this to our illustration of four workshops in Table 4.1 the coefficients of Q.D. are as given in Table 4.2.

***Table 4.2** Calculation of Quartile Deviation*

		Workshop *A*	*Workshop* *B*	*Workshop* *C*	*Workshop* *D*
Location of Q_2	$\frac{N}{2}$	$\frac{80}{2}=40$	$\frac{80}{2}=40$	$\frac{80}{2}=40$	$\frac{80}{2}=40$
	Q_2	$24.5+\frac{40-30}{22}\times 2$ $= 24.5 + 0.9$ $= 25.4$	$24.5+\frac{40-30}{18}\times 2$ $= 24.5 + 1.1$ $= 25.61$	$24.5+\frac{40-30}{16}\times 2$ $= 24.5 + 0.75$ $= 25.25$	$24.5+\frac{40-30}{16}\times 2$ $= 24.5 + 0.75$ $= 25.25$
Location of Q_1	$\frac{N}{4}$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$

	Q_1	$22.5+\frac{20-10}{22}\times 2$ $= 22.5 + .91$ $= 23.41$	$22.5+\frac{20-16}{14}\times 2$ $= 22.5 + .57$ $= 23.07$	$20.5+\frac{20-10}{10}\times 2$ $= 20.5 + 2$ $= 22.5$	$22.5+\frac{20-18}{16}\times 2$ $= 22.5 + .25$ $= 22.75$
Location of Q_3	$\frac{3N}{4}$	$3\times\frac{80}{4}=60$	60	60	60
	Q_3	$26.5+\frac{60-52}{14}\times 2$ $= 26.5 + 1.14$ $= 27.64$	$26.5+\frac{60-48}{16}\times 2$ $= 26.5 + 1.5$ $= 28.0$	$26.5+\frac{60-50}{12}\times 2$ $= 26.5 + 1.67$ $= 28.17$	$26.5+\frac{60-50}{12}\times 2$ $= 26.5 + 1.67$ $= 28.17$
Quartile Deviation	$\frac{Q_3-Q_1}{2}$	$\frac{27.64-23.41}{2}$ $=\frac{4.23}{2}=$ ₹ 2.12	$\frac{28-23.07}{2}$ $=\frac{4.93}{2}=$ ₹ 2.46	$\frac{28.17-22.5}{2}$ $=\frac{5.67}{2}=$ ₹ 2.83	$\frac{28.17-22.75}{2}$ $=\frac{5.42}{2}=$ ₹. 2.71
Coefficient of Quartile Deviation $=\frac{27.64-23.41}{27.64+23.41}$	$\frac{Q_3-Q_1}{Q_3+Q_1}$	$= 0.083$	$\frac{28-23.07}{28+23.07}$ $= 0.097$	$\frac{28.17-22.5}{28.17+22.5}$ $= 0.112$	$\frac{28.17-22.75}{28.17+22.75}$ $= 0.106$

Characteristics of Quartile Deviation

The following are the characteristics of quartile deviation:

(i) The size of the quartile deviation gives an indication about the uniformity or otherwise of the size of the items of a distribution. If the quartile deviation is small, it denotes large uniformity. Thus, a coefficient of quartile deviation may be used for comparing uniformity or variation in different distributions.

(ii) Quartile deviation is not a measure of dispersion in the sense that it does not show the scatter around an average, but only a distance on scale. Consequently, quartile deviation is regarded as a measure of partition.

(iii) It can be computed when the distribution has open-end classes.

Limitation of Quartile Deviation

Except for the fact that its computation is simple and it is easy to understand, a quartile deviation does not satisfy any other test of a good measure of variation.

4.6.2 Mean Deviation

In this section, you will study that a weakness of the measures of dispersion, based upon the range or a portion thereof, is that the precise size of most of the variants has no effect on the result. As an illustration, the quartile deviation will be the same whether the variates between Q_1 and Q_3 are concentrated just above Q_1 or they are spread uniformly from Q_1 to Q_3. This is an important defect from the viewpoint of measuring the divergence of the distribution from its typical value. The mean deviation is employed to answer the objection.

Mean deviation, also called average deviation, of a frequency distribution is the mean of the absolute values of the deviation from some measure of central tendency. In other words, mean deviation is the arithmetic average of the variations (deviations) of the individual items of the series from a measure of their central tendency.

NOTES

We can measure the deviations from any measure of central tendency, but the most commonly employed ones are the median and the mean. The median is preferred because it has the important property that the average deviation from it is the least.

Calculation of mean deviation then involves the following steps:

(i) Calculate the median (or the mean) Me (or $\overline{X}$).

(ii) Record the deviations $|\,d\,| = |\,x - Me\,|$ of each of the items, ignoring the sign.

(iii) Find the average value of deviations.

$$\text{Mean Deviation} = \frac{\Sigma|d|}{N} \qquad \text{...(4.3)}$$

Example 4.5 explains it better.

Example 4.5: Calculate the mean deviation from the following data giving marks obtained by 11 students in a class test.

14, 15, 23, 20, 10, 30, 19, 18, 16, 25, 12

Solution: The mean deviation is obtained as follows:

$$\text{Median} = \text{Size of } \frac{11+1}{2} \text{ th item}$$

$$= \text{Size of 6th item} = 18.$$

Serial No.	*Marks*	*| x – Median |* *| d |*
1	10	8
2	12	6
		$\Sigma\,|d| = 50$

$$\text{Mean deviation from median} = \frac{\Sigma\,|d|}{N}$$

$$= \frac{50}{11} = 4.5 \text{ marks}$$

For grouped data, it is easy to see that the mean deviation is given by:

$$\text{Mean deviation} = \frac{\Sigma f\,|d|}{\Sigma f} \qquad \text{...(4.4)}$$

Where,

$|\,d\,| = |\,x - \text{median}\,|$ for grouped discrete data.

$|\,d\,| = |\,M - \text{median}\,|$ for grouped continuous data with M as the mid-value of a particular group.

Examples 4.6 and 4.7 illustrate the use of this formula.

Example 4.6: Calculate the mean deviation from the following data:

Size of item	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4

Solution: The mean deviation is calculated as follows:

Size	*Frequency* *f*	*Cumulative* *Frequency*	*Deviations* *From median* *(9)* $\lvert d \rvert$	$f\lvert d \rvert$
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12
	48			60

Median = The size of $\frac{48+1}{2}$ = 24.5th item which is 9.

Therefore, deviations d are calculated from 9, i.e., $\lvert d \rvert = \lvert x - 9 \rvert$.

$$\text{Mean deviation} = \frac{\Sigma f \lvert d \rvert}{\Sigma f} = \frac{60}{48} = 1.25$$

Example 4.7: Calculate the mean deviation from the following data:

x	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	18	16	15	12	10	5	2	2

Solution: This is a frequency distribution with continuous variable. Thus, deviations are calculated from mid-values.

x	*Mid-Value*	f	*Less than* *c.f.*	*Deviation* *from Median* $\lvert d \rvert$	$f\lvert d \rvert$
0–10	5	18	18	19	342
10–20	15	16	34	9	144
20–30	25	15	49	1	15
30–40	35	12	61	11	132
40–50	45	10	71	21	210
50–60	55	5	76	31	155
60–70	65	2	78	41	82
70–80	75	2	80	51	102
		80			1182

Median = The size of $\frac{80}{2}$ th item

$$= 20 + \frac{6}{15} \times 10 = 24$$

And then, mean deviation $= \frac{\Sigma f\,|d|}{\Sigma f}$

$$= \frac{1182}{80} = 14.775$$

Merits and Demerits of the Mean Deviation

Merits

The merits are as follows:

(i) It is easy to understand.

(ii) As compared to standard deviation (discussed later), its computation is simple.

(iii) As compared to standard deviation, it is less affected by extreme values.

(iv) Since it is based on all values in the distribution, it is better than range or quartile deviation.

Demerits

The demerits are as follows:

(i) It lacks those algebraic properties which would facilitate its computation and establish its relation to other measures.

(ii) Due to this, it is not suitable for further mathematical processing.

Coefficient of Mean Deviation

The coefficient or relative dispersion is found by dividing the mean deviations recorded. Thus,

$$\text{Coefficient of MD} = \frac{\text{Mean Deviation}}{\text{Mean}} \qquad \text{...(4.5)}$$

(when deviations were recorded from the mean)

$$= \frac{\text{MD}}{\text{Median}} \qquad \text{...(4.6)}$$

(when deviations were recorded from the median)

Applying the above formula to Example 4.6.

$$\text{Coefficient of MD} = \frac{14.775}{24}$$

$$= 0.616$$

4.6.3 Standard Deviation

By far, the most universally used and the most useful measure of dispersion is the standard deviation or root mean square deviation about the mean. We have seen that all the methods of measuring dispersion so far discussed are not universally adopted for adequacy and accuracy. The range is not satisfactory as its magnitude is determined by most extreme cases in the entire group. Further, the range is notable because it is dependent on the item whose size is largely a matter of chance. Mean deviation method is also an unsatisfactory measure of scatter, as it ignores the algebraic signs of deviation. We desire a measure of scatter which is free from these shortcomings. To some extent, standard deviation is one such measure.

The calculation of standard deviation differs in the following respects from that of mean deviation. First, in calculating standard deviation, the deviations are squared. This is done so as to get rid of negative signs without committing algebraic violence. Further, the squaring of deviations provides added weight to the extreme items, a desirable feature for certain types of series.

Second, the deviations are always recorded from the arithmetic mean, because although the sum of deviations is the minimum from the median, the sum of squares of deviations is minimum when deviations are measured from the arithmetic average. The deviation from $\bar{x}$ is represented by *d*.

Thus, standard deviation, σ (sigma) is defined as the square root of the mean of the squares of the deviations of individual items from their arithmetic mean.

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}} \qquad \text{...(4.7)}$$

For grouped data (discrete variables),

$$\sigma = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{\Sigma f}} \qquad \text{...(4.8)}$$

And, for grouped data (continuous variables),

$$\sigma = \sqrt{\frac{\Sigma f(M-\bar{x})}{\Sigma f}} \qquad \text{...(4.9)}$$

Where *M* is the mid-value of the group.

The use of these formulae is illustrated in Examples 4.8 and 4.9.

Example 4.8: Compute the standard deviation for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution: Here, Formula (4.7) is appropriate. We first calculate the mean as $\bar{x} = \Sigma x/N = 176/11 = 16$, and then calculate the deviation as follows:

x	$(x-\bar{x})$	$(x-\bar{x})^2$
11	–5	25
12	–4	16
13	–3	9
14	–2	4
15	–1	1
16	0	0
17	+1	1
18	+2	4
19	+3	9
20	+4	16
21	+5	25
176		110

Thus, by Formula (4.7),

$$\sigma = \sqrt{\frac{110}{11}} = \sqrt{10} = 3.16$$

Example 4.9: Find the standard deviation of the data in the following distributions:

x	12	13	14	15	16	17	18	20
f	4	11	32	21	15	8	6	4

Solution: For this discrete variable grouped data, we use Formula (4.8). Since for calculation of $\bar{x}$, we need $\sum fx$ and then for σ we need $\sum f(x-\bar{x})^2$, the calculations are conveniently made in the following format.

x	f	fx	$d = x - \bar{x}$	d^2	fd^2
12	4	48	–3	9	36
13	11	143	–2	4	44
14	32	448	–1	1	32
15	21	315	0	0	0
16	15	240	1	1	15
17	8	136	2	4	32
18	5	90	3	9	45
20	4	80	5	25	100
	100	1500			304

Here, $\bar{x} = \sum fx / \sum f = 1500/100 = 15$

And, $$\sigma = \sqrt{\frac{\sum fd^2}{\sum f}}$$

$$= \sqrt{\frac{304}{100}} = \sqrt{3.04} = 1.74$$

Calculation of Standard Deviation by Short-Cut Method

In most cases, it is very unlikely that $\bar{x}$ will turn out to be an integer simplifying problems. In such cases, the calculation of d and d^2 becomes quite time-consuming. Short-cut methods have consequently been developed. These are on the same lines as those for calculation of mean itself.

In the short-cut method, we calculate deviations x' from an assumed mean A. Then for ungrouped data,

$$\sigma = \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2} \qquad ...(4.10)$$

And for grouped data:

$$\sigma = \sqrt{\frac{\sum fx'^2}{\sum f} - \left(\frac{fx'}{\sum f}\right)^2} \qquad ...(4.11)$$

This formula is valid for both discrete and continuous variables. In case of continuous variables, x in the equation $x' = x - A$ stands for the mid-value of the class in question.

Note that the second term in each of the formulae is a correction term because of the difference in the values of A and $\bar{x}$. When A is taken as $\bar{x}$ itself, this correction is automatically reduced to zero. Examples 4.11 to 4.12 explain the use of these formulae.

NOTES

Example 4.10: Compute the standard deviation by the short-cut method for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution: Let us assume that $A = 15$.

	$x' = (x - 15)$	x'^2
11	−4	16
12	−3	9
13	−2	4
14	−1	1
15	0	0
16	1	1
17	2	4
18	3	9
19	4	16
20	5	25
21	6	36
$N = 11$	$\sum x' = 11$	$\sum x'^2 = 121$

$$\sigma = \sqrt{\frac{\sum x'^2}{N} - \left(\frac{\sum x'}{N}\right)^2}$$

$$= \sqrt{\frac{121}{11} - \left(\frac{11}{11}\right)^2}$$

$$= \sqrt{11 - 1}$$

$$= \sqrt{10}$$

$$= 3.16$$

Another Method

If we assume A as zero, then the deviation of each item from the assumed mean is the same as the value of item itself. Thus, 11 deviates from the assumed mean of zero by 11, 12 deviates by 12, and so on. As such, we work with deviations without having to compute them and the formula takes the following shape:

x	x^2
11	121
12	144
13	169
14	196
15	225
16	256
17	289
18	324
19	361
20	400
21	441
176	2,926

NOTES

$$\sigma = \sqrt{\frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2}$$

$$= \sqrt{\frac{2926}{11} - \left(\frac{176}{11}\right)^2} = \sqrt{266-256} = 3.16$$

Combining Standard Deviations of Two Distributions

If we were given two sets of data of N_1 and N_2 items with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2, respectively, we can obtain the mean and standard deviation $\bar{x}$ and σ of the combined distribution by the following formulae:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} \quad ...(4.12)$$

And, $$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x}-\bar{x}_1)^2 + N_2(\bar{x}-\bar{x}_2)^2}{N_1+N_2}} \quad ...(4.13)$$

Example 4.11: The mean and standard deviations of two distributions of 100 and 150 items are 50, 5 and 40, 6, respectively. Find the standard deviation of all taken together.

Solution: The standard deviation is obtained as follows:

Combined mean,

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{100 \times 50 + 150 \times 40}{100 + 150}$$

$$= 44$$

Combined standard deviation,

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x}-\bar{x}_1)^2 + N_2(\bar{x}-\bar{x}_2)^2}{N_1+N_2}}$$

$$= \sqrt{\frac{100 \times (5)^2 + 150\,(6)^2 + 100\,(44-50)^2 + 150\,(44-40)^2}{100+150}}$$

$$= 7.46$$

Example 4.12: A distribution consists of three components with 200, 250, 300 items having mean 25, 10 and 15 and standard deviation 3, 4 and 5, respectively. Find the standard deviation of the combined distribution.

Solution: In the usual notations, we are given here:

$N_1 = 200, N_2 = 250, N_3 = 300$

$\bar{x}_1 = 25, \bar{x}_2 = 10, \bar{x}_3 = 15$

The Formulae (4.12) and (4.13) can easily be extended for combination of three series as:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

$$= \frac{200 \times 25 + 250 \times 10 + 300 \times 15}{200 + 250 + 300}$$

$$= \frac{12000}{750} = 16$$

And,

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1(\overline{x} - \overline{x}_1)^2 + N_2(\overline{x} - \overline{x}_2)^2 + N_3(\overline{x} - \overline{x}_3)^2}{N_1 + N_2 + N_3}}$$

$$= \sqrt{\frac{200 \times 9 + 250 \times 16 + 300 \times 25 + 200 \times 81 + 250 \times 36 + 300 \times 1}{200 + 250 + 300}}$$

$$= \sqrt{51.73} = 7.19$$

4.6.4 Range

The crudest measure of dispersion is the range of the distribution. The range of any series is the difference between the highest and the lowest values in the series. If the marks received in an examination taken by 248 students are arranged in ascending order, then the range will be equal to the difference between the highest and the lowest marks.

In a frequency distribution, the range is taken to be the difference between the lower limit of the class at the lower extreme of the distribution and the upper limit of the class at the upper extreme.

Consider the data on weekly earnings of worker on four workshops given in Table 4.1.

From these figure in Table 4.1, it is clear that the greater the range, the greater is the variation of the values in the group.

The range is a measure of absolute dispersion and as such cannot be usefully employed for comparing the variability of two distributions expressed in different units. The amount of dispersion measured, say, in pounds, is not comparable with dispersion measured in inches. Thus, the need of measuring relative dispersion arises.

An absolute measure can be converted into a relative measure if we divide it by some other value regarded as standard for the purpose. We may use the mean of the distribution or any other positional average as the standard.

For Table 4.1, the relative dispersion would be,

$$\text{Workshop } A = \frac{9}{25.5} \qquad \text{Workshop } C = \frac{23}{25.5}$$

$$\text{Workshop } B = \frac{15}{25.5} \qquad \text{Workshop } D = \frac{15}{25.5}$$

An alternate method of converting an absolute variation into a relative one would be to use the total of the extremes as the standard. This will be equal to dividing the difference of the extreme items by the total of the extreme items. Thus,

$$\text{Relative Dispersion} = \frac{\text{Difference of extreme items, i.e, Range}}{\text{Sum of extreme items}}$$

The relative dispersion of the series is called the coefficient or ratio of dispersion. In our example of weekly earnings of workers considered earlier, the coefficients would be,

$$\text{Workshop } A = \frac{9}{21+30} = \frac{9}{51} \qquad \text{Workshop } B = \frac{15}{17+32} = \frac{15}{49}$$

$$\text{Workshop } C = \frac{23}{15+38} = \frac{23}{53} \qquad \text{Workshop } D = \frac{15}{19+34} = \frac{15}{53}$$

Merits and Limitations of Range

Merits

Among the various characteristics that a good measure of dispersion should possess, the range has only two, which are as follows:

(i) It is easy to understand.

(ii) Its computation is simple.

Limitations

Besides the aforesaid two qualities, the range does not satisfy the other test of a good measure and, hence, it is often termed as a crude measure of dispersion.

The following are the limitations that are inherent in the range as a concept of variability:

(i) Since it is based upon two extreme cases in the entire distribution, the range may be considerably changed if either of the extreme cases happens to drop out, while the removal of any other case would not affect it at all.

(ii) It does not tell anything about the distribution of values in the series relative to a measure of central tendency.

(iii) It cannot be computed when distribution has open-end classes.

(iv) It does not take into account the entire data. These can be illustrated by the following illustration. Consider the data given in Table 4.3.

Table 4.3 *Distribution with the Same Number of Cases but Different Variability*

Class	*No. of students*		
	Section A	*Section B*	*Section C*
0–10	...	...	...
10–20	1	...	...
20–30	12	12	19
30–40	17	20	18
40–50	29	35	16
50–60	18	25	18
60–70	16	10	18
70–80	6	8	21
80–90	11	...	...
90–100	...	...	...
Total	110	110	110
Range	80	60	60

Table 4.3 is designed to illustrate three distributions with the same number of cases but different variability. The removal of two extreme students from Section *A* would make its range equal to that of *B* or *C*.

The greater range of *A* is not a description of the entire group of 110 students, but of the two most extreme students only. Further, though sections *B* and *C* have the same range, the students in Section *B* cluster more closely around the central tendency of the group than they do in Section *C*. Thus, the range fails to reveal the greater homogeneity of *B* or the greater dispersion of *C*. Due to this defect, it is seldom used as a measure of dispersion.

NOTES

Specific Uses of Range

In spite of the numerous limitations of the range as a measure of dispersion, it is the most appropriate under the following circumstances:

(i) In situations where the extremes involve some hazard for which preparation should be made, it may be more important to know the most extreme cases to be encountered than to know anything else about the distribution. For example, an explorer, would like to know the lowest and the highest temperatures on record in the region he is about to enter; or an engineer would like to know the maximum rainfall during 24 hours for the construction of a storm water drain.

(ii) In the study of prices of securities, range has a special field of activity. Thus, to highlight fluctuations in the prices of shares or bullion, it is a common practice to indicate the range over which the prices have moved during a certain period of time. This information, besides being of use to the operators, gives an indication of the stability of the bullion market, or that of the investment climate.

(iii) In statistical quality control, range is used as a measure of variation. We, for example, determine the range over which variations in quality are due to random causes, which is made the basis for the fixation of control limits.

4.7 CORRELATION ANALYSIS AND REGRESSION ANALYSIS

Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable. In fact, the word 'correlation' refers to the relationship or interdependence between two variables. There are various phenomenons which have relation to each other. For instance, when demand of a certain commodity increases, then its price goes up and when its demand decreases, its price comes down. Similarly, with age, the height of the children; with height, the weight of the children; and with money, the supply and the general level of prices, go up. Such sort of relationship can as well be noticed for several other phenomena. The theory by means of which quantitative connections between two sets of phenomena are determined is called the *Theory of Correlation.*

On the basis of the theory of correlation, one can study the comparative changes occurring in two related phenomena and their cause-effect relation can be examined. It should, however, be borne in mind that ideas like 'black cat causes bad luck', 'filled up pitchers result in good fortune' and similar other beliefs of the people cannot be explained by the theory of correlation, since they are all imaginary and are incapable of being justified mathematically. Thus, correlation is concerned with relationship between two related and quantifiable variables. If two quantities vary in sympathy, so that a movement

Check Your Progress

9. Define measure of dispersion.
10. List the merits of mean deviation.
11. State the crudest measure of dispersion.

(an increase or decrease) in one tends to be accompanied by a movement in the same or opposite direction in the other and the greater the change in the one, the greater is the change in the other, the quantities are said to be correlated. This type of relationship is known as correlation or what is sometimes called, in statistics, as covariation.

For correlation, it is essential that the two phenomena should have cause-effect relationship. If such relationship does not exist, then one should not talk of correlation. For example, if the height of the students as well as the height of the trees increases, then one should not call it a case of correlation because the two phenomena, viz., the height of students and the height of trees are not even casually related. However, the relationship between the price of a commodity and its demand, the price of a commodity and its supply, the rate of interest and savings, and so on, are examples of correlation, since in all such cases the change in one phenomenon is explained by a change in the other phenomenon.

It is appropriate here to mention that correlation in case of phenomena pertaining to natural sciences can be reduced to absolute mathematical term, e.g., heat always increases with light. However, in phenomena pertaining to social sciences, it is often difficult to establish any absolute relationship between two phenomena. Hence, in social sciences, we must take the fact of correlation being established if in a large number of cases, two variables always tend to move in the same or opposite direction.

Correlation can either be positive or it can be negative. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive, but when the variations in the two variables take place in opposite direction, the correlation is termed as negative (see Table 4.4).

Table 4.4 *Nature of Correlation*

Changes in Independent Variable	*Changes in Dependent Variable*	*Nature of Correlation*
Increase (+)↑	Increase (+)↑	Positive (+)
Decrease (–)↓	Decrease (–)↓	Positive (+)
Increase (+)↑	Decrease (–)↓	Negative (–)
Decrease (–)↓	Increase (+)↑	Negative (–)

Statisticians have developed *two measures for describing the correlation* between two variables, viz., the coefficient of determination and the coefficient of correlation. We now explain, illustrate and interpret the said two coefficients concerning the relationship between two variables as under.

4.7.1 The Coefficient of Determination

The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say X and Y, one of which happens to be independent variable and the other being dependent variable. This coefficient is based on the following two kinds of variations:

(i) The variation of the Y values around the fitted regression line viz., $\Sigma\left(Y-\hat{Y}\right)^2$, technically known as the unexplained variation.

(ii) The variation of the Y values around their own mean viz., $\Sigma\left(Y-\bar{Y}\right)^2$, technically known as the total variation.

If we subtract the unexplained variation from the total variation, we obtain what is known as the explained variation, i.e., the variation explained by the line of regression. Thus, Explained Variation = (Total Variation) – (Unexplained Variation).

$$= \Sigma\left(Y-\bar{Y}\right)^2 - \Sigma\left(Y-\hat{Y}\right)^2$$

$$= \Sigma\left(\hat{Y}-\bar{Y}\right)^2$$

The Total and Explained as well as Unexplained variations can be shown as given in Figure 4.3.

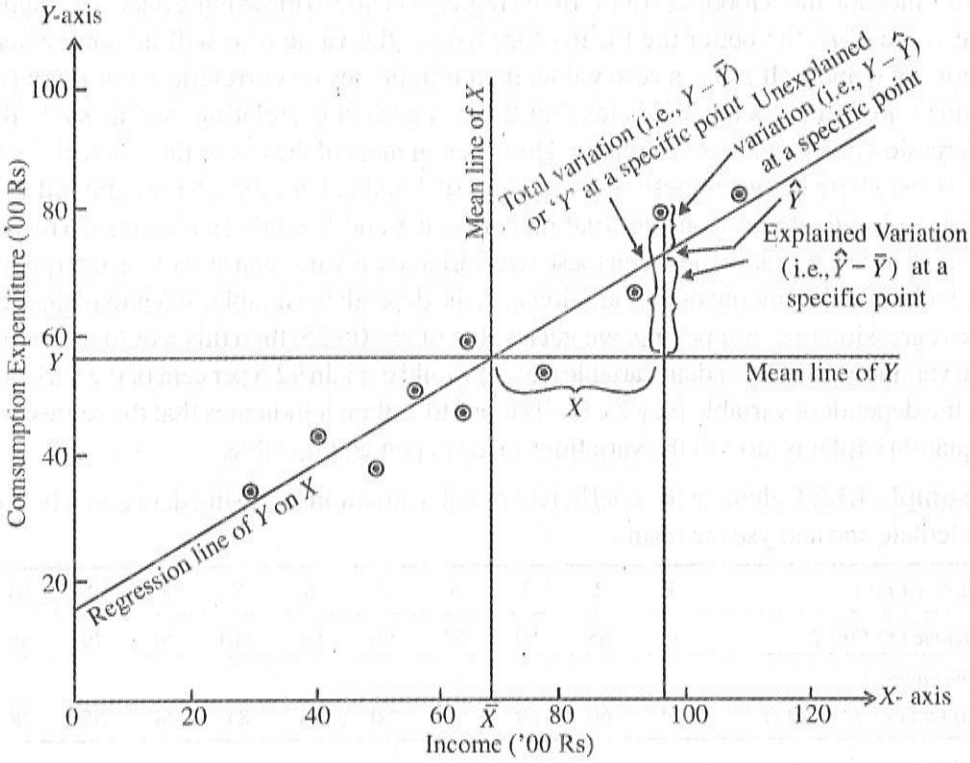

Y-axis
100
80
60
$\bar{Y}$
40
20
0
Consumption Expenditure ('00 Rs)
Mean line of X
Total variation (i.e., $Y-\bar{Y}$) or 'Y' at a specific point
Unexplained variation (i.e., $Y-\hat{Y}$) at a specific point
$\hat{Y}$
Explained Variation (i.e., $\hat{Y}-\bar{Y}$) at a specific point
Mean line of Y
X
Regression line of Y on X
X- axis
20
40
60
$\bar{X}$
80
100
120
Income ('00 Rs)

Fig. 4.3 Diagram Showing Total, Explained and Unexplained Variations

Coefficients of determination is that fraction of the total variation of Y which is explained by the regression line. In other words, coefficient of determination is the ratio of explained variation to total variation in the Y variable related to the X variable. Coefficient of determination algebraically can be stated as under:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{\Sigma\left(\hat{Y}-\bar{Y}\right)^2}{\Sigma\left(Y-\bar{Y}\right)^2}$$

Alternatively, r^2 can also be stated as under:

$$r^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\Sigma\left(\hat{Y}-\overline{Y}\right)^2}{\Sigma\left(Y-\overline{Y}\right)^2}$$

4.7.2 Interpreting r^2

The coefficient of determination can have a value ranging from zero to one. The value of one can occur only if the unexplained variation is zero, which simply means that all the data points in the Scatter diagram fall exactly on the regression line. For a zero value to occur, $\Sigma(Y-\overline{Y})^2 = \Sigma(Y-\hat{Y})^2$, which simply means that *X* tells us nothing about *Y* and, hence, there is no regression relationship between *X* and *Y* variables. Values between 0 and 1 indicate the 'Goodness of fit' of the regression line to the sample data. The higher the value of r^2, the better the fit. In other words, the value of r^2 will lie somewhere between 0 and 1. If r^2 has a zero value, then it indicates no correlation, but if it has a value equal to 1, then it indicates that there is perfect correlation, and as such, the regression line is a perfect estimator. However, in most of the cases, the value of r^2 will lie somewhere between these two extremes of 1 and 0. One should remember that r^2 close to 1 indicates a strong correlation between *X* and *Y*, while an r^2 near zero means there is little correlation between these two variables. r^2 value can as well be interpreted by looking at the amount of the variation in *Y*, the dependant variable, that is explained by the regression line. Supposing, we get a value of $r^2 = 0.925$, then this would mean that the variations in independent variable (say *X*) would explain 92.5 per cent of the variation in the dependent variable (say *Y*). If r^2 is close to 1, then it indicates that the regression equation explains most of the variations in the dependent variable.

Example 4.13: Calculate the coefficient of determination (r^2) using data given below. Calculate and analyse the result.

Observations	1	2	3	4	5	6	7	8	9	10
Income (X) ('00 ₹)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (Y) ('00 ₹)	44	60	39	51	80	68	84	34	55	48

Solution: r^2 can be worked out as shown below:

Since,
$$r^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}} = 1 - \frac{\Sigma\left(Y-\hat{Y}\right)^2}{\Sigma\left(Y-\overline{Y}\right)^2}$$

As, $\Sigma\left(Y-\overline{Y}\right)^2 = \Sigma Y^2 = \left(\Sigma Y^2 - n\overline{Y}^2\right)$, we can write,

$$r^2 = 1 - \frac{\Sigma\left(Y-\hat{Y}\right)^2}{\Sigma Y^2 - n\overline{Y}^2}$$

Calculating and putting the various values, we have the following equation:

$$r^2 = 1 - \frac{260.54}{34223 - 10(56.3)^2} = 1 - \frac{260.54}{2526.10} = 0.897$$

Analysis of the Result: The regression equation used to calculate the value of the coefficient of determination (r^2) from the sample data shows that about 90 per cent of the variations in consumption expenditure can be explained. In other words, it means that the variations in income explain about 90 per cent of variations in consumption expenditure.

Observation	1	2	3	4	5	6	7	8	9	10
Income (X) ('00 ₹)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (Y) ('00 ₹)	44	60	39	51	80	68	84	34	55	48

4.7.3 Regression Equations

The term 'regression' was first used in 1877 by English Victorian Statistician Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move back or 'regress' toward the mean height of the population. He designated the word 'regression' as the name of the process of predicting one variable from the another variable. He coined the term 'multiple regression' to describe the process by which several variables are used to predict another. Thus, when there is a well-established relationship between variables, it is possible to make use of this relationship in making estimates and to forecast the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s). A banker, for example, could predict deposits on the basis of per capita income in the trading area of bank. A marketing manager may plan his advertising expenditures on the basis of the expected effect on total sales revenue of a change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population. Such predictions may be made by using regression analysis. An investigator may employ regression analysis to test his theory having the cause and effect relationship. All this explains that regression analysis is an extremely useful tool, specially in problems of business and industry involving predictions.

Assumptions in Regression Analysis

While making use of the regression techniques for making predictions, it is always assumed that:

(a) There is an actual relationship between the dependent and independent variables.

(b) The values of the dependent variable are random but the values of the independent variable are fixed quantities without error and are chosen by the experimentor.

(c) There is clear indication of direction of the relationship. This means that dependent variable is a function of independent variable. (For example, when we say that advertising has an effect on sales, then we are saying that sales has an effect on advertising.)

(d) The conditions (that existed when the relationship between the dependent and independent variable was estimated by the regression) are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed.

(e) The analysis is to be used to predict values within the range (and not for values outside the range) for which it is valid.

4.7.4 Simple Linear Regression Model

In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (i.e., relationship of the type defined by $Y = a + bX$) between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

NOTES

Simple linear regression model (or the Regression Line) is stated as,

$$Y_i = a + bX_i + e_i$$

Where, Y_i is the dependent variable.

X_i is the independent variable.

e_i is unpredictable random element (usually called as residual or error term).

(a) a represent the Y-intercept, i.e., the intercept specifies the value of the dependent variable when the independent variable has a value of zero. (But this term has practical meaning only if a zero value for the independent variable is possible.)

(b) b is a constant, indicating the slope of the regression line. Slope of the line indicates the amount of change in the value of the dependent variable for a unit change in the independent variable.

If the two constants (viz., a and b) are known, the accuracy of our prediction of Y (denoted by $\hat{Y}$ and read as Y-hat) depends on the magnitude of the values of e_i. If in the model, all the e_i tend to have very large values, then the estimates will not be very good but if these values are relatively small, then the predicted values ($\hat{Y}$) will tend to be close to the true values (Y_i).

Estimating the Intercept and Slope of the Regression Model (or Estimating the Regression Equation)

The two constants or the parameters, viz., 'a' and 'b' in the regression model for the entire population or universe are generally unknown and as such are estimated from sample information. The following are the two methods used for estimation:

(a) Scatter diagram method

(b) Least squares method

Scatter Diagram Method

This method makes use of the Scatter diagram also known as Dot diagram. Scatter diagram is a diagram representing two series with the known variable, i.e., independent variable plotted on the X-axis and the variable to be estimated, i.e., dependent variable to be plotted on the Y-axis on a graph paper (see Figure 4.4) to get the following information:

Income X *(Hundreds of Rupees)*	*Consumption Expenditure* Y *(Hundreds of Rupees)*
41	44
65	60
50	39
57	51
96	80
94	68
110	84
30	34
79	55
65	48

The scatter diagram by itself is not sufficient for predicting values of the dependent variable. Some formal expression of the relationship between the two variables is necessary for predictive purposes. For the purpose, one may simply take a ruler and draw a straight line through the points in the scatter diagram, and this way can determine the intercept and the slope of the said line, and then the line can be defined as $\hat{Y} = a + bX_i$, with the help of which we can predict Y for a given value of X. However, there are shortcomings in this approach. For example, if five different persons draw such a straight line in the same scatter diagram, it is possible that there may be five different estimates of a and b, specially when the dots are more dispersed in the diagram. Hence, the estimates cannot be worked out only through this approach. A more systematic and statistical method is required to estimate the constants of the predictive equation. The least squares method is used to draw the best fit line.

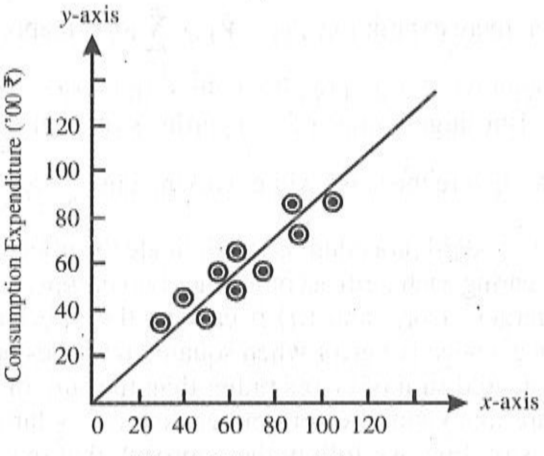


Fig. 4.4 Scatter Diagram

Least Square Method

Least squares method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.

The meaning of the least squares criterion can be easily understood through reference to Figure 4.5 drawn below, where Figure 4.4 in scatter diagram has been reproduced along with a line which represents the least squares line fit to the data.

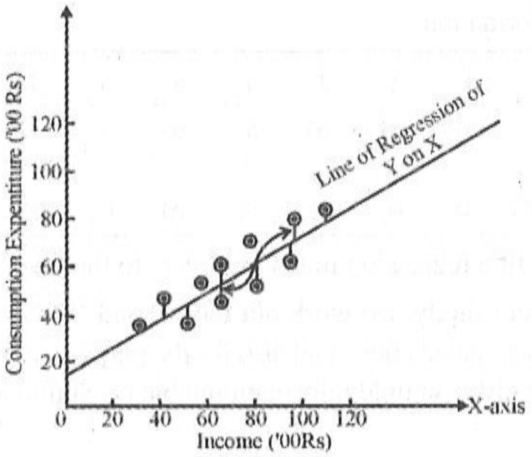


Fig. 4.5 Scatter Diagram, Regression Line and Short Vertical Lines Representing 'e'

NOTES

In Figure 4.4, the vertical deviations of the individual points from the line are shown as the short vertical lines joining the points to the least squares line. These deviations will be denoted by the symbol '*e*'. The value of '*e*' varies from one point to another. In some cases, it is positive, while in others, it is negative. If the line drawn happens to be least squares line, then the values of $\sum e_i$ is the least possible. It is so because of this feature, the method is known as Least Squares Method.

Why we insist on minimizing the sum of squared deviations is a question that needs explanation. If we denote the deviations from the actual value *Y* to the estimated value $\hat{Y}$ as $(Y - \hat{Y})$ or e_i, it is logical that we want the $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, to be as small as possible. However, mere examining $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, is inappropriate, since any e_i can be positive or negative. Large positive values and large negative values could cancel one another. But large values of e_i regardless of their sign, indicate a poor prediction. Even if we ignore the signs while working out $\sum_{i=1}^{n} | e_i |$, the difficulties may continue. Hence, the standard procedure is to eliminate the effect of signs by squaring each observation. Squaring each term accomplishes two purposes, viz., (*i*), it magnifies (or penalizes) the larger errors, and (*ii*) it cancels the effect of the positive and negative values (since a negative error when squared becomes positive). The choice of minimizing the squared sum of errors rather than the sum of the absolute values implies that there are many small errors rather than a few large errors. Hence, in obtaining the regression line, we follow the approach that the sum of the squared deviations be minimum and on this basis work out the values of its constants viz., '*a*' and '*b*' also known as the intercept and the slope of the line. This is done with the help of the following two normal equations:

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

In the above two equations, '*a*' and '*b*' are unknowns and all other values viz., ΣX, ΣY, ΣX^2, ΣXY, are the sum of the products and cross products to be calculated from the sample data, and '*n*' means the number of observations in the sample.

The following examples that explain the least squares method.

Example 4.14: Fit a regression line $\hat{Y} = a + bX_i$ by the method of least squares to the given sample information.

Observations	1	2	3	4	5	6	7	8	9	10
Income (*X*) ('00 ₹)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (*Y*) ('00 ₹)	44	60	39	51	80	68	84	34	55	48

Solution: We are to fit a regression line $\hat{Y} = a + bX_i$ to the given data by the method of least squares. Accordingly, we work out the '*a*' and '*b*' values with the help of the normal equations as stated above and also for the purpose, work out ΣX, ΣY, ΣXY, ΣX^2 values from the given sample information table on Summations for Regression Equation.

NOTES

Summations for Regression Equation

Observations	*Income* X ('00 ₹)	*Consumption Expenditure* Y ('00 ₹)	XY	X^2	Y^2
1	41	44	1804	1681	1936
2	65	60	3900	4225	3600
3	50	39	1950	2500	1521
4	57	51	2907	3249	2601
5	96	80	7680	9216	6400
6	94	68	6392	8836	4624
7	110	84	9240	12100	7056
8	30	34	1020	900	1156
9	79	55	4345	6241	3025
10	65	48	3120	4225	2304
$n = 10$	$\Sigma X = 687$	$\Sigma Y = 563$	$\Sigma XY = 42358$	$\Sigma X^2 = 53173$	$\Sigma Y^2 = 34223$

Putting the values in the required normal equations, we have,

$$563 = 10a + 687b$$

$$42358 = 687a + 53173b$$

Solving these two equations for a and b, we obtain,

$$a = 14.000 \quad \text{and} \quad b = 0.616$$

Hence, the equation for the required regression line is,

$$\hat{Y} = a + bX_i$$

or,

$$\hat{Y} = 14.000 + 0.616X_i$$

This equation is known as the regression equation of Y on X from which Y values can be estimated for given values of X variable.

Checking the Accuracy of Equation

After finding the regression line as stated above, one can check its accuracy also. The method to be used for the purpose follows from the mathematical property of a line fitted by the method of least squares, viz., the individual positive and negative errors must sum to zero. In other words, using the estimating equation, one must find out whether the term $\Sigma\left(Y - \hat{Y}\right)$ is zero, and if this is so, then one can reasonably be sure that he has not committed any mistake in determining the estimating equation.

The Problem of Prediction

When we talk about prediction or estimation, we usually imply that if the relationship $Y_i = a + bX_i + e_i$ exists, then the regression equation, $\hat{Y} = a + bX_i$ provides a base for making estimates of the value for Y which will be associated with particular values of X. In Example 4.14, we worked out the regression equation for the income and consumption data as,

$$\hat{Y} = 14.000 + 0.616X_i$$

NOTES

On the basis of this equation, we can make a *point estimate* of Y for any given value of X. Suppose we wish to estimate the consumption expenditure of individuals with income of ₹ 10,000. We substitute $X = 100$ for the same in our equation and get an estimate of consumption expenditure as follows:

$$\hat{Y} = 14.000 + 0.616(100) = 75.60$$

Thus, the regression relationship indicates that individuals with ₹ 10,000 of income may be expected to spend approximately ₹ 7,560 on consumption. However, this is only an expected or an estimated value, and it is possible that actual consumption expenditure of same individual with that income may deviate from this amount. If so, then our estimate will be an error, the likelihood of which will be high, if the estimate is applied to any one individual. The *interval estimate* method is considered better and it states an interval in which the expected consumption expenditure may fall. Remember that the wider the interval, the greater the level of confidence we can have, but the width of the interval (or what is technically known as the precision of the estimate) is associated with a specified level of confidence and is dependent on the variability (consumption expenditure in our case) found in the sample. This variability is measured by the standard deviation of the error term, 'e', and is popularly known as the standard error of the estimate.

Standard Error of the Estimate

Standard error of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. Like the standard deviation, the Standard Error (S.E.) of $\hat{Y}$ measures the variability or scatter of the observed values of Y around the regression line. Standard Error of Estimate (S.E. of $\hat{Y}$) is worked out as under:

$$\text{S.E. of } \hat{Y} \text{ (or } S_e) = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

Where, S.E. of $\hat{Y}$ (or S_e) = Standard error of the estimate

Y = Observed value of Y

$\hat{Y}$ = Estimated value of Y

e = The error term = $(Y - \hat{Y})$

n = Number of observations in the sample

Note: In the above Formula, $n - 2$ is used instead of n because of the fact that two degrees of freedom are lost in basing the estimate on the variability of the sample observations about the line with two constants, viz., 'a' and 'b' whose position is determined by those same sample observations.

The square of the S_e, also known as the variance of the error term, is the basic measure of reliability. The larger the variance, the more significant are the magnitudes of the e's and the less reliable is the regression analysis in predicting the data.

Interpreting the standard error of estimate and finding the confidence limits for the estimate in large and small samples

The larger the S.E. of estimate (SE_e), the greater happens to be the dispersion, or scattering, of given observations around the regression line. But if the S.E. of estimate happens to be zero, then the estimating equation is a 'perfect' estimator (i.e., cent per cent correct estimator) of the dependent variable.

NOTES

In case of large samples, i.e., where $n > 30$ in a sample, it is assumed that the observed points are normally distributed around the regression line and we may find,

68% of all points within $\hat{Y} \pm 1\ \text{SE}_e$ limits

95.5% of all points within $\hat{Y} \pm 2\ \text{SE}_e$ limits

99.7% of all points within $\hat{Y} \pm 3\ \text{SE}_e$ limits

This can be stated as,

(*i*) The observed values of Y are normally distributed around each estimated value of $\hat{Y}$.

(*ii*) The variance of the distributions around each possible value of $\hat{Y}$ is the same.

In case of small samples, i.e., where $n \leq 30$ in a sample the '*t*' distribution is used for finding the two limits more appropriately.

This is done as follows:

$$\text{Upper limit} = \hat{Y} + \text{'}t\text{'}\ (\text{SE}_e)$$

$$\text{Lower limit} = \hat{Y} - \text{'}t\text{'}\ (\text{SE}_e)$$

Where, $\hat{Y}$ = The estimated value of Y for a given value of X

SE_e = The standard error of estimate

't' = Table value of 't' for given degrees of freedom for a specified confidence level

Some Other Details Concerning Simple Regression

Sometimes the estimating equation of Y, also known as the regression equation of Y on X, is written as follows:

$$\left(\hat{Y} - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right)$$

or,

$$\hat{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right) + \overline{Y}$$

Where, r = Coefficient of simple correlation between X and Y

σ_Y = Standard deviation of Y

σ_X = Standard deviation of X

$\overline{X}$ = Mean of X

$\overline{Y}$ = Mean of Y

$\hat{Y}$ = Value of Y to be estimated.

X_i = Any given value of X for which Y is to be estimated

This is based on the formula we have used, i.e., $\hat{Y} = a + bX_i$. The coefficient of X_i is defined as,

$$\text{Coefficient of } X_i = b = r\frac{\sigma_Y}{\sigma_X}$$

Also known as regression coefficient of *Y* on *X* or slope of the regression line of *Y* on *X* or b_{YX}:

$$= \frac{\Sigma XY - n\overline{X}\overline{Y} \times \sqrt{\Sigma Y^2 - n\overline{Y}^2}}{\sqrt{\Sigma Y^2 - n\overline{Y}^2}\sqrt{\Sigma X^2 - n\overline{X}^2}\sqrt{\Sigma X^2 - n\overline{X}^2}}$$

$$= \frac{\Sigma XY - n\overline{X}\overline{Y}}{\Sigma X^2 - n\overline{X}^2}$$

And,

$$= -r\frac{\sigma_Y}{\sigma_X}\overline{X} + \overline{Y}$$

$$= \overline{Y} - b\overline{X} \qquad \left(\text{since } b = r\frac{\sigma_Y}{\sigma_X}\right)$$

Similarly, the estimating equation of *X*, also known as the regression equation of *X* on *Y*, can be stated as:

$$\left(\hat{X} - \overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right)$$

or,

$$\hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right) + \overline{X}$$

And the regression coefficient of *X* on *Y* (or b_{XY}) $= r\dfrac{\sigma_X}{\sigma_Y} = \dfrac{\Sigma XY - n\overline{X}\overline{Y}}{\Sigma Y^2 - n\overline{Y}^2}$

If we are given the two regression equations as stated above, along with the values of '*a*' and '*b*' constants to solve the same for finding the value of *X* and *Y*, then the values of *X* and *Y* so obtained, are the mean value of *X* (i.e., $\overline{X}$) and the mean value of *Y* (i.e., $\overline{Y}$).

If we are given the two regression coefficients (viz., b_{XY} and b_{YX}), then we can work out the value of coefficient of correlation by just taking the square root of the product of the regression coefficients as shown below:

$$r = \sqrt{b_{YX} . b_{XY}}$$

$$= \sqrt{r\frac{\sigma_Y}{\sigma_X} . r\frac{\sigma_X}{\sigma_Y}}$$

$$= \sqrt{r.r} \quad = r$$

The (±) sign of *r* will be determined on the basis of the sign of the regression coefficients given. If regression coefficients have minus sign, then *r* will be taken with minus (–) sign and if regression coefficients have plus sign, then *r* will be taken with plus (+) sign. (Remember that both regression coefficients will necessarily have the same sign whether it is minus or plus, for their sign is governed by the sign of coefficient of correlation.)

NOTES

Example 4.15: Given is the following information:

	$\overline{X}$	$\overline{Y}$
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Simple correlation coefficient between X and Y is = + 0.42

Find the estimating equation of Y and X.

Solution: Estimating equation of Y can be worked out as,

$$\because \quad \left(\hat{Y}-\overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X_i-\overline{X}\right)$$

$$\text{or,} \quad \hat{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X_i-\overline{X}\right)+\overline{Y}$$

$$= 0.42\frac{17.8}{10.8}(X_i-39.5)+47.5$$

$$= 0.69X_i - 27.25 + 47.5$$

$$= 0.69X_i + 20.25$$

Similarly, the estimating equation of X can be worked out as under:

$$\because \quad \left(\hat{X}-\overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y_i-\overline{Y}\right)$$

$$\text{or,} \quad \hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y_i-\overline{Y}\right)+\overline{X}$$

$$\text{or,} \quad = 0.42\frac{10.8}{17.8}(Y_i-47.5)+39.5$$

$$= 0.26Y_i - 12.35 + 39.5$$

$$= 0.26Y_i + 27.15$$

Example 4.16: Given is the following data:

Variance of $X = 9$

Regression equations:

$$4X - 5Y + 33 = 0$$

$$20X - 9Y - 107 = 0$$

Find: (*i*) Mean values of X and Y

(*ii*) Coefficient of Correlation between X and Y

(*iii*) Standard deviation of Y

Solution: The solution is obtained as follows:

(i) For finding the mean values of X and Y, we solve the two given regression equations for the values of X and Y as follows:

$$4X - 5Y + 33 = 0 \qquad (1)$$

$$20X - 9Y - 107 = 0 \qquad (2)$$

If we multiply Equation (1) by 5, we have the following equations:

$$20X - 25Y = -165 \quad (3)$$

$$20X - 9Y = 107 \quad (2)$$

$$\underline{- \quad + \quad\quad -}$$

$$-16Y = -272$$

Subtracting Equation (2) from Equation (3)

or, $Y = 17$

Putting this value of Y in Equation (1) we have,

$$4X = -33 + 5(17)$$

or, $$X = \frac{-33+85}{4} = \frac{52}{4} = 13$$

Hence, $\overline{X} = 13$ and $\overline{Y} = 17$

(ii) For finding the coefficient of correlation, first of all, we presume one of the two given regression equations as the estimating equation of X. Let equation $4X - 5Y + 33 = 0$ be the estimating equation of X, then we have:

$$\hat{X} = \frac{5Y_i}{4} - \frac{33}{4}$$

And,

From this, we can write $b_{XY} = \frac{5}{4}$

The other given equation is then taken as the estimating equation of Y and can be written as:

$$\hat{Y} = \frac{20X_i}{9} - \frac{107}{9}$$

and from this, we can write $b_{YX} = \frac{20}{9}$

If the above equations are correct, then r must be equal to,

$$r = \sqrt{5/4 \times 20/9} = \sqrt{25/9} = 5/3 = 1.6$$

which is an impossible equation, since r can in no case be greater than 1. Hence, we change our supposition about the estimating equations and by reversing it, we re-write the estimating equations as under:

$$\hat{X} = \frac{9Y_i}{20} + \frac{107}{20}$$

And, $$\hat{Y} = \frac{4X_i}{5} + \frac{33}{5}$$

Hence, $$r = \sqrt{9/20 \times 4/5}$$

$$= \sqrt{9/25}$$

$$= 3/5$$

$$= 0.6$$

Since, regression coefficients have plus signs, we take $r = + 0.6$

(*iii*) Standard deviation of Y can be calculated as follows:

$\because$ Variance of $X = 9$ $\therefore$ Standard deviation of $X = 3$

$$\because \qquad b_{YX} = r\frac{\sigma_Y}{\sigma_X} = \frac{4}{5} = 0.6\frac{\sigma_Y}{3} = 0.2\sigma_Y$$

Hence, $\sigma_Y = 4$

Alternatively, we can work it out as under:

$$\because \qquad b_{XY} = r\frac{\sigma_X}{\sigma_Y} = \frac{9}{20} = 0.6\frac{\sigma_Y}{3} = \frac{1.8}{\sigma_Y}$$

Hence, $\sigma_Y = 4$

4.8 TIME SERIES

The time series analysis method is quite accurate where future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner as in the past. These techniques are fairly sophisticated and require experts to use these methods.

The classical approach is to analyse a time series in terms of four distinct types of variations or separate components that influence a time series.

1. Secular Trend or Simply Trend (*T*)

Trend is a general long-term movement in the time series value of the variable (*Y*) over a fairly long period of time. The variable (*Y*) is the factor that we are interested in evaluating for the future. It could be sales, population, crime rate, and so on.

Trend is a common word, popularly used in day-to-day conversation, such as population trends, inflation trends and birth rate. These variables are observed over a long period of time, and any changes related to time are noted and calculated and a trend of these changes is established. There are many types of trends; the series may be increasing at a slow rate or at a fast rate or these may be decreasing at various rates. Some remain relatively constant and some reverse their trend from growth to decline or from decline to growth over a period of time. These changes occur as a result of the general tendency of the data to increase or decrease as a result of some identifiable influences.

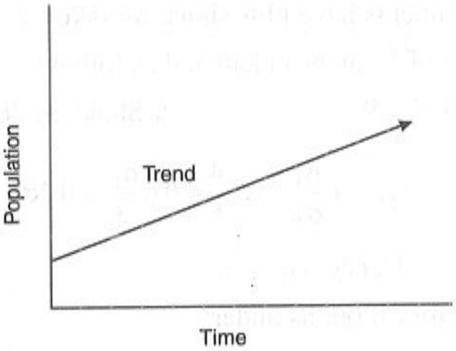
If a trend can be determined and the rate of change can be ascertained, then tentative estimates on the same series values into the future can be made. However, such forecasts are based upon the assumption that the conditions affecting the steady growth or decline are reasonably expected to remain unchanged in the future. A change in these conditions would affect the forecasts. As an example, a time series involving increase in population over time can be shown as:

NOTES

Check Your Progress

12. State the theory of correlation.
13. Name the methods used for estimating the regression equation.

NOTES



2. Cyclical Fluctuations (*C*)

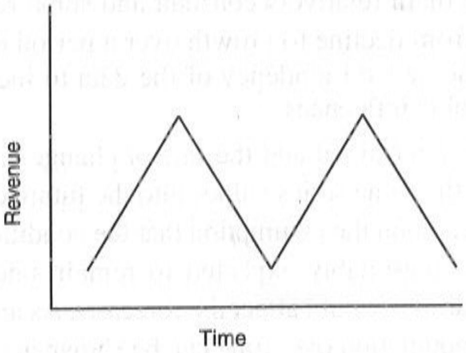
Cyclical fluctuations refer to regular swings or patterns that repeat over a long period of time. The movements are considered cyclical only if they occur after time intervals of more than one year. These are the changes that take place as a result of economic booms or depressions. These may be up or down, and are recurrent in nature and have a duration of several years—usually lasting for two to ten years. These movements also differ in intensity or amplitude and each phase of movement changes gradually into the phase that follows it. Some economists believe that the business cycle completes four phases every twelve to fifteen years. These four phases are prosperity, recession, depression and recovery. However, there is no agreement on the nature or causes of these cycles.

Even though measurement and prediction of cyclical variation is very important for strategic planning, the reliability of such measurements is highly questionable due to the following reasons:

(*i*) These cycles do not occur at regular intervals. In the twenty-five years from 1956 to 1981 in America, it is estimated that the peaks in the cyclical activity of the overall economy occurred in August 1957, April 1960, December 1969, November 1973 and January 1980. This shows that they differ widely in timing, intensity and pattern, thus, making reliable evaluation of trends very difficult.

(*ii*) The cyclic variations are affected by many erratic, irregular and random forces which cannot be isolated and identified separately, nor can their impact be measured accurately.

The cyclic variation for revenues in an industry against time is shown graphically as follows:



3. Seasonal Variation (*S*)

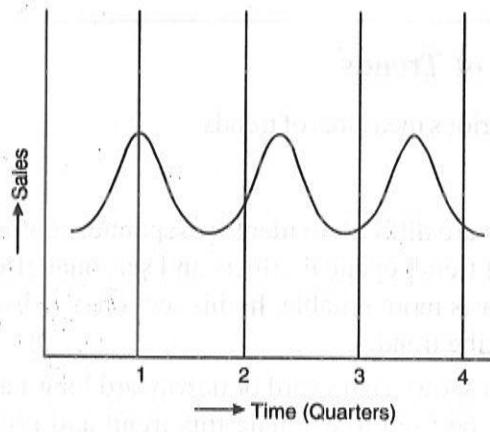
Seasonal variation involves patterns of change that repeat over a period of one year or less. Then they repeat from year to year and they are brought about by fixed events. For example, sales of consumer items increase prior to Christmas due to gift giving tradition. The sale of automobiles in America are much higher during the last three to four months of the year due to the introduction of new models. This data may be measured monthly or quarterly.

Since these variations repeat during a period of twelve months, they can be predicted fairly and accurately. Some factors that cause seasonal variations are as follows:

(i) Season and Climate: Changes in the climate and weather conditions have a profound effect on sales. For example, the sale of umbrellas in India is always more during monsoons. Similarly, during winter, there is a greater demand for woollen clothes and hot drinks, while during summer months, there is an increase in the sales of fans and air conditioners.

(ii) Customs and Festivals: Customs and traditions affect the pattern of seasonal spending. For example, Mother's Day or Valentine's Day in America see increase in gift sales preceding these days. In India, festivals such as Baisakhi and Diwali mean a big demand for sweets and candy. It is customary all over the world to give presents to children when they graduate from high school or college. Accordingly, the month of June, when most students graduate, is a time for the increase of sale for presents befitting the young.

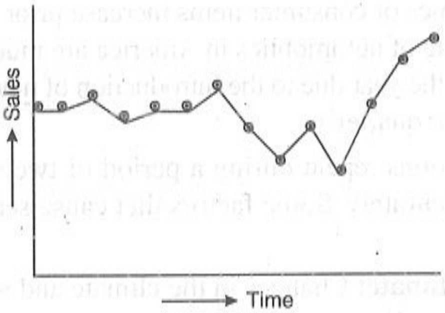
An accurate assessment of seasonal behaviour is an aid in business planning and scheduling such as in the area of production, inventory control, personnel, advertising, and so on. The seasonal fluctuations over four repeating quarters in a given year for sale of a given item is illustrated as:



4. Irregular or Random Variation (*I*)

These variations are accidental, random or simply due to chance factors. Thus, they are wholly unpredictable. These fluctuations may be caused by such isolated incidents as floods, famines, strikes or wars. Sudden changes in demand or a breakthrough in technological development may be included in this category. Accordingly, it is almost impossible to isolate and measure the value and the impact of these erratic movements on forecasting models or techniques. This phenomenon may be graphically shown as follows:

NOTES



It is traditionally acknowledged that the value of the time series (Y) is a function of the impact of variable trend (T), seasonal variation (S), cyclical variation (C) and irregular fluctuation (I). These relationships may vary depending upon assumptions and purposes. The effects of these four components might be additive, multiplicative, or combination thereof in a number of ways. However, the traditional time series analysis model is characterized by multiplicative relationship, so that:

$$Y = T \times S \times C \times I$$

This model is appropriate for those situations where percentage changes best represent the movement in the series and the components are not viewed as absolute values but as relative values.

Another approach to define the relationship may be additive, so that:

$$Y = T + S + C + I$$

This model is useful when the variations in the time series are in absolute values and can be separated and traced to each of these four parts and each part can be measured independently.

4.8.1 Measurement of Trends

The following are the various measures of trends:

Trend Analysis

While chance variations are difficult to identify, separate, control or predict, a more precise measurement of trend, cyclical effects and seasonal effects can be made in order to make the forecasts more reliable. In this section, we discuss techniques that would allow us to describe trend.

When a time series shows an upward or downward long-term linear trend, then regression analysis can be used to estimate this trend and project the trends into forecasting the future values of the variables involved. The equation for the straight line used to describe the linear relationship between the independent variable X and the dependent variable Y is:

$$Y = b_0 + b_1X$$

where, b_0 = Intercept on the Y-axis and b_1 = Slope of the straight line

In time series analysis, the independent variable is time, so we will use the symbol t in place of X and we will use the symbol Y_t in place of Y_c, which we have used previously.

NOTES

Hence, the equation for linear trend is given as:

$$Y_t = b_0 + b_1 t$$

where, Y_t = Forecast value of the time series in period t

b_0 = Intercept of the trend line on Y-axis

b_1 = Slope of the trend line

t = Time period

As discussed earlier, we can calculate the values of b_0 and b_1 by the following formulae:

$$b_1 = \frac{n\Sigma(ty) - (\Sigma t)(\Sigma y)}{n(\Sigma t^2) - (\Sigma t)^2}, \text{ and } b_0 = \overline{y} - b_1\overline{t}$$

where, y = Actual value of the time series in period time t

n = Number of periods

$$\overline{y} = \text{Average value of time series} = \frac{\Sigma y}{n}$$

$$\overline{t} = \text{Average value of } t = \frac{\Sigma t}{n}$$

Knowing these values, we can calculate the value of y

Example 4.17:

A car fleet owner has 5 cars which have been in the fleet for several different years. The manager wants to establish if there is a linear relationship between the age of the car and the repairs in hundreds of dollars for a given year. This way, he can predict the repair expenses for each year as the cars become older. The information for the repair costs he had collected for the last year on these cars is as follows:

Car #	*Age (t)*	*Repairs (Y)*
1	1	4
2	3	6
3	3	7
4	5	7
5	6	9

The manager wants to predict the repair expenses for the next year for the two cars that are 3 years old.

Solution:

The trend in repair costs suggests a linear relationship with the age of the car, so that the linear regression equation is given as:

$$Y_t = b_0 + b_1 t$$

where
$$b_1 = \frac{n\Sigma(ty) - (\Sigma t)(\Sigma y)}{n(\Sigma t^2) - (\Sigma t)^2}$$

and
$$b_0 = \overline{y} - b_1\overline{t}$$

To calculate the various values, let us form a new table as follows:

Age of Car (t)	*Repair Cost* (Y)	tY	t^2
1	4	4	1
3	6	18	9
3	7	21	9
5	7	35	25
6	9	54	36
Total 18	33	132	80

Knowing that $n = 5$, let us substitute these values to calculate the regression coefficients b_0 and b_1.

Then, $$b_1 = \frac{5(132)-(18)(33)}{5(80)-(18)^2}$$

$$= \frac{660-594}{400-324}$$

$$= \frac{66}{76} = 0.87$$

and $$b_0 = \bar{y} - b_1\bar{t}$$

where $$\bar{y} = \frac{\Sigma y}{n} = \frac{33}{5} = 6.6$$

and $$\bar{t} = \frac{t}{n} = \frac{18}{5} = 3.6$$

Then, $$b_0 = 6.6 - 0.87(3.6)$$

$$= 6.6 - 3.13$$

$$= 3.47$$

Hence, $$Y_t = 3.47 + 0.87t$$

The cars that are 3 years old now will be 4 years old next year, so that $t = 4$.

Hence, $$Y_{(4)} = 3.47 + 0.87(4)$$

$$= 3.47 + 3.48$$

$$= 6.95$$

Accordingly, the repair costs on each car that is 3 years old are expected to be \$695.00.

Smoothing Techniques

Smoothing techniques improve the forecasts of future trends provided that the time series is fairly stable with no significant trend, cyclical or seasonal effect and the objective is to smooth out the irregular component of the time series through the averaging process. There are two techniques that are generally employed for such smoothing:

1. Moving averages 2. Exponential smoothing

These are discussed as follows:

1. Moving Averages: The concept of the moving averages is based on the idea that any large irregular component of time series at any point in time will have a less significant impact on the trend, if the observation at that point in time is averaged with such values immediately before and after the observation under consideration. For example, if we are interested in computing the three-period moving average for any time period, then we will take the average of the value in such time period, the value in the period immediately preceding it and the value in the time period immediately following it. Let us illustrate this concept with the help of an example.

Example 4.18: Let the following table represent the number of cars sold in the first 6 weeks of the first two months of the year by a given dealer. Our objective is to calculate the three-week moving average.

Week	*Sales*
1	20
2	24
3	22
4	26
5	21
6	22

Solution: The moving average for the first three-week period is given as:

$$\text{Moving average} = \frac{20 + 24 + 22}{3} = \frac{66}{3} = 22$$

This moving average can then be used to forecast the sale of cars for week 4. Since the actual number of cars sold in week 4 is 26, we note that the error in the forecast is (26 – 22) = 4.

The calculation for the moving average for the next three periods is done by adding the value for week 4 and dropping the value for week 1, and taking the average for weeks 2, 3 and 4. Hence,

$$\text{Moving average} = \frac{24 + 22 + 26}{3} = \frac{72}{3} = 24$$

Then, this is considered to be the forecast of sales for week 5. Since the actual value of the sales for week 5 is 21, we have an error in our forecast of (21 – 24) = – (3).

The next moving average for weeks 3 to 5, as a forecast for week 6 is given as:

$$\text{Moving average} = \frac{22 + 26 + 21}{3} = \frac{69}{3} = 23$$

The error between the actual and the forecast value for week 6 is (22 – 23) = – (1). (Since the actual value of the sales for week 7 is not given, there is no need to forecast such values.)

Our objective is to predict the trend and forecast the value of a given variable in the future as accurately as possible so that the forecast is reasonably free from random variations. To do that, we must have the sum of individual errors, as discussed earlier, as little as possible. However, since errors are irregular and random, it is expected that some errors would be positive in value and others negative, so that the sum of these errors would be highly distorted and would be closer to zero. This difficulty can be avoided by squaring each of the individual forecast errors and then taking the average. Naturally, the minimum values of these errors would also result in the minimum value of the 'average of the sum of squared errors'. This is shown as follows:

Week	*Time Series Value*	*Moving Average*	*Error*	*Error Squared*
1	20			
2	24			
3	22			
4	26	22	4	16
5	21	24	− 3	9
6	22	23	− 1	1

Then the average of the sum of squared errors, also known as *Mean Squared Error* (MSE), is given as:

$$\text{MSE} = \frac{16+9+1}{3} = \frac{26}{3} = 8.67$$

The value of MSE is an often-used measure of the accuracy of the forecasting method, and the method which results in the least value of MSE is considered more accurate than others. The value of MSE can be manipulated by varying the number of data values to be included in the moving average. For example, if we had calculated the value of MSE by taking 4 periods into consideration for calculating the moving average, rather than 3, then the value of MSE would be less. Accordingly, by using trial and error method, the number of data values selected for use in forecasting would be such that the resulting MSE value would be minimum.

2. Exponential Smoothing using Least Square Method: In the moving average method, each observation in the moving average calculation receives the same weight. In other words, each value contributes equally towards the calculation of the moving average, irrespective of the number of time periods taken into consideration. In most actual situations, this is not a realistic assumption. Because of the dynamics of the environment over a period of time, it is more likely that the forecast for the next period would be closer to the most recent previous period than the more distant previous period, so that the more recent value should get more weight than the previous value, and so on. The exponential smoothing technique uses the moving average with appropriate weights assigned to the values taken into consideration in order to arrive at a more accurate or smoothed forecast. It takes into consideration the decreasing impact of the past time periods as we move further into the past time periods. This decreasing impact as we move down into the time period is exponentially distributed and, hence, the name exponential smoothing.

In this method, the smoothed value for period t, which is the weighted average of that period's actual value and the smoothed average from the previous period $(t - 1)$, becomes the forecast for the next period $(t + 1)$. Then the exponential smoothing model for time period $(t + 1)$ can be expressed as follows:

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

where $F_{(t+1)}$ = The forecast of the time series for period $(t + 1)$

Y_t = Actual value of the time series in period t

NOTES

α = Smoothing factor $(0 \leq \alpha \leq 1)$

F_t = Forecast of the time series for period t

The value of α is selected by the decision-maker on the basis of the degree of smoothing required. A small value of α means a greater degree of smoothing. A large value of α means very little smoothing. When $\alpha = 1$, then there is no smoothing at all so that the forecast for the next time period is exactly the same as the actual value of times series in the current period. This can be seen by:

$$F_{(t+1)} = \alpha Y_t + (1-\alpha)F_t$$

when $\alpha = 1$

$$F_{(t+1)} = Y_t + 0F_t = Y_t$$

The exponential smoothing approach is simple to use and once the value of α is selected, it requires only two pieces of information, namely Y_t and F_t to calculate $F_{(t+1)}$.

To begin with the exponential smoothing process, we let F_t equal the actual value of the time series in period t, which is Y_1. Hence, the forecast for period 2 is written as:

$$F_2 = \alpha Y_1 + (1-\alpha)\, F_1$$

But since we have put $F_1 = Y_1$, hence:

$$F_2 = \alpha Y_1 + (1-\alpha)\, Y_1$$
$$= Y_1$$

Let us now apply exponential smoothing method to the problem of forecasting car sales as discussed in the case of moving averages. The data once again is given as follows:

Week	*Time Series Value* (Y_t)
1	20
2	24
3	22
4	26
5	21
6	22

Let $\alpha = 0.4$

Since F_2 is calculated earlier as equal to $Y_1 = 20$, we can calculate the value of F_3 as follows:

$$F_3 = 0.4Y_2 + (1-0.4)\, F_2$$

Since $F_2 = Y_1$, we get

$$F_3 = 0.4(24) + 0.6(20) = 9.6 + 12$$
$$= 21.6$$

Similar values can be calculated for subsequent periods, so that:

$$F_4 = 0.4Y_3 + 0.6F_3$$
$$= 0.4(22) + 0.6(21.6)$$
$$= 8.8 + 12.96$$
$$= 21.76$$

NOTES

$$\begin{aligned} F_5 &= 0.4Y_4 + 0.6F_4 \\ &= 0.4(26) + 0.6(21.76) \\ &= 10.4 + 13.056 \\ &= 23.456 \\ F_6 &= 0.4Y_5 + 0.6F_5 \\ &= 0.4(21) + 0.6(23.456) \\ &= 8.4 + 14.07 \\ &= 22.47 \end{aligned}$$

and,

$$\begin{aligned} F_7 &= 0.4Y_6 + 0.6F_6 \\ &= 0.4(22) + 0.6(22.47) \\ &= 8.8 + 13.48 \\ &= 22.28 \end{aligned}$$

Now we can compare the exponential smoothing forecast value with the actual values for the six time periods and calculate the forecast error.

Week	*Time Series Value* (Y_t)	*Exponential Smoothing Forecast Value* (F_t)	*Error* $(Y_t - F_t)$
1	20	–	–
2	24	20.000	4.0
3	22	21.600	0.4
4	26	21.760	4.24
5	21	23.456	– 2.456
6	22	22.470	– 0.47

(The value of F_7 is not considered because the value of Y_7 is not given.)

Let us now calculate the value of MSE for this method with selected value of $\alpha = 0.4$. From the previous table:

Forecast errors $(Y_t - F_t)$	Squared Forecast Error $(Y_t - F_t)$
4	16
0.4	0.16
4.24	17.98
– 2.456	6.03
– 0.47	0.22
	Total = 40.39

Then,

$$\begin{aligned} \text{MSE} &= 40.39/5 \\ &= 8.08 \end{aligned}$$

The previous value of MSE was 8.67. Hence, the current approach is a better one.

The choice of the value for α is very significant. Let us look at the exponential smoothing model again.

$$\begin{aligned} F_{(t+1)} &= \alpha Y_t + (1-\alpha)F_t \\ &= \alpha Y_t + F_t - \alpha F_t \\ &= F_t + \alpha(Y_t - F_t) \end{aligned}$$

where $(Y_t - F_t)$ is the forecast error during the time period t.

The accuracy of the forecast can be improved by carefully selecting the value of α. If the time series contains substantial random variability, then a small value of α (known as smoothing factor or smoothing constant) is preferable. On the other hand, a larger value of α would be desirable for time series with relatively little random variability $(Y_t - F_t)$.

NOTES

Measuring Cyclical Effect

Cyclic variation, as we have discussed earlier, is a pattern that repeats over time periods longer than one year. These variations are generally unpredictable in relation to the time of occurrence, duration as well as amplitude. However, these variations have to be separated and identified. The measure we use to identify cyclical variation is the *percentage of trend* and the procedure used is known as the *residual trend.*

As we have discussed earlier, there are four components of time series. These are secular trend (T), seasonal variation (S), cyclical variation (C) and irregular (or chance) variation (I). Since the time period considered for seasonal variation is less than one year, it can be excluded from the study, because when we look at time series consisting of annual data spread over many years, then only the secular trend, cyclical variation and irregular variation are considered.

Since secular trend component can be described by the trend line (usually calculated by line of regression), we can isolate cyclical and irregular components from the trend. Furthermore, since irregular variation occurs by chance and cannot be predicted or identified accurately, it can be reasonably assumed that most of the variation in time series left unexplained by the trend component can be explained by the cyclical component. In that respect, cyclical variation can be considered as the *residual*, once other causes of variation have been identified.

The measure of cyclic variation as percentage of trend is calculated as follows:

(1) Determine the trend line (usually by regression analysis).

(2) Compute the trend value Y_t for each time period (t) under consideration.

(3) Calculate the ratio Y/Y_t for each time period.

(4) Multiply this ratio by 100 to get the percentage of trend, so that:

$$\text{Percentage of trend} = \left(\frac{Y}{Y_t}\right)100$$

Freehand Curve Method

This is a simple method of studying trends. In this method, the given time series data are plotted on graph paper by taking time on X-axis and the other variable on Y-axis. The graph obtained will be irregular as it would include short-run oscillations. We may observe the up and down movement of the curve and if a smooth freehand curve is drawn passing approximately all points of a curve previously drawn, it would eliminate the short-run oscillations (seasonal, cyclical and irregular variations) and show the long-period general tendency of the data.

This is exactly what is meant by **trend**. However, it is very difficult to draw a freehand smooth curve and different persons are likely to draw different curves from the same data. The following points must be kept in mind in drawing a freehand smooth curve:

NOTES

1. That the curve is smooth.
2. That the numbers of points above the line or curve are equal to the points below it.
3. That the sum of vertical deviations of the points above the smoothed line is equal to the sum of the vertical deviations of the points below the line. In this way, the positive deviations will cancel the negative deviations. These deviations are the effects of seasonal cyclical and irregular variations, and by this process, they are eliminated.
4. The sum of the squares of the vertical deviations from the trend line curve is minimum. (This is one of the characteristics of the trend line fitted by the method of lest squares.)

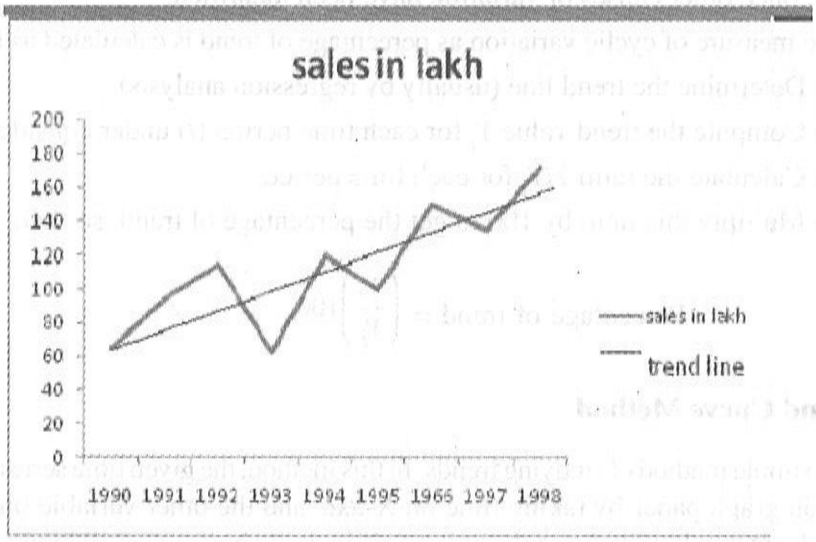
The trend values can be read for various time periods by locating them on the trend line against each time period. The following example will illustrate the fitting of a freehand curve to set of time series values:

Example:

The table below shows the data of sale of nine years:

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998
Sales in (lakh units)	65	95	115	63	120	100	150	135	172

If we draw a graph taking year on x-axis and sales on y- axis, it will be irregular as shown below. Now drawing a freehand curve passing approximately through all this points will represent trend line (shown below by black line).



Merits:

The following are the merits of freehand curve method:

1. It is a simple method of estimating trend which requires no mathematical calculations.
2. It is a flexible method as compared to rigid mathematical trends and, therefore, a better representative of the trend of the data.
3. This method can be used even if trend is not linear.

4. If the observations are relatively stable, the trend can easily be approximated by this method.
5. Being a non-mathematical method, it can be applied even by a common man.

Demerits:

The following are the demerits of freehand curve method:

1. It is a subjective method. The values of trend obtained by different statisticians would be different and, hence, not reliable.
2. Predictions made on the basis of this method are of little value.

4.9 TESTING OF HYPOTHESIS AND STATISTICAL TESTING

A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning. However, it is often referred as a convenient mathematical approach for simplifying cumbersome calculation. Setting up and testing hypothesis is an integral art of statistical inference. Hypotheses are often statements about population parameters like variance and expected value. During the course of hypothesis testing, some inference about population like the mean and proportion are made. Any useful hypothesis will enable predictions by reasoning including deductive reasoning. According to well-known philosopher Karl Popper, a hypothesis must be falsifiable and that a proposition or theory cannot be called scientific if it does not admit the possibility of being shown false. Hypothesis might predict outcome of an experiment in a lab, setting the observation of a phenomenon in nature. Thus, hypothesis is a explanation of a phenomenon proposal, suggesting a possible correlation between multiple phenomena.

Hypothesis is put forward as a proposition. It may even be a set of more than one proposition. A proposition is the antecedent of a conditional proposition which may be an assumption or a guess. It is something yet to be proved, but taken to be temporarily true.

Hypothesis is applied in Natural Sciences, as a tentative theory provisionally adopted to explain few facts that provide guidance in proving other facts. This is often known as a working hypothesis.

A hypothesis may be a proposal that explains certain facts based on certain observations. It may be a message which is an opinion, and it may be based on certain evidences which may be incomplete.

Formalized hypotheses have two variables, independent and dependent. The independent variable is the person, may be the scientist, who is going to put the hypothesis and the dependent variable is one that the person observes.

A good hypothesis has three characteristics. It is testable as it is based on sound rationale, and it is practical and ethical to conduct the test.

Statistical Hypothesis cannot ascertain the truth of the population parameter. To do this, truth table for entire population is required to be examined which is time consuming and impractical. Researchers examine a random sample and after judging its consistency, the hypothesis is accepted.

Check Your Progress

14. List the factors that cause seasonal variation.
15. State the concept of moving averages.

The characteristics of hypothesis are as follows:

- **Clear and Accurate:** Hypothesis should be clear and accurate so as to draw a consistent conclusion.
- **Statement of Relationship between Variables:** If a hypothesis is relational, it should state the relationship between different variables.
- **Testability:** A hypothesis should be open to testing so that other deductions can be made from it and can be confirmed or disproved by observation. The researcher should do some prior study to make the hypothesis a testable one.
- **Specific with Limited Scope:** A hypothesis, which is specific with limited scope, is easily testable than a hypothesis with limitless scope. Therefore, a researcher should pay more time to do research on such kind of hypothesis.
- **Simplicity:** A hypothesis should be stated in the most simple and clear terms to make it understandable.
- **Consistency:** A hypothesis should be reliable and consistent with established and known facts.
- **Time Limit:** A hypothesis should be capable of being tested within a reasonable time. In other words, it can be said that the excellence of a hypothesis is judged by the time taken to collect the data needed for the test.
- **Empirical Reference:** A hypothesis should explain or support all the sufficient facts needed to understand what the problem is all about.

A hypothesis is a statement or assumption concerning a population. For the purpose of decision-making, a hypothesis has to be verified and then accepted or rejected. This is done with the help of observations. We test a sample and make a decision on the basis of the result obtained. Decision-making plays significant role in different areas such as marketing, industry and management.

Statistical Decision-Making

Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enable us to accept or reject the hypothesis. Since the sample data give incomplete information about the population, the result of the test need not be considered to be final or unchallengeable. The procedure, on which the basis of sample results, enables to decide whether a hypothesis is to be accepted or rejected. This is called Hypothesis Testing or Test of Significance.

Note 1: If a test provides evidence, if any, against a hypothesis, it is usually called a null hypothesis. The test cannot prove the hypothesis to be correct. It can give some evidence against it.

The test of hypothesis is a procedure to decide whether to accept or reject a hypothesis.

Note 2: The acceptance of a hypotheses implies if there is no evidence from the sample that we should believe otherwise.

The rejection of a hypothesis leads us to conclude that it is false. This way of putting the problem is convenient because of the uncertainty inherent in the problem. In view of this, we must always briefly state a hypothesis that we *hope to reject.*

A hypothesis stated in the hope of being rejected is called a *null hypothesis* and is denoted by H_0.

If H_0 is rejected, it may lead to the acceptance of an alternative hypothesis denoted by H_1.

For example, new fragrance soap is introduced in the market. The null hypothesis H_0, which may be rejected, is that the new soap is not better than the existing soap.

Similarly, a dice is suspected to be rolled. Roll the dice a number of times to test.

The Null Hypothesis H_0: $p = 1/6$ for showing six.

The Alternative hypothesis H_1: $p \neq 1/6$.

For example, skulls found at an ancient site may all belong to race *X* or race *Y* on the basis of their diameters. We may test the hypothesis that the mean is μ of the population from which the present skulls came. We have the hypotheses.

$H_0 : \mu = \mu_x, H_1 : \mu = \mu_y$

Here, we should not insist on calling either hypothesis null and the other alternative since the reverse could also be true.

Simple and Composite Hypotheses

A simple hypothesis is one that specifies complete population distribution.

For example,

1. H_0: $X \sim Bi(150,1/2)$, i.e., p is given ($p = ½$)
2. H_0: $X \sim N(8,30)$, i.e., μ and s^2 are given.

In composite hypothesis, population distribution is *not* specified completely.

The following examples will make the concept clear:

1. $X \sim Bi(150,p)$ and H_1: $p > 0.5$
2. $X \sim N(0, s^2)$ and H_1: s^2 is not specified.

Composite hypothesis has one or more free parameters. We can cite an example of a hypothesis that the decay of a particle of a radioactive element is purely exponential with unknown lifetime. This is a composite hypothesis.

Testing of Simple Hypothesis

Many applications in language engineering require testing of hypotheses. Suppose we have to differentiate between a person's 'reading a speech' and 'giving spontaneous speech'. If testing aims at differentiating between read and spontaneous speech with respect to selected statistics and the criteria is put as mean vowel duration in the two conditions where speech was recorded, then this is simple hypothesis testing since it involves a parameter of a single population.

Concept involved in such a testing is making alternative assertions about the likely outcome of an analysis. One assertion is there is no difference between the two conditions. This is null hypothesis, denoted as H_0, which asserts that the mean tone unit duration in the read speech is the same as that in the spontaneous speech.

There may be other assertions which are called *alternative hypotheses*, denoted *as* H_1 or H_a. An alternative hypothesis may assert that the tone unit duration of the read speech will be less than that of the spontaneous speech. A second may be the converse of it.

One-Tailed or Two-Tailed Hypotheses

The decision on selection of alternate hypotheses, to propose, depends on factors leading the language engineer to note differences in one direction or the other. These instances are referred to as *one-tailed or two-tailed hypotheses* depending on whether differentiation is being done for one direction or both the directions. Here, large differences between the means of the read and spontaneous speech, regardless of the direction followed may give evidence in favour of the alternative hypothesis.

It is important to make distinction between one-tailed and a two-tailed test. This affects the decision to assert a significant difference and this supports the null hypothesis. One-tailed test needs smaller differences between means in comparison to that needed for a two-tailed test.

Time is noted for both the cases of read and spontaneous speech, and the samples are from the same speaker. But if it is required to have a related groups test instead of an independent group, the *t*-statistic is calculated as:

$$t = \frac{\text{Mean of condition}_1 - \text{Mean of condition}_2}{S.E.\text{ of differences}}$$

A statistic collected for 20 speakers records a mean tone unit duration of 38.6 centiseconds and the spontaneous speech 33.4 centiseconds, and the standard deviation of the difference between the means is 2.65, then *t* value is 1.96 [=(38.6 – 33.4)/2.65]. This *t* value is used to establish whether two sample means differing so much, might have come from the same (null hypothesis) or different (alternate hypothesis) distributions.

Decision rules are formulated to assess a level of support for the alternate hypothesis. Basically, this involves stipulations that assume the samples from the same distribution. But if the probability of the means differ so much, then one may think of an alternative conclusion that the samples are drawn from different populations.

Such a stipulation is done at discrete probability levels. If there is a less than 5 per cent chance of samples belonging to the same distribution, then the hypothesis that the samples were drawn from different distributions is supported as alternative hypothesis at that level of significance. If there is a chance of more than 5 per cent, in which case the samples are drawn from the same distribution, then the null hypothesis is supported. In the worked example, with 19 degrees of freedom, a *t* value of 1.96 does not lead to the conclusion that samples are drawn from different populations; thus, the null hypothesis is accepted.

Test Statistic

These are quantified parameters calculated from sample of data, which is used to decide the acceptance or rejection of the null hypothesis.

Test statistic of a hypothesis test is given by:

$$Z = \frac{Y - \mu_0}{\sigma_7} = \frac{Y - \mu_0}{\sigma / \sqrt{n}}$$ where symbols have their usual meaning.

Y stands for mean, μ_0 for claim level = H_0, σ = Standard deviation. σ^2 is variance and *Z* gives a test statistic.

Critical Value(s)

This is defined as a threshold value which is used to decide the criteria of accepting or rejecting null hypothesis. This depends on the significance level of the test.

Significance Level

This is given as a fixed probability of wrongfully rejecting the null hypothesis H_0 and is the probability of a type I error. This decision is taken by a person or agency carrying out the investigation.

Critical Region

This region is defined as a set of test statistic leading to rejection of the null hypothesis in a hypothesis test. For this, the sample space is split into two mutually exclusive regions. This region provides basis to reject the null hypothesis H_0.

P-Value

This is a probability value and known as p-value. This is the probability of getting extreme value of the test statistic in comparison to that observed by chance alone, provided the null hypothesis H_0 is true. It is the probability of wrongly rejecting the null hypothesis.

Power

The power of a statistical hypothesis test measures the ability of such a test to reject the null hypothesis when it is actually false. Thus, it is power to make a correct decision. This can be retold as the power of not committing a type II error. It is given by subtracting the probability of a Type II error from 1. Mathematically, it is given as:
Power = 1 – P (type II error) = $(1-\beta)$?

The maximum power a test can have is 1 and the minimum is 0. Ideally, we want a test to have high power, close to 1. Normally 0.8 is considered as good for correct decision-making.

Power of a statistical Test

A statistical hypothesis test is a test in which a hypothesis is tested. It analyses gathered data to decide on hypothesis. In decision-making, such analysis is highly desired. It calculates values of some key variables and compares with a critical value for which hypothesis is assumed to be true. In case value of such variable(s) are far away from the critical value, it rejects the hypothesis.

Hypothesis Tests

Two hypotheses are made; one is 'null hypothesis' the other is called 'alternative hypothesis'. A formal process is followed defining some standards and then it is decided whether to accept the hypothesis or reject it. This is known as hypothesis testing which has four steps:

1. **Statements:** Make two statements which are hypotheses. These are: 'null hypothesis' and 'alternative hypotheses'. Null hypothesis is denoted by H_0 and alternative hypothesis by H_a. H_0 and H_a are mutually exclusive. If H_0 is true, then H_a must be false and vice-versa.

2. **Deciding Strategy:** Set formula to compute salient parameters and chalk out analysis plan describes the use of sample data and compute the critical parameters to decide whether to accept or reject the null hypothesis.
3. **Sample Data Analysis:** Values of critical parameters like mean, proportion, *t*-score, Z-score and other parameters are decided by the researcher.
4. **Interpretation on Results:** Based on decision rules, null hypothesis is accepted or rejected.

In any test, error can take place and this is also true for statistical tests. Statistic defines two types of errors: type I and type II:

- **Type I Error:** A Type I error is defined as rejection of a null hypothesis when it is true. It signifies wrong decision. Probability of a Type I error is known as **significance level**, named alpha and denoted by Greek letter α.
- **Type II Error:** A Type II error is acceptance of null hypothesis when it is false. This too signifies wrong decision. Probability of a Type II error is known as **Beta** and is denoted by Greek letter β. Complement of this probability is power of the test.

The ***critical region*** is defined as a set of all outcomes of a hypothesis test that leads to the rejection of a null hypothesis and acceptance of an alternative hypothesis accepted, and is denoted by C.

Power of a Test

A test is more powerful if it is able to worked out a criteria that directs to a clear decision. Probability of producing significant difference for such decisions is known as the power of test. This difference must be found at a significance level decided by the researcher to asses the power of the test.

This probability signifies the power of making correct decision. It is complement of probability of occurrence of Type II error, β (Beta). Thus, power of a test = $(1-\beta)$.

Power of a statistical test is affected by differences between the sample size and the specified significance level. A power of 1 is ideal, but in practice, 0.80 or more is taken as good value to decide departure from the null hypothesis.

A significance criterion is a statement of unlikelihood of a result. Here the null hypothesis is considered significant. Commonly used criteria take probabilities as 0.05, 0.01 and 0.001. The power of a test is increased by weakening the significance level, putting this under a narrow limit. This increases the chance of obtaining a statistically significant result by rejecting the null hypothesis correctly and, thus, taking a correct decision. But it increases the risk of a Type I error.

There is no formal standard for power. In practice, a power of 0.80 or more is considered good to detect a reasonable departure from the null hypothesis.

Size/Significance Level of a Test (α)

Significance level in simple hypothesis test is the probability of *incorrectly* rejecting the null hypothesis. In a composite hypothesis, it is the upper bound of the probability that serves the basis of rejecting the null hypothesis.

The greatest power for a given *significance level* is known as **most powerful test.**

A test that has greatest *power* for all values of the parameter under test is **Uniformly Most Powerful (UMP) test.**

When power of test is near unity, the test is consistent. This is termed as 'consistent test'.

Unbiased test for a H_a in which the probability of rejection of H_0 > significance level for H_a to be true and d^0 the significance level when H_0 is true.

Uniformly Most Powerful Unbiased (UMPU) test is a UMP in the set of all unbiased tests.

NOTES

4.10 SPSS OR STATISTICAL PACKAGE FOR SOCIAL SCIENCE

Amongst the student community as well as with most research agencies, SPSS is the most widely used package. It is adaptable to most business problems and is extremely user friendly. A reference URL for SPSS is http://www.spss.com/.

There are a number of specific software programs like E Views for business forecasting and LISREL (Linear Structural Relations) for structural equation modelling. However, for most purposes, SPSS is the most widely used software.

SPSS

SPSS is abbreviated term for Statistical Package for Social Science, and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies. The statistical package SPSS is used to perform quantitative research in social science because it is easy to use. The SPSS Data Editor is very valuable and is specifically designed for performing statistical tests, such as correlation, regression, t-test, hypotheses, chi-square and Analysis of Variance or ANOVA. It also helps a researcher to make useful data entries, find frequency counts, sort and rearrange data and so on.

The SPSS features available with the software package can be accessed with the help of pull-down menus or can be programmed using a licensed 4GL (Fourth Generation Language) command syntax language. The advantage of command syntax programming language is that it helps in data reproducibility, simplifying repetitive tasks, performing complex data manipulations and analyses. In addition, the user can program specific syntax for some complex applications which are not available in the predefined menu structure. The *command syntax* can also be generated by pull-down menu interface and can be displayed in the output. *To syntax* can be made visible to the user by changing the default settings. It can also be pasted into a syntax file with the help of 'paste' button which is available in each menu.

SPSS can read and write data from ASCII (American Standard Code for Information Interchange) text files including hierarchical files, other statistics packages, spreadsheets and databases. SPSS can also be used to read and write to external relational database tables using ODBC (Open Database Connectivity) and SQL (Sequential Query Language). Statistical output is in the licensed file format with the file extension name as **.spv** which supports pivot tables. The output can be exported to Microsoft Word and can be acquired as data, as text, PDF, XLS, HTML, XML, SPSS dataset or in the graphic image formats (JPEG, PNG, BMP and EMF).

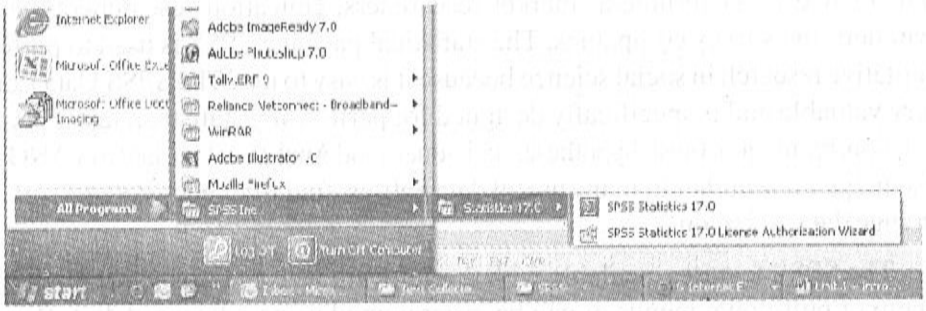
Check Your Progress

16. What are the two variables of formalized hypothesis?
17. What do you mean by statistical decision-making?

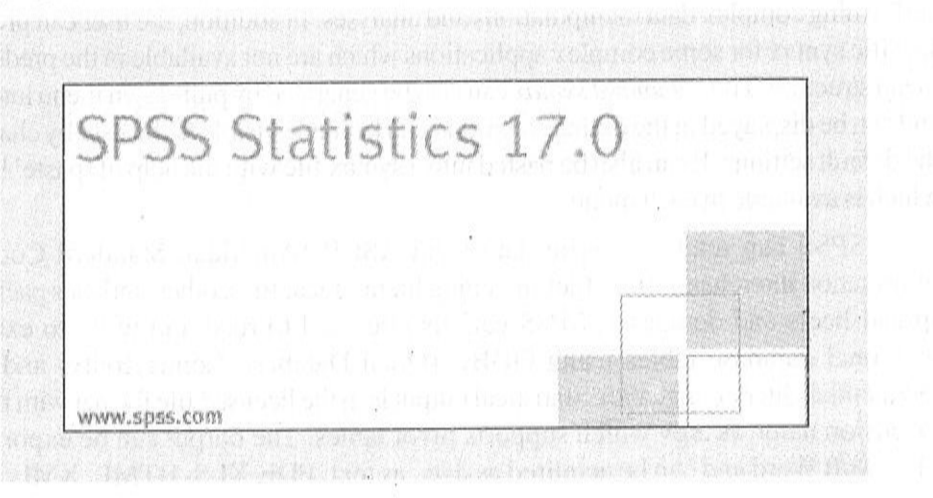
The SPSS is based on Graphical User Interface (GUI) which supports the two data editor views, the Data View and the Variable View. The user can toggle between the two views just by selecting one of the two tabs that appear in the bottom left of the SPSS window and clicking on it. The 'Data View' exhibits a view in the form of a spreadsheet as the cases (rows) and variables (columns). Only two data types can be defined in SPSS Statistics, i.e., the numeric data type and the text or 'string' data type. All data processing processes appears in sequence case-by-case through the file. You can match the files on the basis of one-to-one and one-to-many, but not many-to-many. In SPSS, the data cells simply hold numbers or text. You cannot store the formulas in these cells. The 'Variable View' exhibits the metadata dictionary in which each row represents a variable to display the variable name, variable label, value label(s), print width, measurement type and other associated characteristics. In both views, you can manually edit the cells, define file structure and do data entry without using the command syntax for smaller datasets. Large datasets, such as statistical surveys, are created using data entry software or entered by scanning using Optical Character Recognition (OCR) and Optical Mark Recognition (OMR) software. Using a 'macro' language, command language subroutines can be written. A Python programmability extension is used to access the information in the data dictionary and dynamically build command syntax programs.

Working With SPSS

Install SPSS on your computer and create a **shortcut** menu on your desktop and directly start the package by clicking on the SPSS icon. Alternatively, you can go to the **Start → All Programs → SPSS Inc → Statistics 17.0 → SPSS Statistics 17.0** as shown:

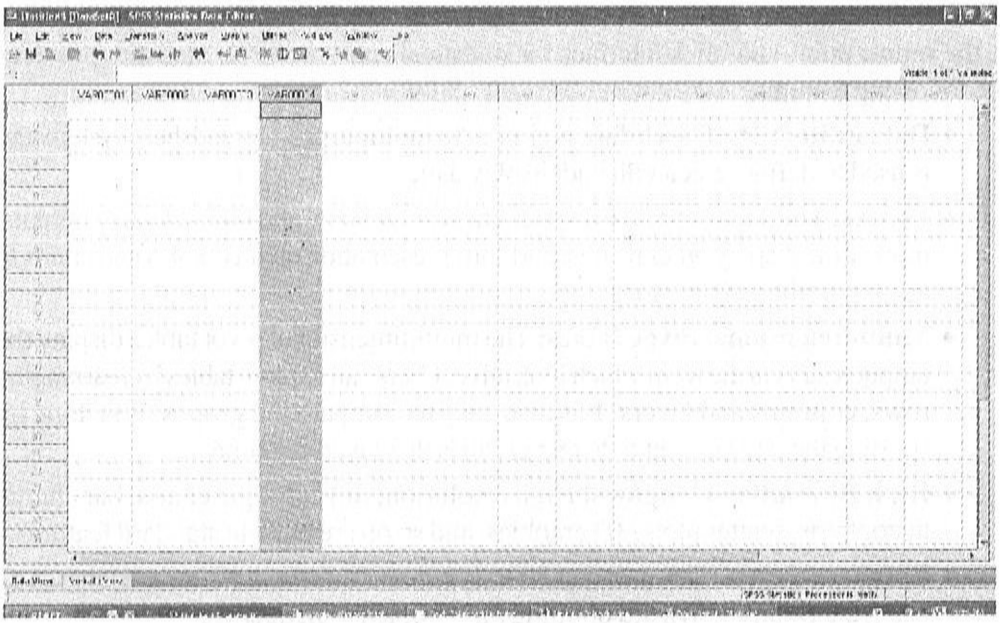


When you click on the SPSS Statistics 17.0, the following screen will appear to start SPSS program:

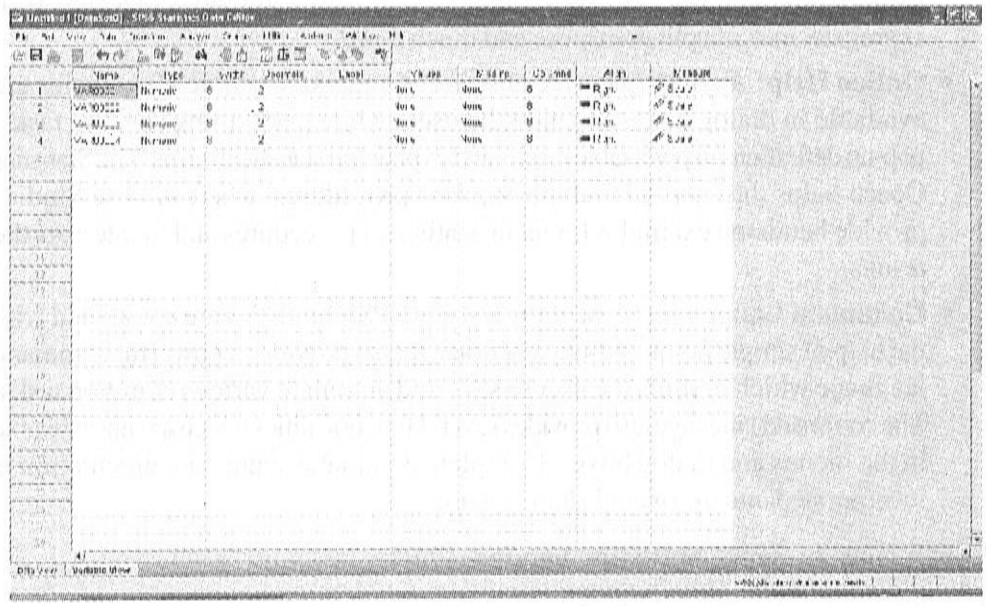


As discussed earlier, the SPSS Data Editor has two main 'views'. The user can enter data in the Data View, while in the Variable View, the user can select the name, type, maximum number of letters per cell ('width'), number of decimal points, label, width of cell ('columns'), alignment within the cell ('align') and whether or not the variable is nominal, ordinal or 'scale' ('measure'). The user can also categorize entries into labels (in the 'Values' column) and mark entries as invalid (in the 'Missing' column) in the Variable View. Using SPSS for Windows, the user can perform almost any statistics calculation in combination with pointing and clicking on the menus, and various specific interactive dialog boxes. Both the SPSS views are shown as follows:

Data View


Untitled1 [DataSet0] - SPSS Statistics Data Editor
VAR00001
VAR00002
VAR00003
VAR00004
Data View
Variable View
SPSS Statistics Processor is ready

Variable View


Untitled1 [DataSet0] - SPSS Statistics Data Editor
Data View
Variable View
SPSS Statistics Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	VAR00001	Numeric	8	2		None	None	8	Right	Scale
2	VAR00002	Numeric	8	2		None	None	8	Right	Scale
3	VAR00003	Numeric	8	2		None	None	8	Right	Scale
4	VAR00004	Numeric	8	2		None	None	8	Right	Scale

SPSS Statistics 17.0

SPSS Statistics 17.0 is a comprehensive system for analysing data based on the Graphical User Interface (GUI). SPSS Statistics can acquire data from almost any form of file, and use them to create tabulated reports, charts, plots of distributions and trends, descriptive statistics and complex statistical analyses. SPSS Statistics Base 17.0 provides examples in the Help system which is automatically installed with the software. In addition, below the menus and dialog boxes, SPSS Statistics uses a command language for data analysis.

SPSS Statistics has a powerful statistical analysis and data management system in a graphical environment. It also has descriptive menus and simple dialog boxes which help the users to accomplish the task just by pointing and clicking the mouse. In addition to the simple point-and-click interface for statistical analysis, SPSS Statistics provides the following features:

- **Data Editor:** The Data Editor is similar to multipurpose spreadsheet system and is used to define, enter, edit and display data.
- **Viewer:** The Viewer helps to browse the results, show and hide selective outputs, modify the display order results, and shift presentation quality tables and charts to and from other applications.
- **Multidimensional Pivot Tables:** The multidimensional pivot tables display the output results in the form which look alive. Users can explore tables by rearranging rows, columns and layers. It is also easy to compare the groups. It is done by splitting the table so that only one group is displayed at a time.
- **High Resolution Graphics:** High resolution, full color pie charts, bar charts, histograms, scatter plots, 3D graphics, and so on are built-in standard features.
- **Database Access:** The user can directly recover information from databases by using the Database Wizard omitting the complex SQL queries.
- **Data Transformations:** Transformation features help to find the data organized for analysis. You can also subset data and files to combine categories, add, aggregate, merge, split, transpose and much more.
- **Online Help:** A comprehensive abstract of context sensitive Help topics are available in dialog boxes to guide the users while performing specific tasks, pop-up definitions in pivot table results and explaining statistical terms. The Statistics Coach helps the users to find the required procedures, whereas Case Studies provide hands-on examples for using statistical procedures and to interpret the results.
- **Command Language:** Most of the tasks in SPSS Statistics are completed with the help of simple point-and-click actions. It also provides a powerful command language which permits the user to save and automate various common tasks. The command language also provides several functionalities which are not available in the menus and dialog boxes. Complete command syntax documentation is incorporated into the overall Help system.

What's New in SPSS Statistics Version 17.0?

The following are the enhanced features available in the new version of SPSS Statistics 17.0:

NOTES

New Syntax Editor: The syntax editor in SPSS Statistics 17.0 has been entirely redesigned including features such as auto completion, color coding, bookmarks and breakpoints. Auto completion feature provides the user a list of valid command names, subcommands and keywords, hence, the user spends less time referring to syntax charts. Color coding feature permits the user to spot unrecognized terms and some common syntactical errors quickly. Bookmarks feature permits the user to speedily navigate huge command syntax files. Breakpoints feature permits the user to stop execution at specific points for inspecting data or output prior to proceeding.

- **Custom Dialog Builder:** The Custom Dialog Builder permits the user to create and manage custom dialogs for creating command syntax.
- **Multiple Language Support:** In addition to the capability to modify the output language, the user can now modify the user interface language.
- **Codebook:** The Codebook method accounts the dictionary information, such as variable names, variable labels, value labels, missing values and summary statistics for all or specific variables and manifold response sets in the active dataset. For nominal and ordinal variables, and manifold response sets, summary statistics include counts and percents. For scale variables, summary statistics include mean, standard deviation and quartiles.
- **Nearest Neighbor Analysis:** Nearest Neighbor analysis is a technique for categorizing cases based on their similarity to other cases. In machine learning, it was used to identify patterns of data without involving an accurate match to any stored patterns or cases. Similar cases are close to each other and dissimilar cases are isolated from each other. Thus, the distance between the two cases is a measure of their dissimilarity.
- **Multiple Imputation:** The Multiple Imputation method executes multiple imputation of missing data values. The dataset having missing values give outputs as one or more datasets in which missing values are substituted with plausible estimates. The pooled results are obtained when other procedures are run. This technique also summarizes missing values in the working dataset. This feature is available in the Missing Values add-on option.
- **RFM Analysis:** RFM analysis is the abbreviated form of Recency, Frequency, and Monetary analysis. This method is used to recognize existing customers who are most probable to respond to a new offer and is frequently used in direct marketing. This feature is available in the EZ RFM add-on option. The fundamental principle of RFM analysis is for customers who have purchased recently, have made more purchases and are more likely to respond to your offering than other customers who have purchased less recently, less often and in smaller amounts.

  Basically, RFM analysis uses information about customers' past behaviour that is *Recency* is how long ago the customer last made a purchase. *Frequency* is how many purchases the customer has made (sometimes within a specified time period, such as average number of purchases per year). *Monetary* is the total amount spent by the customer (sometimes within a specified time period).
- **Categorical Regression Enhancements:** Categorical Regression has been enhanced and included regularization and resampling techniques for accurately assessing and improving predictions. Jointly, these new methods feasibly create state-of-the-art models even for high volume data where there are more variables than observations. This feature is available in the Categories add-on option.

NOTES

- **Graphboard:** Graphboard are visualizations which include graphs, charts and plots created using a visualization template. SPSS Statistics 17.0 provides built-in new visualization templates which are effectively custom visualization types.
- **Exporting Output:** The following output export format options and control over. exported contents are available in SPSS Statistics version 17.0:
  - o To wrap or shrink wide table in Word documents.
  - o To create new worksheets or append data to existing worksheets in an Excel workbook.
  - o To save output export specifications in the form of command syntax with the OUTPUT EXPORT command. All the features for exporting output in the Export Output dialog are available in command syntax.
  - o The Output Management System (OMS) supports the additional output formats, such as Word, Excel and PDF.
- **Shift Values:** Shift Values generates new variables that hold the values of existing variables from preceding or subsequent cases.
- **Aggregate Enhancements:** This feature allows the user to use the aggregate method without specifying a break variable.
- **Median Function:** A median function is available for computing the median value across selected variables for each case.

Windows in SPSS Statistics

The following are the different types of windows in SPSS Statistics:

- **Data Editor:** The Data Editor displays the contents of the data file. You can create new data files or modify existing data files using the Data Editor. When you open more than one data file, then there is a separate Data Editor window for each opened data file.
- **Viewer:** The Viewer displays all statistical results, tables and charts. The user can edit the output and save it for later use. A Viewer window opens automatically the first time the user runs a procedure to generate output.
- **Pivot Table Editor:** The Pivot Table Editor modifies the output in various ways that is displayed in pivot tables. The user can edit text, swap data in rows and columns, add color, create multidimensional tables, and hide and show selective results.
- **Chart Editor:** The high resolution charts and plots can be modified in chart windows. The user can change the colors, select different font types or sizes, switch the horizontal and vertical axes, rotate 3D scatter plots, and even change the chart type.
- **Text Output Editor:** Text output which is not displayed in pivot tables can be modified using the Text Output Editor. The user can edit the output and modify font characteristics, such as type, style, color and size.
- **Syntax Editor:** The user can paste the dialog box choices into a syntax window, where the selections appear in the form of command syntax. Now the user can edit the command syntax to use special features that are not available through dialog boxes. The user can also save these commands in a file for use in subsequent sessions.

Check Your Progress

18. State the advantage of command syntax programming language.
19. What is the function of SPSS Statistics 17.0?

4.11 CHI-SQUARE TEST

NOTES

For the use of a chi-square test, data is required in the form of frequencies. The data expressed in percentages or proportion can also be used, provided it could be converted into frequencies. The majority of the applications of chi-squarc (χ^2) are with the discrete data. The test could also be applied to continuous data, provided it is reduced to certain categories and tabulated in such a way that the chi-square may be applied.

Some of the important properties of the chi-square distribution are as follows:

- Unlike the normal and t distribution, the chi-square distribution is not symmetric (see Figure 4.6).

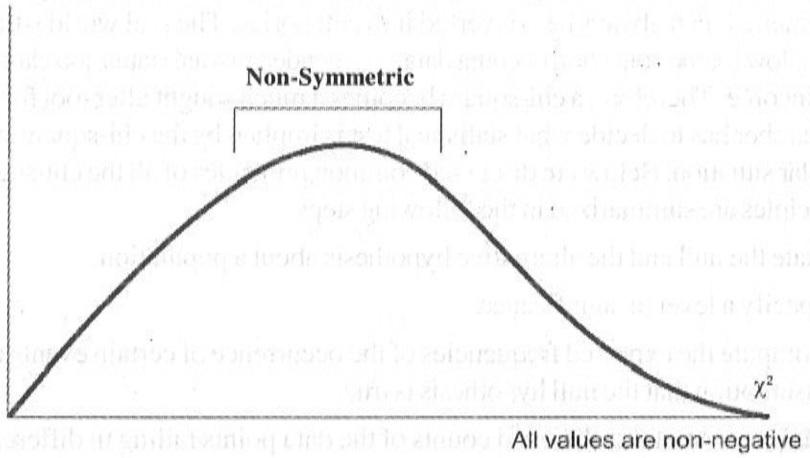


Fig. 4.6 Shape of Chi-Square (χ^2) Distribution

- The values of a chi-square are greater than or equal to zero.
- The shape of a chi-square distribution depends upon the degrees of freedom. With the increase in degrees of freedom, the distribution tends to normal (see Figure 4.7).

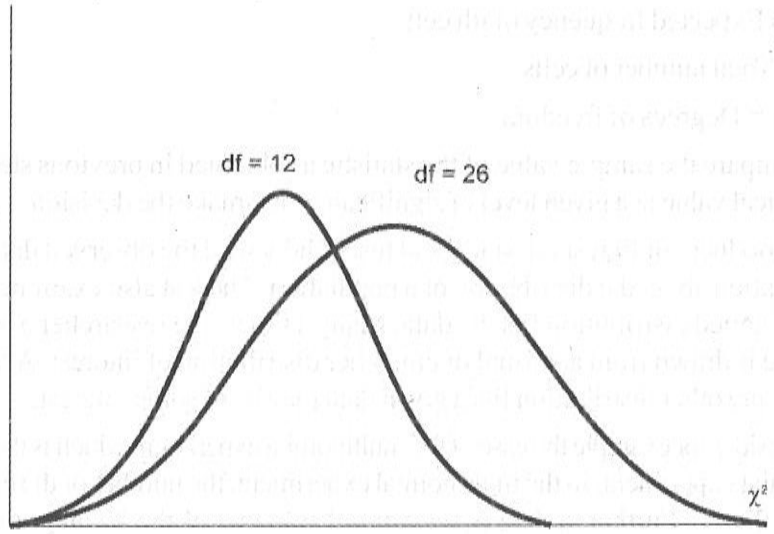


Fig. 4.7 Shape of Chi-Square Distribution with Varying Degrees of Freedom

NOTES

Application of Chi-Square

There are many applications of a chi-square test. Some of them are explained below:

- A chi-square test for the goodness of fit
- A chi-square test for the independence of variables
- A chi-square test for the equality of more than two population proportions

Each of the above-mentioned applications are discussed in the following sections.

1. A Chi-Square Test for the Goodness of Fit

The data in chi-square tests is often in terms of counts or frequencies. The actual survey data may be on a nominal or higher scale of measurement. If it is on a higher scale of measurement, it can always be converted into categories. The real world situations in business allow for the collection of count data, e.g., gender, marital status, job classification, age and income. Therefore, a chi-square becomes a much sought after tool for analysis. The researcher has to decide what statistical test is implied by the chi-square statistic in a particular situation. Below are discussed common principles of all the chi-square tests. The principles are summarized in the following steps:

- State the null and the alternative hypothesis about a population.
- Specify a level of significance.
- Compute the expected frequencies of the occurrence of certain events under the assumption that the null hypothesis is true.
- Make a note of the observed counts of the data points falling in different cells.
- Compute the chi-square value given by the formula.

  $$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  Where,

  O_i = Observed frequency of ith cell

  E_i = Expected frequency of ith cell

  k = Total number of cells

  k–1 = Degrees of freedom

- Compare the sample value of the statistic as obtained in previous step with the critical value at a given level of significance and make the decision.

A goodness of fit test is a statistical test of how well the observed data supports the assumption about the distribution of a population. The test also examines that how well an assumed distribution fits the data. Many a times, the researcher assumes that the sample is drawn from a normal or any other distribution of interest. A test of how normal or any other distribution fits a given data may be of some interest.

Consider for example the case of the multinomial experiment which is the extension of a binomial experiment. In the multinomial experiment, the number of the categories k is greater than 2. Further, a data point can fall into one of the k categories and the probability of the data point falling in the ith category is a constant and is denoted by p_i where i = 1, 2, 3, 4, ..., k. In summary, a multinomial experiment has the following features:

- There are fixed number of trials.
- The trials are statistically independent.
- All the possible outcomes of a trial get classified into one of the several categories.
- The probabilities for the different categories remain constant for each trial.

Consider as an example that a respondent can fall into any one of the four non-overlapping income categories. Let the probabilities that the respondent will fall into any of the four groups may be denoted by the four parameters, p_1, p_2, p_3 and p_4. Given these, the multinomial distribution with these parameters and n the number of people in a random sample specifies the probabilities of any combination of the cell counts.

Given such a situation, we may use a multinomial distribution to test how well the data fits the assumption of k probability p_1, p_2, ..., p_k of falling into the k cells. The hypothesis to be tested is:

H_0: Probabilities of the occurrence of events E_1, E_2, ..., E_k are given by the specified probabilities p_1, p_2, ..., p_k.

H_1: Probabilities of the k events are not the p_i stated in the null hypothesis.

Such hypothesis could be tested using the chi-square statistics. Below are given a set of illustrated examples.

Example 4.18: The manager of ABC ice-cream parlour has to take a decision regarding how much of each flavour of ice-cream he should stock so that the demands of the customers are satisfied. The ice-cream suppliers claim that among the four most popular flavours, 62 per cent customers prefer vanilla, 18 per cent chocolate, 12 per cent strawberry and 8 per cent mango. A random sample of 200 customers produces the results below. At the $\alpha = 0.05$ significance level, test the claim that the percentages given by the suppliers are correct.

Flavour	Vanilla	Chocolate	Strawberry	Mango
Number preferring	120	40	18	22

Solution:

Let

p_v : Proportion of customers preferring vanilla flavour

p_c : Proportion of customers preferring chocolate flavour

p_s : Proportion of customers preferring strawberry flavour

p_m : Proportion of customers preferring mango flavour

H_0 : $p_v = 0.62, p_c = 0.18, p_s = 0.12, p_m = 0.08$

H_1 : Proportions are not that specified in the null hypothesis

The expected frequencies corresponding to the various flavours under the assumption that the null hypothesis is true are as follows:

Vanilla $= 200 \times 0.62 = 124$

Chocolate $= 200 \times 0.18 = 36$

Strawberry $= 200 \times 0.12 = 24$

Mango $= 200 \times 0.08 = 16$

The computations for χ_3^2 are as under: $\sum_{i=1}^{k}\frac{(O_i-E_i)^2}{E_i}$

Flavour	O (Observed Frequencies)	E (Expected Frequencies)	O – E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
Vanilla	120	124	– 4	16	0.129
Chocolate	40	36	4	16	0.444
Strawberry	18	24	– 6	36	1.500
Mango	22	16	6	36	2.250
				Total	**4.323**

The computed value of chi-square is 4.323.

Table χ_3^2 (5 per cent) = 9.488

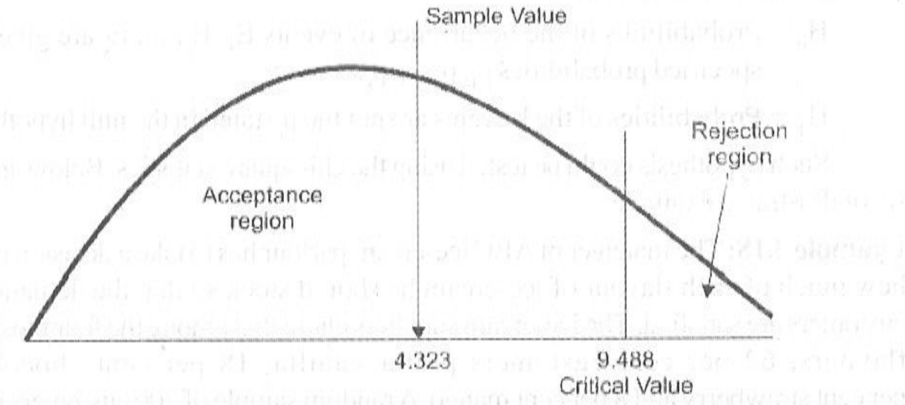


Fig. 4.8 Rejection Region for Example 4.18

As sample χ^2 lies in the acceptance region, accept H_0. Therefore, the customer preference rates are as stated. Using the p value approach, we find that the sample χ^2 value lies as shown below:

χ^2 with 3 d.f.	11.345	7.815	6.251	↓
Level of significance	1 per cent	5 per cent	10 per cent	4.323 (sample χ^2)

It is seen that the sample χ^2 corresponds to a p value greater than 10 per cent. Therefore, there is not enough evidence to reject the null hypothesis. This means that the customer preference rates are as stated in the null hypothesis.

It may be worth pointing out that for the application of a chi-square test, the expected frequency in each cell should be at least 5.0. In case, it is found that one or more cells have the expected frequency less than 5, one could still carry out the chi-square analysis by combining them into meaningful cells so that the expected number has a total of at least 5. Another point worth mentioning is that the degree of freedom, usually denoted by *df* in such cases, is given by k – 1, where k denotes the number of cells (categories).

It may be noted that in Example 4.18, the hypothesized probabilities were not equal. There are situations where the hypothesized probabilities in each category are equal or in other words, the interest is in investigating the uniformity of the distribution. The following example would illustrate it.

A Chi-Square Test for Independence of Variables

The chi-square test can be used to test the independence of two variables each having at least two categories. The test makes a use of contingency tables, also referred to as cross-tabs, with the cells corresponding to a cross classification of attributes or events. Layout of contingency table given below.

Second Classification Category	**First Classification Category**				
	1	**2**	**3**	**4**	**Total**
1	O_{11}	O_{12}	O_{13}	O_{14}	R_1
2	O_{21}	O_{22}	O_{23}	O_{24}	R_2
3	O_{31}	O_{32}	O_{33}	O_{34}	R_3
Total	C_1	C_2	C_3	C_4	n

Assuming that there are r rows and c columns, the count in the cell corresponding to the ith row and the jth column is denoted by O_{ij}, where i = 1, 2, ..., r and j = 1, 2, ..., c. The total for row i is denoted by R_i, whereas that corresponding to column j is denoted by C_j. The total sample size is given by n, which is also the sum of all the r row totals or the sum of all the c column totals.

The hypothesis test for independence is:

H_0 : Row and column variables are independent of each other.

H_1 : Row and column variables are not independent.

The hypothesis is tested using a chi-square test statistic for independence given by:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

The degrees of freedom for the chi-square statistic are given by (r – 1) (c – 1).

For a given level of significance α, the sample value of the chi-square is compared with the critical value for the degree of freedom (r – 1) (c – 1) to make a decision.

The expected frequency in the cell corresponding to the ith row and the jth column is given by:

$$E_{ij} = \frac{R_i \times C_j}{n}$$

Where, R_i = Total for the ith row

C_j = Total for the jth column

n = Total sample size

Let us consider a few examples:

Example 4.19: The following table gives the number of good and defective parts produced by each of the three shifts in a factory:

Shift	**Good**	**Defective**	**Total**
Day	900	130	1030
Evening	700	170	870
Night	400	200	600
Total	2000	500	2500

Is there any association between the shift and the equality of the parts produced? Use a 0.05 level of significance.

Solution:

NOTES

H_0 : There is no association between the shift and the quality of parts produced.

H_1 : There is an association between the shift and quality of parts.

The computations of the expected frequencies corresponding to the ith row and the jth column of the contingency table are shown below: (i = 1, 2, 3) and (j = 1, 2).

$$E_{1,1} = \frac{1030 \times 2000}{2500} = 824$$

$$E_{1,2} = \frac{1030 \times 500}{2500} = 206$$

$$E_{2,1} = \frac{870 \times 2000}{2500} = 696$$

$$E_{2,2} = \frac{870 \times 500}{2500} = 174$$

$$E_{3,1} = \frac{600 \times 2000}{2500} = 480$$

$$E_{3,2} = \frac{600 \times 500}{2500} = 120$$

The table of the observed and expected frequencies corresponding to the ith row and the jth column and the computation of the chi-square is given below:

Row, Column	O_{ij}	E_{ij}	$(O_{ij} - E_{ij})^2$	$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$
1,1	900	824	5776	7.010
1,2	130	206	5776	28.039
2,1	700	696	16	0.023
2,2	170	174	16	0.092
3,1	400	480	6400	13.333
3,2	200'	120	6400	53.333
			Total	**101.83**

The sample chi-square is $\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 101.83$

The critical value of the chi-square with 2 degrees of freedom at 5 per cent level of significance is given by 5.991. The null hypothesis is rejected as the sample chi-square lies in the rejection region, as shown in the Figure 4.9. Therefore, the quality of parts produced is related to the shifts in which they were produced.

NOTES

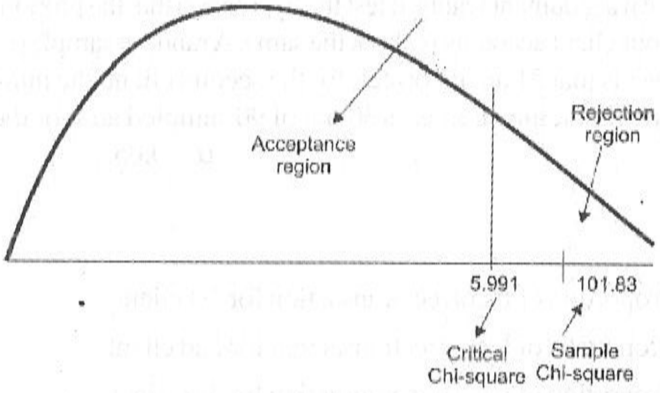


Fig. 4.9 Rejection Region for Example 4.19

Using a p value approach, the same decision would be arrived at. It is left for the readers to show it.

It may be worth mentioning again that for the application of a chi-square test of independence, the sample should be selected at random and the expected frequency in each cell should be at least 5.

Chi-Square Test for the Equality of More Than Two Population Proportions

In certain situations, the researchers may be interested to test whether the proportion of a particular characteristic is the same in several populations. The interest may lie in finding out whether the proportion of people liking a movie is the same for the three age groups, i.e., 25 and under, over 25 and under 50, and 50 and over. To take another example, the interest may be in determining whether in an organization, the proportion of the satisfied employees in four categories—class I, class II, class III and class IV employees—is the same. In a sense, the question of whether the proportions are equal is a question of whether the three age populations of different categories are homogeneous with respect to the characteristics being studied. Therefore, the tests for equality of proportions across several populations are also called tests of homogeneity.

The analysis is carried out exactly in the same way as was done for the other two cases. The formula for a chi-square analysis remains the same. However, two important assumptions here are different:

(i) We identify our population (e.g., age groups or various class employees) and the sample directly from these populations.

(ii) As we identify the populations of interest and the sample from them directly, the sizes of the sample from different populations of interest are fixed. This is also called a chi-square analysis with fixed marginal totals. The hypothesis to be tested is as under:

H_0 : The proportion of people satisfying a particular characteristic is the same in population.

H_1 : The proportion of people satisfying a particular characteristic is not the same in all populations.

The expected frequency for each cell could also be obtained by using the formula as explained early. There is an alternative way of computing the same, which would give identical results. This is shown in the following example:

Example 4.20: An accountant wants to test the hypothesis that the proportion of incorrect transactions at four client accounts is about the same. A random sample of 80 transactions of one client reveals that 21 are incorrect; for the second client, the number is 25 out of 100; for the third client, the number is 30 out of 90 sampled and for the fourth, 40 are [illegible] $\alpha = 0.05$.

Solution:

Let

p_1 = Proportion of incorrect transaction for 1st client

p_2 = Proportion of incorrect transaction for 2nd client

p_3 = Proportion of incorrect transaction for 3rd client

p_4 = Proportion of incorrect transaction for 4th client

Let

$H_0 : p_1 = p_2 = p_3 = p_4$

H_1 : All proportions are not the same.

The observed data in the problem can be rewritten as:

Transactions	Client 1	Client 2	Client 3	Client 4	Total
Incorrect transactions	21	25	30	40	116
Correct transactions	59	75	60	70	264
Total	80	100	90	110	380

An estimate of the combined proportion of the incorrect transactions under the assumption that the null hypothesis is true:

$$p = \frac{21+25+30+40}{80+100+90+110} = \frac{116}{380} = 0.305$$

q = combined proportion of the correct transaction

$= 1 - p = 1 - 0.305 = 0.695$

Using the above, the expected frequencies corresponding to the various cells are computed as shown below:

Transactions	Client 1	Client 2	Client 3	Client 4	Total
Incorrect transactions	80 × 0.305 = 24.4	100 × 0.305 = 30.5	90 × 0.305 = 27.45	110 × 0.305 = 33.55	115.9
Correct transactions	80 × 0.695 = 55.6	100 × 0.695 = 69.5	90 × 0.695 = 62.55	110 × 0.695 = 76.45	264.1
Total	80	100	90	110	380

In fact, the sum of each row/column in both the observed and expected frequency tables should be the same. Here, a bit of discrepancy is found because of the rounding of the error. It can be easily verified that the expected frequencies in each cell would be the same using the formula $E_{ij} = \frac{R_i \times C_j}{n}$ as already explained. Now the value of the chi-square statistic can be calculated as:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{4}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \frac{(21-24.4)^2}{24.4} + \frac{(25-30.5)^2}{30.5} + \frac{(30-27.45)^2}{27.45} + \frac{(40-33.55)^2}{33.55}$$

$$+\frac{(59-55.6)^2}{55.6}+\frac{(75-69.5)^2}{69.5}+\frac{(60-62.55)^2}{62.55}+\frac{(70-76.45)^2}{76.45}$$

$$= 0.474+0.992+0.237+1.240+0.208+0.435+0.104+0.544$$

$$= 4.234$$

NOTES

Degrees of freedom (df) = (2 – 1) × (4 – 1) = 3

The critical value of the chi-square with 3 degrees of freedom at 5 per cent level of significance equals 7.815. Since the sample value of χ^2 is less than the critical value, there is not enough evidence to reject the null hypothesis. Therefore, the null hypothesis is accepted. Therefore, there is no significant difference in the proportion of incorrect transaction for the four clients.

4.12 REPORT WRITING

Since research is a demanding and painstaking effort, sufficient time and attention should be devoted to writing the report of the work carried out.

4.12.1 Importance of Report Writing

On completion of the research study and after obtaining the research results, the real skill of the researcher lies in terms of analysing and interpreting the findings and linking them with the propositions formulated in the form of research hypotheses at the beginning of the study. The statistical or qualitative summary of results would be little more than numbers or conclusions unless one is able to present the documented version of the research endeavour.

Depending on the business researcher's orientation, the intention might be different and would be reflected in the form of the presentation but the significance is critical to both. Essentially, this is so because of the following reasons:

- The research report fulfils the historical task of serving as a concrete proof of the study that was undertaken. This serves the purpose of providing a framework for any work that can be conducted in the same or related areas.
- It is the complete detailed report of the research study undertaken by the researcher; thus, it needs to be presented in a comprehensive and objective manner. This is a one-way communication of the researcher's study and analysis to the reader/manager, and, thus, needs to be all-inclusive and yet neutral in its reporting.
- For academic purpose, the recorded document presents a knowledge base on the topic under study, and for the business manager seeking help in taking more informed decisions, the report provides the necessary guidance for taking appropriate action.
- As the report documents all the steps followed and the analysis carried out, it also serves to authenticate the quality of the work carried out and establishes the strength of the findings obtained.

Thus, effective recording and communicating of the results of the study becomes an extremely critical step of the research process. Based on the nature of the research study and the researcher's orientation, the report can take different forms.

Check Your Progress

20. What is a goodness of fit test?
21. State the use of a chi-square test.

Steps Involved in Report Writing

Whatever the type of report, the reporting and dissemination of the study and its findings require a structured format and by and large, the process is standardized. As stated above, the major difference amongst the types of reports is that all the elements that essentially constitute a research report would be present only in a detailed technical report. In the management report, the information on the sampling techniques follows the research intention, and the questionnaire design details need not be reported. The review of past literature would be perfunctory in the management report; however, they would be detailed and accompanied with the bibliography in the technical report. Usage of theoretical and technical jargon would be higher in the technical report and visual presentation of data would be higher in the management report.

The preliminary section of report formulation and presentation includes the rudimentary parts, for example the title page, followed by the letter of authorization, acknowledgements, executive summary and the table of contents. Then come the background section, which includes the problem statement, introduction, study background, scope and objectives of the study, and the review of literature (depends on the purpose). This is followed by the methodology section, which, as stated earlier, is again specific to the technical report. This is followed by the findings section and then come the conclusions. The technical report would have a detailed bibliography at the end.

In the management report, the sequencing of the report might be reversed to suit the needs of the decision-maker, as here the reader needs to review and absorb the findings. Thus, instead of simply summarizing the statistical results, the findings need to be presented in such a way that they can be used directly as inputs for decision-making. Thus, the last section would be presented immediately after the study objectives and a short reporting on methodology could be presented in the appendix.

Thus, the entire research project needs to be recorded either as a single written report or into several reports, depending on the need of the readers. The researcher would need to assist the business manager in deciphering the report, executing the findings, and in case of need, to revise the report to suit the specific actionable requirements of the manager.

Some of the steps involved in the planning of writing a good report are as follows:

- **Step 1:** The entire study should be analysed very logically in a thorough manner.
- **Step 2:** A rough or preliminary draft of the outline of the proposed research should be prepared with thought and care.
- **Step 3:** The final outline should be developed.
- **Step 4:** This can then be polished and refined, before being re-written.
- **Step 5:** This can be followed by the compiling of the bibliography.
- **Step 6:** Then the final draft can be written.

4.12.2 Style and Types of Report Writing

Reports have to be clear, concise, complete and easy to read. The object of reporting a study is to help others who read it to be able to follow the procedure and carry out a similar activity. The report should be clear in description and explanation. The reference works must mention authors by last name only. Abbreviations should be avoided. Only

NOTES

the very well-known abbreviations, such as IQ, can be used. Detailed formulae and computations can be eliminated.

The report has to be written in the past tense. All sources must be acknowledged, correctly and completely. Certain permissible and well-known acronyms alone are allowed to be used in report writing, e.g., STM (Short Term Memory). This can also figure in the midst of a sentence, but not at the beginning. Ideas must be paraphrased and written. As a rule, whenever quotes are used, the relevant study should be cited.

As far as possible, only the accepted terminology should be used. Numbers that come in the report which are from zero to nine should be written in words, while ten or those larger than ten can be written in digits.

Types of Reports

Reports vary in length, format and purpose. The research report presented here is one of a scientific study. Other forms of reports are as follows:

(i) **Business Reports:** These are largely in the form of a letter. They are short, precise and factual in detail.

(ii) **Balance Sheets:** Banks and other financial institutions present their reports in this form. This is largely a statement of accounts for their customers and shareholders.

(iii) **Formulae and Symbols:** Chemists present their reports in terms of formulae and other symbols to explain the preparation.

(iv) **Manuals:** Manufacturing companies prepare reports in the form of manuals of operation, servicing, spares and so on.

(v) **Literature, Linguistics, Philosophy and Other Subjects Reports:** Students of literature, linguistics, philosophy and other subjects write reports that critically analyse a choice subject matter. Here quotations and descriptions are common.

(vi) **Newspapers, Magazines and Other Media Reports**: Newspapers, magazines and other media personnel present items or articles in the form of reports. These include first hand account of events or opinions on a topic. Sometimes interviews of experts in a given field are reported on a theme. This kind of report is largely fact-based, with some details to explain the happenings. Some of these reports have eye-witness accounts also. The implications of these events form the later parts of the report writing.

(vii) **Book-review Reports:** Book-review reports analyse the content, style, format, language and other details of a book, so as to inform the readers about the book. This is often a short report.

(viii) **Government Reports/Reports of Commissions:** Government reports/reports of commissions are very elaborate reports that are comprehensive in nature. These reports are the result of a detailed inquiry carried out on a topic of interest, nationwide or about a specific concern/matter.

(ix) **Case History Reports:** Case history reports submitted by doctors and other healthcare personnel involve in-depth studies of cases. They are based on diagnostic findings and include plans of treatment.

4.12.3 Methods Used in Report Writing

The following are the various methods used in writing research proposal and report:

Research Proposal

In simple terms, a research proposal means a written application that proposes to pursue or conduct a research study. It aims at presenting the idea around which the research study revolves. A research proposal should be able to communicate that the researcher has applied deep thought to the subject of research, and has put considerable effort in collecting the required information, scrutinizing the available data and contemplated a well-organized plan for the research. It tries to emphasize the need for conducting the research, and, thus, necessarily involves formulation of a good research question. The basic components of a research proposal are as follows:

- Title
- Abstract
- Background
- Objective
- Technical approach
- Bibliography

Based on this format, the desired form and features of the contents contained in a research proposal can be enumerated as follows:

- A research proposal starts with a foreword that contains the core question which the researcher aims to answer. Thus, it is written with the purpose of explaining something.
- It also contains a concise review of the prevailing literature related to the researcher's subject. Here, the researcher aims at reviewing the major works related to his/her topic and specify the arguments that have been formulated.
- The research proposal includes a statement regarding the argument or explanation that the researcher aims to present.
- The proposal should also indicate the way in which the researcher's argument is going to be different from the arguments made by other authors. In other words, it should emphasize the aspects in which the argument is unique.
- The proposal should also include a short summary of the different parts of the research.
- A short bibliography containing the important sources being used should also be written in the proposal. This can also mean including databases, websites and interviews.
- The researcher should opt for quality rather than quantity in writing his/her proposal. Thus, a proposal need not be long and an approximately 3–4 pages of research proposal is quite sufficient.

The research proposal is supposed to communicate the researcher's overall effort that is involved in conducting the research. As such, a researcher should take ample care while writing the research proposal. You should keep in mind while writing such a proposal that the ideas involved in the research study need to be presented in a comprehensive and reliable format. The reader should get a clear-cut idea of what the research is all about and what argument it aims to convey. It should also emphasize the researcher's thought process and the depth of his/her knowledge of the concerned subject matter of research.

Research Process

The process of research can be implemented as a series of actions or steps that are essential to be performed in a specific order. These actions or activities usually overlap each other rather than pursuing a specific sequence. A brief description of the steps is given as follows:

- **Selecting the Topic:** The first step of a researcher is to select a topic of research. While doing so, he/she should restrict it to the most potential topic that is open for extensive research out of several alternatives. The factors to be considered for topic selection are as follows:
    - o Relevance of the topic
    - o Scope for research, i.e., the required data should be available and accessible
    - o Contribution to knowledge in the specific field
    - o Required cooperation from the research guide
- **Define the Research Problem:** The research problems can be related to either the state of nature or to the relationship of variables. In defining the research problem, the researcher should study the existing literature including books and journals available in the field with an interdisciplinary perspective to base his/her research topic on some reliable background. He/she should also concentrate on the relevance of the present research with the past works.
- **Mention the Objective of Research:** After selecting the topic and defining the research problem, the researcher should mention the objective of research. This means that he/she should explain what he/she aims to achieve through the research. His/her objective should also include an explanation of the extent to which the research work is related to the specific field.
- **Survey Existing Literature:** To understand the basis of research, it is important for the researcher to review the existing literature. This involves:
    - o Surveying the existing books available in the field
    - o Reviewing other published materials like articles, journals, reports and conference proceedings

  The researcher should then prepare his/her own index for a period, in a chronological order, in addition to his/her consultation of various indices.
- **Determine the Sample Design:** Often, we select only a few items for universal study purposes, for example, blood testing on sample basis to perform census inquiry. The item selected is technically known as a sample. The researcher must decide the way of selecting a sample or decide about the sample design. A sample design is a definite plan determined for data collection to obtain a sample from a given population. The various types of sample designs are as follows:
    - o Deliberate sampling
    - o Simple random sampling
    - o Systematic sampling
    - o Stratified sampling

NOTES

- o Quota Sampling
- o Cluster Sampling
- o Multi-Stage Sampling
- o Sequential Sampling

The researcher should decide the sample design after considering the nature of inquiry and other related factors. Sometimes, several of these methods of sampling are used in the same study, which in turn is called 'mixed sampling'.

- **Data Collection:** There are a variety of ways to collect data. Primary data can be collected through experiments or through surveys. If the researcher performs an experiment, he/she observes some quantitative measurements. This helps him/her to examine the truth in his/her hypothesis. In the case of survey, however, the researcher can adopt one or more of the following ways to collect data:
  - o By observation
  - o Through personal interview
  - o Through telephone interview
  - o By Mailing of questionnaires
  - o Through schedules
- **Execute the Project:** This is the most important step in the research process. The researcher should ensure that the project is performed in a logical way and in time. If a survey is to be carried out, steps should be taken to ensure that it is under statistical control so that the collected data is in accordance with the predetermined standard of accuracy.
- **Analysis of Data:** After data collection, the researcher turns to the task of analysing them. The bulk data should be compressed into a few manageable groups and tables for further analysis. The researcher can analyse the collected data by using various statistical measures.
- **Hypothesis Testing:** After analysing the data, the researcher should test the hypothesis, if any. He/she should check if the facts support the hypothesis or are contrary to the hypothesis. Statisticians have developed tests like chi-square test, *t*-test and F-test, for hypothesis testing. This testing further results in either acceptance or rejection of the hypothesis.
- **Generalizations and Interpretations:** The real value of research lies in its ability to arrive at certain generalizations. If the researcher cannot find a hypothesis to start with, he/she might seek to explain his/her findings on the basis of some theory. This is called 'interpretation'. This may give rise to new questions and further lead to more research.
- **Preparation of Report or Thesis:** This is the concluding step of research, where the researcher has to prepare the report of what has been done by him/her. Generally, the report should be designed in accordance with the following layout:
  - o **The Preliminary Pages:** Here the title, date, acknowledgements and foreword with the table of contents should be mentioned.
  - o **The Main Text:** This should be divided into introduction, summary, main report and conclusion.
  - o **End Matter:** This should contain appendices, bibliography and index.

A report should be written in a precise and objective style in simple language. Charts and illustrations should be included to lay emphasis on the study of research.

4.12.4 Introduction to Report Writing

A report can be defined as a written document which presents information in a specialized and concise manner. For example, a list of employees prepared by the HR department for salary distribution can be termed as a report. In other words, a report is information presented in a logical and concise manner.

There is a difference between report writing and other compositions because a report is written in a very short and conventional form. A report should cover all mandatory matters but nothing extra should be written. For writing a report, first the relevant data is collected, and then it is presented in a concise and objective manner. Then after successfully establishing the structure of the report, the formatting features that improve the look and readability of the report are added.

Reports can be divided into different categories. The two main types of reports are as follows:

1. Informational Report

The report that consists of a collection of data or facts and is written in an orderly way is called an informational report. The main purpose of this type of report is to present the information in its original form without any conclusion and recommendation. Informational reports are further divided into four parts, which are as follows:

- **Inspection Report:** The report which shows the outcome of a product or equipment to assure its proper functioning or to describe its quality is called an inspection report. This type of report is mainly used in manufacturing organizations.
- **Inventory Report:** The report which is made to keep the stock of various things like furniture, equipment, stationery, utensils and other accessories is called an inventory report.
- **Assessment Report:** These reports are made to maintain the database of the employees in an organization. Generally, these reports are useful for the HR department.
- **Performance Report:** The report which is made to measure the performance of the employees in an organization for purposes like appraisal or promotion are called performance reports.

2. Interpretive Report

Interpretive reports are those reports which contain a collection of data with its interpretation or any recommendation explicitly specified by the writer. This type of report also includes data analysis and conclusions made by the report writer. Writing interpretive reports is different from writing an informational report because it contains different elements. The possible elements that can be used in the interpretive reports are as follows:

- Cover
- Frontpiece
- Title Page
- Copyright Notice

NOTES

- Forwarding Letter
- Preface
- Acknowledgements
- Table of Contents
- List of Illustrations
- Abstract and Summary
- Introduction
- Discussion
- Conclusions
- Recommendations
- Appendices
- List of References
- Bibliography
- Glossary
- Index

Characteristics of a Good Report

Reports are used for various purposes by various departments of an organization. Industries, governments, businesses and scientific projects, all resort to report writing to collect information and keep track of their performance and progress. The most important aspect of a report is to convey the information in clear-cut terms. It should provide facts in a direct, straightforward and accurate style. In this light, the characteristics of a good report can be classified under four heads, which are as follows:

- Language and style of the report
- Structure of the report
- Presentation of the report
- References in the report

Each of these aspects of report writing needs to be given due attention, as they are interrelated to each other. A report given with a lucid style but with very less and hypothetical information is of no use to the reader. Similarly, the report writer needs to avoid overcrowding of information that may make the reader feel confused and lost in reading data, thereby, losing its charm. A systematic scrutiny of each of these aspects of a report is, therefore, necessary.

Language and Style of Report

A report must have a clear logical structure with clear indication of where the ideas are leading. It should be able to make a good first impression. The presentation of the report is very important. All reports must be written in good language, using short sentences and correct grammar and spellings. The main points to be kept in mind in this light are as follows:

- **Context and Style**
  - o Appropriate, informative title for the content of report
  - o Crisp, specific, unbiased writing with minimal jargon

- o Adequate analysis of prior relevant research

- **Questions/Hypotheses**
  - o Clearly stated questions or hypotheses
  - o Thorough operational definitions of key concepts along with exact wording or measurement of key variables
- **Research Procedures**
  - o Full and clear description of the research design
  - o Demographic profile of the participants/subjects
  - o Specific data gathering procedures
- **Data Analysis**
  - o Appropriate inferential statistics for sample or experimental data and appropriate use of descriptive statistics
  - o Clear and reasonable interpretation of the statistical findings, accompanied by effective tables and figures
- **Summary**
  - o Fair assessment of the implications and limitations of the findings
  - o Effective commentary on the overall implications of the findings for theory and/or policy

Structure of Report

Before you write a report, you should define the high level structure of the report. Defining a clear logical structure will make the report easier to write and to read. There are two types of report structures, which are listed as follows:

- **Report Structure I:** In general, the report writing structure comprises the following sub-headings:
  - o Title Page
  - o Abstract
  - o Table of Contents
  - o Introduction
  - o Technical Detail and Results
  - o Discussion and Conclusions
  - o References
  - o Appendices
- **Report Structure II:** There is also a specific structure of report writing pertaining to technical or scientific reports which is as follows:
  - o Introduction
  - o Background and Context
  - o Technical Details
  - o Results
  - o Discussion and Conclusion
- **Order of Writing:** The following is the correct order of writing:
  - o Start with the technical chapters/sections.
  - o Follow with the discussion.

NOTES

- o Finally, write the conclusions, introduction and abstract, if you are including any.
- **Appendix:** The appendix should contain the following:
  - o Material that suits or goes well with the flow of the main report but cannot be included in the main text of the report either because it is too long or is not essential reading. For example, lists of parameter values.
  - o Bibliography, i.e., list of all the sources of material, you referred to in your report.

Presentation of Report

As stated earlier, mere data overloading or just a lucid style of writing may not be a plus point to good report writing. Both the aspects need to be given due consideration, so that they interact to give a simple, easy-to-read and comprehensive type of report. Same goes with the presentation of the contents of the report. Printing mistakes, informal use of font size and style can distract the attention of the reader. On the other hand, effective use of tables and figures for better understanding of data and writing its conclusions facilitate easy comprehension. The main points of focus where due attention is required on part of the report writer are as follows:

- **Capitals:** This requires taking care of the following aspects:
  - o Using capitals only for proper nouns, place names, organization names and so on
  - o Defining acronyms at the first point of usage. For example, Incorporated (Inc.)
  - o Using bold, italics or underlines for emphasis, instead of capitals
- **Headings:** The basic points to be kept in mind for headings are as follows:
  - o Differentiate headings from the rest of the text using different fonts, bold, italics or underlines
  - o Maintain consistency in formatting headings using predefined styles.
  - o Avoid headings beyond three levels
- **Tables, Figures and Equations:** In general, certain formatting standards are pursued while giving tables and figures that are as follows:
  - o Descriptive labelling of all tables at the top with reference in the text
  - o All figures must be labelled descriptively at the top and must be referenced in the text
  - o All equations must be numbered consecutively
- **General Presentation:** The following points must be considered for preparing any general presentation:
  - o Sheets should be plain like white A4 size, printed on one side only
  - o Text should be justified on both sides and leave a blank line between paragraphs
  - o A staple in the top right hand corner is sufficient for most of the reports
- **References in the Report:** Several report types like scientific, engineering, technical and census reports contain either original writing or text adopted from previous work. As such, a report writer should be careful and avoid the violation of copyright laws and plagiarism. The necessary rule of thumb in this regard can be stated as follows:

- *Citations and Referencing*
  - A citation is the acknowledgement in your writing of the work of other authors and includes paraphrasing and making direct quotes.
  - Unless citation is very necessary, you should write the material in your own words. This shows that you understand what you have read and know how to apply it to your own context.
  - Direct quotes should be used sparingly.
- *Direct Quotes*
  - *Short Direct Quotes:* These need to be placed between quotation marks. For example, Research theorist S. Rosenfield defines a cluster as a 'geographically bounded concentration of similar, related or complementary businesses, with active channels for business transactions, communications and dialogue that share specialized infrastructure, common opportunities and threats'. This shows clearly that the words being used are not your own words.
  - *Longer Direct Quotes:* There are occasions when it is useful to include longer direct quotes. If you are quoting more than about 40 words, you should again use quotation marks but also indent the text. For example, the sustainability of higher value added industry is grounded in the diminishing significance of cost structures. At the level of the European Union, a weak capacity to innovate has been identified as an innovation, in the sense of product, process and organizational innovation, accounts for a very large amount, perhaps 80–90 per cent of the growth in productivity in advanced economies.

Mechanics of Writing a Report

There are several mechanics of writing a report, which are strictly followed for preparing technical reports. The following points should be considered for writing a technical report:

- **Size and Physical Design:** The manuscript, if handwritten, should be in black or blue ink and on unruled paper of 8½" × 11" size. A margin of at least one-and-half inches is set at the left side and half inch at the right side of the paper. The top and bottom margins should be of one inch each. If the manuscript is to be typed, then all typing should be double spaced and on one side of the paper, except for the insertion of long quotations.
- **Layout:** According to the objective and nature of the research, the layout of the report should be decided and followed in a proper manner.
- **Quotations:** Quotations should be punctuated with quotation marks and double spaces, forming an immediate part of the text. However, if a quotation is too lengthy, then it should be single spaced and indented at least half an inch to the right of the normal text margin.
- **Footnotes:** Footnotes are meant for cross-references. They are placed at the bottom of the page, separated from the textual material by a space of half an inch as a line that is around one-and-a-half inches long. Footnotes are always typed in a single space, though they are divided from one another by double space.
- **Documentation Style:** The first footnote reference to any given work should be complete, giving all essential facts about the edition used. Such footnotes follow a general sequence and order:

- In case of the single volume reference:
  - Author's name in normal order
  - Title of work, underlined to indicate italics
  - Place and date of publication
  - Page number reference

  For example:

  John Gassner, *Masters of the Drama*. New York: Dover Publications, Inc.1954, p.315.
- In case of a multi-volume reference:
  - Author's name in the normal order
  - Title of work, underlined to indicate italics
  - Place and date of publication
  - Number of the volume
  - Page number reference

  For example:

  George Birkbeck Hill, *Life Of Johnson*. Whitefish, June 2004, Volume 2, p.124.
- In case of works arranged alphabetically:
  - For works arranged alphabetically, such as encyclopaedias and dictionaries, page reference is usually not needed. In such cases, order is illustrated according to the names of the topics.
  - Name of the encyclopaedia.
  - Number of editions.

  For example:

  'Salamanca', *Encyclopaedia Britannica*, 14th Edition.
- In case of periodicals reference:
  - Name of the author in normal order.
  - Title of article, in quotation marks.
  - Name of the periodical, underlined to indicate italics.
  - Volume number.
  - Date of issuance.
  - Pagination.

  For example:

  P.V. Shahad, 'Rajesh Jain's Ecosystem', in *Business Today*, Vol.14, 18 December 2005, p. 28.
- In case of multiple authorship:

  If there are more than two authors or editors, then in the documentation, the name of only the first is given and the multiple authorship is indicated by '*et al*' or 'and others'.
  - Author's name in normal order.

NOTES

- Title of work, underlined to indicate italics.
- Place and date of publication.
- Pagination.

For example:

Alexandra K. Wigdor, *Ability Testing: Uses Consequences and Controversies*, 1981, p.23.

Subsequent references to the same work need not be detailed. If the work is cited again without any other work intervening, it may be indicated as *ibid*, followed by a comma and the page number.

- **Punctuations and Abbreviations in Footnotes:** Punctuation concerning the book and author names has already been discussed. They are general rules to be strictly adhered. Some English and Latin abbreviations are often used in bibliographies and footnotes to eliminate any repetition.
- **Use of Statistics, Charts and Graphs:** Statistics contribute to clarity and simplicity in a report. They are usually presented in the form of tables, charts, bars, line-graphs and pictograms.
- **Final Draft:** It requires careful scrutiny with regard to grammatical errors, logical sequence and coherence in the sentences of the report.
- **Index:** An index acts as a good guide to the reader. It can be prepared both as subject index and author index, giving names of subjects and names of authors, respectively. The names are followed by the page numbers of the report, where they have appeared or been discussed.

4.12.5 Research Report: An Overview

In simple terms, a research report means a written document, which describes the findings of some individual or a group of individuals. It gives an account of something seen, heard, done, and so on. The findings may comprise such information like data, surveys, resolutions, or policies on which the concerned individual or individuals have to submit their reports about the proceedings along with the relevant conclusions.

The preparation and presentation of a research report is the most important part of the research process. No matter how well designed the research study is, it is of little value, unless communicated effectively to others in the form of a research report. Moreover, if the report is confusing or poorly written, then the time and effort spent on gathering and analysing data would be wasted. It is therefore, essential to summarize and communicate the result to the management of an organization with the help of an understandable and logical research report.

Research reports are helpful during the research study, in the sense that they facilitate maintenance of vast data in a logical way. Thus, in case the researcher experiences any difficulty during the course of the study, it becomes easier to refer to the contents of the report to get the relevant data. Research report writing essentially involves systematic arrangement of data. This helps in discovering flaws in reasoning, which may have been missed earlier while conducting research.

NOTES

Format of Research Report

The layout of the research report is of utmost importance because the reader should be able to grasp logically, what has been said and not feel lost in the bulk findings mentioned in the research. This requires preparing a proper layout of the report. Report layout means allotting the research findings in a comprehensible format. The layout should contain the following points:

- **Preliminary Pages:** In the preliminary pages, the report should carry a 'title' and a 'date', followed by acknowledgements in the form of 'Preface' or 'Foreword'. The 'Table of Contents' should come next, followed by a 'list of tables and illustrations'. This entails the reader to an easy reading and quick location of the required information.
- **Main Text:** The main text comprises the complete outline of the research report with all the details. The title of the research study is repeated at the top of the first page of the main text, and then followed with the other details on the pages numbered consecutively, beginning with the second page. The main text can be classified into the following sections:
    - o **Introduction:** The purpose of introduction is to introduce the research projects to the readers. It should clearly state the objectives of research, i.e., it should make clear, why the problem was considered worth investigating. A brief summary of other relevant research can be included as well to enable the reader to see the present study in that context.
    - o **Methodology used for Performing the Study:** The introduction should contain answers to questions like how was the study carried out, what was the basic design, what were the experimental directions, what questions were asked in the questionnaires used, and so on. Besides this, the scope and limitations of the study must be marked out.
    - o **Statement of Findings and Recommendations:** The research report should comprise a statement of findings and recommendations in a non-technical language so that it is easily comprehensible.
    - o **Results:** A detailed presentation of the findings of the study, with supporting data in tabular forms along with the validation of results, should be given. This section should contain statistical summaries and deductions of the data rather than the raw data. There should be a logical sequence and sectional presentation of the results.
    - o **Implications of the Result:** The researcher should write down his/her results clearly and precisely, again at the end of the main text. The implications derived from the results of the research study should be stated in the research plan. The report should also mention the conclusion drawn from the study, which should be clearly related to the hypothesis stated in the introductory section.
    - o **Summary:** The next step is to conclude the report with a short summary, mentioning in brief the research problem, the methodology, the major findings and the major conclusions drawn from the research results.

- o **End Matter:** The end of the research report should consist of appendices, listed in respect of all technical data such as questionnaires, sample information and mathematical derivations. The bibliography of the referred sources and an index should also be given.

Precautions for Writing Research Reports

A research report is the means of conveying the research study to a specific target audience. The following precautions should be taken while preparing the research report:

- It should be long enough to cover the subject and short enough to preserve interest.
- It should not be dull and complicated.
- It should be simple, without the usage of abstract terms and technical jargons.
- It should offer ready availability of findings with the help of charts, tables and graphs, as readers prefer quick knowledge of the main findings.
- The layout of the report should be in accordance with the objective of the research study.
- There should be no grammatical errors and writing should adhere to techniques of report writing in case of quotations, footnotes and documentations.
- It should be original, intellectual and contribute to the solution of a problem or add knowledge to the concerned field.
- Appendices should be listed with respect to all the technical data in the report.
- It should be attractive, neat and clean, whether handwritten or typed.
- The report writer should be careful about the possessive form of the word 'it is' with 'it's'. The correct possessive form of 'it's' is 'its'. The use of 'it is' is the contractive form of 'it is'.
- A report should not have contractions. Examples are 'didn't' or 'it's'. In report writing, it is best to use the non-contractive form. Hence, the examples would be replaced by 'did not' and 'it is'. Using 'Figure' instead of 'Fig.' and 'Table' instead of 'Tab.' will spare the reader of having to translate the abbreviations while reading. If abbreviations are used, use them consistently throughout the report. For example, do not switch between 'versus' and 'vs'.
- It is advisable to avoid using the word 'very' and other such words that try to embellish a description. They do not add any extra meaning and, therefore, should be dropped.
- Repetition hampers lucidity. The report writer must avoid repeating the same word more than once within a sentence.
- When using the words 'this' or 'these', it must be clear to the reader as to what is being referred to. This reduces ambiguity in the writing and helps to tie sentences together.
- Do not use the word 'they' to refer to a singular person. You can either rewrite the sentence to avoid needing such a reference or use the singular 'he or she.'

NOTES

Check Your Progress

22. List the steps involved in the planning of writing a good report.
23. Identify the characteristics of a good report.

NOTES

4.13 TABLES AND CHARTS

Classification of data is usually followed by tabulation, which is considered as the mechanical part of classification.

Tabulation is the systematic arrangement of data in columns and rows. The analysis of the data is done so by arranging the columns and rows to facilitate analysis and comparisons.

Tabulation has the following objectives:

(*i*) Simplicity. The removal of unnecessary details gives a clear and concise picture of the data.

(*ii*) Economy of space and time.

(*iii*) Ease in comprehension and remembering.

(*iv*) Facility of comparisons. Comparisons within a table and with other tables may be made.

(*v*) Ease in handling of totals, analysis, interpretation, and so on.

Construction of Tables

A table is constructed depending on the type of information to be presented and the requirements of statistical analysis. The following are the essential features of a table:

(*i*) *Title:* It should have a clear and relevant *title*, which describes the contents of the table. The title should be brief and self explanatory.

(*ii*) *Stubs and Captions:* It should have clear headings and sub headings. Column headings are called *captions* and row headings are called *stubs*. The stubs are usually wider than the captions.

(*iii*) *Unit:* It should indicate all the *units* used.

(*iv*) *Body:* The *body* of the table should contain all information arranged according to description.

(*v*) *Headnote:* The *headnote* or prefatory note, placed just below the title, in a less prominent type, gives some additional explanation about the table. Sometimes, the headnote consists of the unit of measurement.

(*vi*) *Footnotes:* A *footnote* at the bottom of the table may clarify some omissions of special features. A source note gives information about the source used, if any.

(*vii*) *Arrangement of Data:* Data may be arranged according to requirements in chronological, alphabetical, geographical or any other order.

(*viii*) *Emphasis:* The items to be emphasized may be put in different print or marked suitably.

(*ix*) *Other Details:* Percentages, ratios, and so on, should be shown in separate columns. Thick and thin lines should be drawn at proper places.

A table should be easy to read and should contain only the relevant details. If the aim of clarification is not achieved, the table should be redesigned.

Types of Tables

Depending on the nature of the data and other requirements, tables may be divided into various types. They are given below:

- *General Tables or Reference Tables:* These contain detailed information for general use and reference, e.g., tables published by government agencies.

- *Specific Purpose or Derivative Tables:* They are usually summarized from general tables, and are useful for comparison and analytical purposes. Averages, percentages, and so on, are incorporated along with information in these tables.
- *Simple and Complex Tables:* A table showing only one characteristic is a simple table (see Table 4.5). The more common tables are complex and show two or more characteristics or groups of items.

***Table 4.5** Simple Table*

Cinema Attendance among Adult Male Factory Workers in Bombay
March 1972

Frequency	*Number of Workers*
Less than once a month	3780
1 to 4 times a month	1652
More than 4 times a month	926

Table 4.6 is the result of a survey on the cinema going habits of adult factory workers.

***Table 4.6** Simple Table*

Cinema Attendance among Adult Male Factory Workers in Bombay
March 1972

Cinema Attendance	*Single*		*Married*	
Frequency	*Under 30*	*Over 30*	*Under 30*	*Over 30*
Less than once a month	122	374	1404	1880
1–4 times a month	1046	202	289	115
More than 4 times a month	881	23	112	10
Total	2049	599	1805	2005

It is obvious that the tabular form of classification of data is a great improvement over the narrative form.

Frequently, table construction involves deciding which attribute should be taken as primary and which as secondary. For the previous table, we can also consider that whether it would be improved further if 'under 30' and '30 and over' had been the main column headings and 'single' and 'married' the sub headings. The modifications depend on the purpose of the table. If the activities of *age groups* are to be compared, it is best left as it stands. However, if a comparison between men of different *marital status* is required, the change would be an improvement.

Advantages of Tabulation of Data

The advantages of tabulation of data are given below:

(*i*) Tabulated data can be more easily understood and grasped than untabulated data.

(*ii*) A table facilitates comparisons between subdivisions and with other tables.

(*iii*) It enables the required figures to be located easily.

(*iv*) It reveals patterns within the figures, which otherwise might not have been obvious, e.g., from the previous table, we can conclude that regular and frequent cinema attendance is mainly confined to younger age group.

NOTES

(*v*) It makes the summation of items and the detection of errors and omissions easier.

(*vi*) It obviates repetition of explanatory phrases and headings and, hence, takes less space.

4.13.1 Presentation of Data

In a graph, the independent variable should always be placed on the horizontal or x-axis and the dependent variable on the vertical or y-axis.

Line Graph

Here, the points are plotted on paper (or graph paper) and joined by straight lines. Generally, continuous variables are plotted by a line graph.

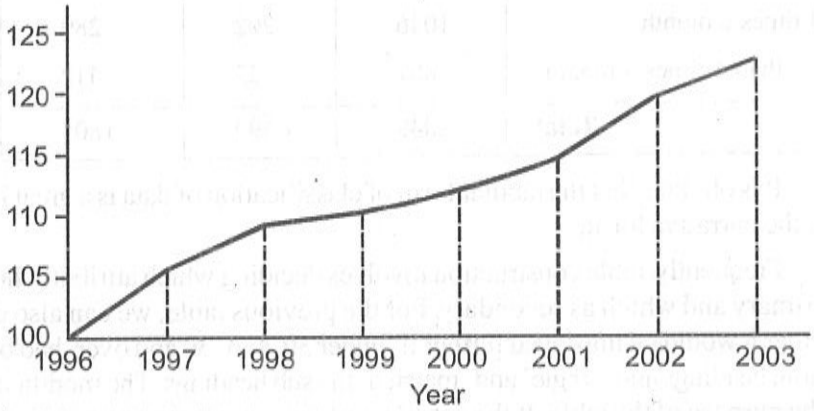
Example 4.21: The monthly averages of Retail Price Index from 1996 to 2003 (Jan. 1996 = 100) were as follows:

Year	1996	1997	1998	1999	2000	2001	2002	2003
Retail Price Index	100	105.8	109.0	109.6	110.7	114.5	119.3	122.3

Draw a diagram to display these figures.

Solution: Here, years are plotted along the horizontal line and the retail price index along the vertical line.

Erect perpendiculars to horizontal line from the points marked as retail price index for the years 1997, 1998, ..., 2003 and cut off these ordinates according to the given data and, thus, various points will be plotted on the paper. Join these points by straight lines.



Frequency Polygon

A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or midpoints of class intervals are plotted against the frequencies, and these plotted points are joined together by straight lines. Since the frequencies generally do not start at zero or end at zero, this diagram as such would not touch the horizontal axis. However, since the area under the entire curve is the same as that of a histogram which is 100 per cent of the data presented, the curve can be enclosed so that the starting point is joined with a fictitious preceding point whose value is zero. This ensures that the start of the curve is at horizontal axis and the last point is joined with a fictitious succeding point whose value is also zero, so that the curve ends at the horizontal axis. This enclosed diagram is known as the frequency polygon.

NOTES

We can construct the frequency polygon from the table presented for the ages of 30 workers as shown in Figure 4.10:

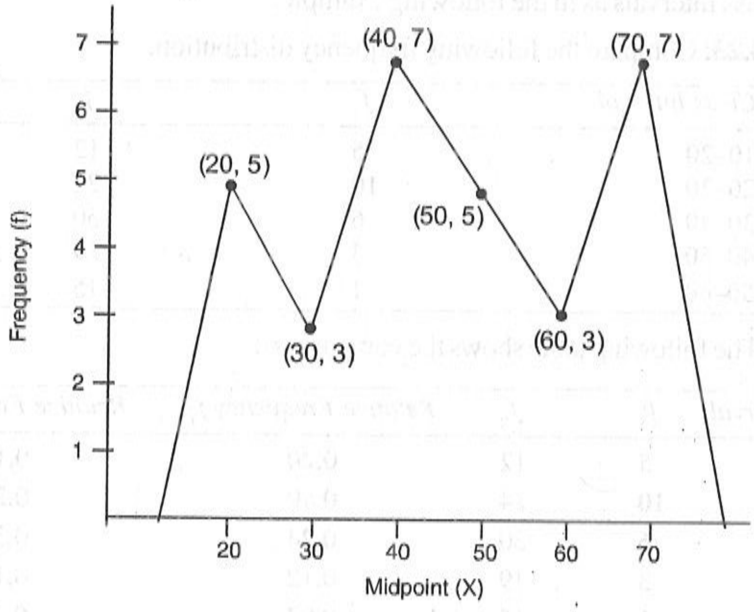


Fig. 4.10 Frequency Polygon

Relative Frequency

In a frequency distribution, if the frequency in each class interval is converted into a proportion, dividing it by the total frequency, we get a series of proportions called *relative frequencies*. A distribution presented with relative frequencies rather than actual frequencies is called a *relative frequency* distribution. The sum of all relative frequencies in a distribution is 1.

Example 4.22: Calculate relative frequency from the table given below:

Class Interval	*Frequency*
25–35	7
35–45	9
45–55	22
55–65	7
65–75	3
75–85	2

Solution: This example shows that the sum of all relative frequencies in a distribution is 1.

Class Interval	*Frequency*	*Relative Frequency*	*Explanation*
25–35	7	0.14	$\frac{7}{50}=0.14$
35–45	9	0.18	$\frac{9}{50}=0.18$
45–55	22	0.44	
55–65	7	0.14	
65–75	3	0.06	
75–85	2	0.04	
Total	50	1.00	

NOTES

The concept of relative frequencies is useful in sampling theory. It can also be used to compare two frequency distributions with unequal total frequency with the same series of class intervals as in the following example.

Example 4.23: Compare the following frequency distribution.

Class Interval	f_1	f_2
10–20	5	12
20–30	10	24
30–40	6	30
40–50	3	19
50–60	1	15

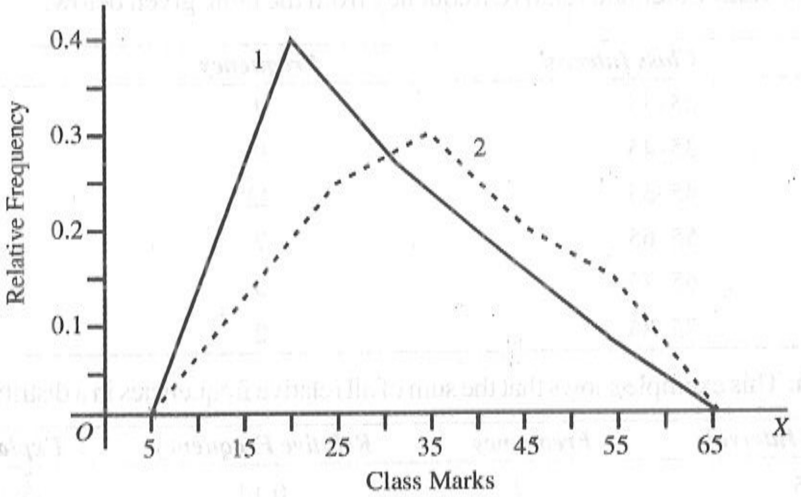
Solution: The following table shows the comparison:

Class Interval	f_1	f_2	*Relative Frequency* f_1	*Relative Frequency* f_2
10–20	5	12	0.20	0.12
20–30	10	24	0.40	0.24
30–40	6	30	0.24	0.30
40–50	3	19	0.12	0.19
50–60	1	15	0.04	0.15
Total	25	100	1.00	1.00

A direct visual comparison of two frequency distributions can be made by drawing their frequency polygons.

Example 4.24: Draw frequency polygons for the relative frequency distributions given in Example 4.23.

Solution: The following is the frequency polygon for the relative frequencies as mentioned in Example 4.23.



4.13.2 Ogive Curves and Histogram

Cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than ogive and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.

Class Interval (Years)	*Mid-Point*	*(f)*	*Cum. Freq. (Less Than)*	*Cum. Freq. (Greater Than)*
15 and upto 25	20	5	5 (less than 25)	30 (more than 15)
25 and upto 35	30	3	8 (less than 35)	25 (more than 25)
35 and upto 45	40	7	15 (less than 45)	22 (more than 35)
45 and upto 55	50	5	20 (less than 55)	15 (more than 45)
55 and upto 65	60	3	23 (less than 65)	10 (more than 55)
65 and upto 75	70	7	30 (less than 75)	7 (more than 65)

(*i*) Less than Ogive: In this case, the less than cumulative frequencies are plotted against the upper boundaries of their respective class intervals. Less than ogive is shown in Figure 4.11

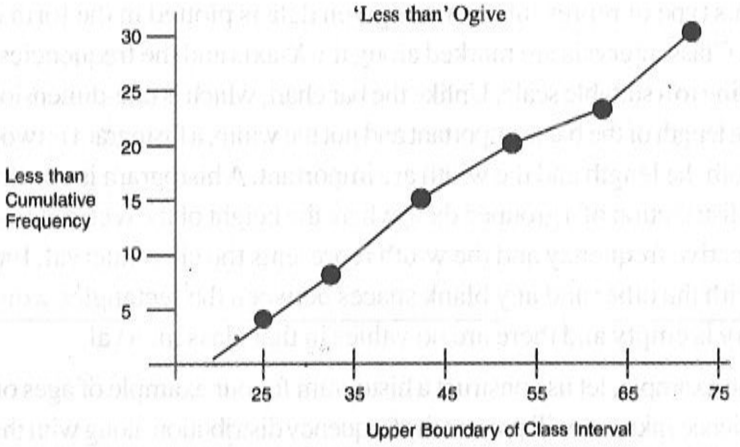


Fig. 4.11 Less than Ogive

(*ii*) Greater than Ogive: In this case, the greater than cumulative frequencies are plotted against the lower boundaries of their respective class intervals. More than ogive is shown in Figure 4.12.

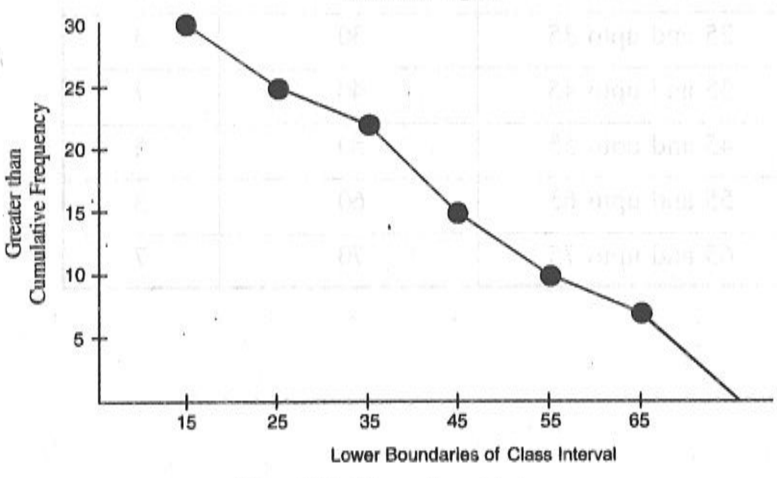


Fig. 4.12 More than Ogive

NOTES

These ogives can be used for comparison purposes. Several ogives can be drawn on the same grid, preferably with different colours for easier visualization and differentiation.

Although diagrams and graphs are powerful and effective media for presenting statistical data, they can only represent a limited amount of information and they are not of much help when intensive analysis of data is required.

Histograms

A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.

The word 'histogram' is derived from the Greek word *histos* which means 'anything set upright' and *gramma* which means 'drawing, record, writing'. It is considered as the most important basic tool of statistical quality control process.

In this type of representation, the given data is plotted in the form of a series of rectangles. Class intervals are marked along the *X*-axis and the frequencies along the *Y*-axis according to a suitable scale. Unlike the bar chart, which is one-dimensional, meaning that only the length of the bar is important and not the width, a histogram is two-dimensional in which both the length and the width are important. A histogram is constructed from a frequency distribution of a grouped data, where the height of the rectangle is proportional to the respective frequency and the width represents the class interval. Each rectangle is joined with the other and any blank spaces between the rectangles would mean that the category is empty and there are no values in that class interval.

As an example, let us construct a histogram for our example of ages of 30 workers. For convenience sake, we will present the frequency distribution along with the midpoint of each interval, where the midpoint is simply the average of the values of the lower and the upper boundary of each class interval. The frequency distribution table is shown as follows:

Class Interval (Years)	*Mid-Point*	*(f)*
15 and upto 25	20	5
25 and upto 35	30	3
35 and upto 45	40	7
45 and upto 55	50	5
55 and upto 65	60	3
65 and upto 75	70	7

The histogram of this data is shown in the Figure 4.13.

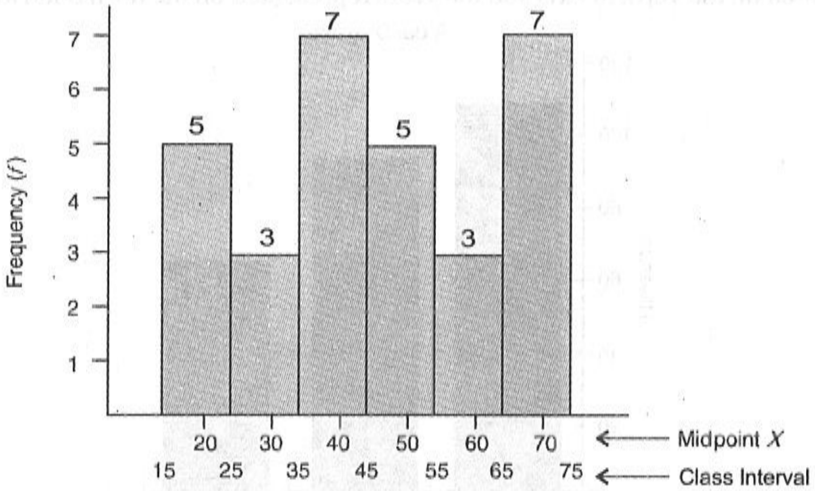


Fig. 4.13 Histograms

NOTES

4.13.3 Diagrams

The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data. These important features are more simply presented in the form of graphs. Also, diagrams facilitate comparisons between two or more sets of data.

The diagrams should be clear and easy to read and understand. Too much information should not be shown in the same diagram; otherwise, it may become cumbersome and confusing. Each diagram should include a brief and self explanatory title dealing with the subject matter. The scale of the presentation should be chosen in such a way that the resulting diagram is of appropriate size. The intervals on the vertical as well as the horizontal axis should be of equal size; otherwise, distortions would occur.

Diagrams are more suitable to illustrate the data which is discrete, while continuous data is better represented by graphs. The following are the diagrammatic and the graphic representation methods that are commonly used:

One-Dimensional Diagrams

Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram. Additionally, the bars should be equally spaced.

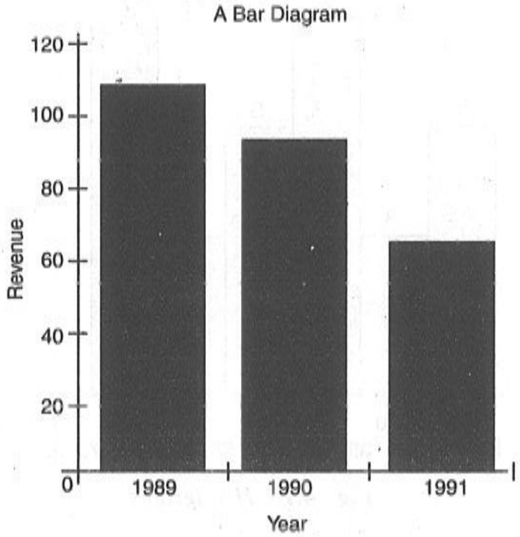
Example 4.25: Suppose that the following were the gross revenues in $100,000.00 for a company *XYZ* for the years 1989, 1990 and 1991.

Year	*Revenue*
1989	110
1990	95
1991	65

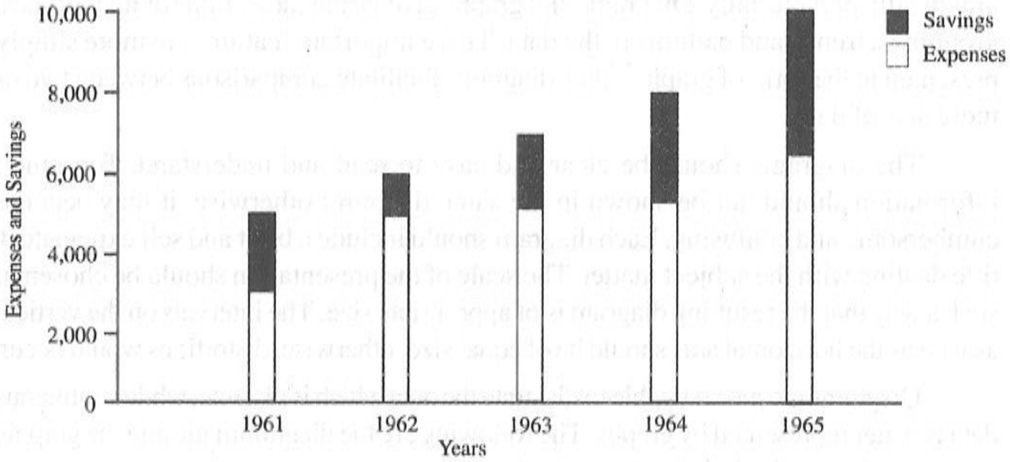
Construct a bar diagram for this data.

NOTES

Solution: The bar diagram for this data can be constructed as follows with the revenues represented on the vertical axis and the years represented on the horizontal axis.



When each figure is made up of two or more component figures, the bars may be subdivided into components. Too many components should not be shown.



Component Bar Chart Showing Expenses and Savings of Mr X

Annual Income, Expenses and Savings of Mr X

Year	*Amounts in ₹ of*			*Percentages of*		
	Income	*Expenses*	*Savings*	*Income*	*Expenses*	*Savings*
1961	5000	3000	2000	100.0	60.0	40.0
1962	6000	5000	1000	100.0	83.3	16.7
1963	7000	5000	2000	100.0	71.4	28.6
1964	8000	5000	3000	100.0	62.5	37.5
1965	10000	6000	4000	100.0	60.0	40.0

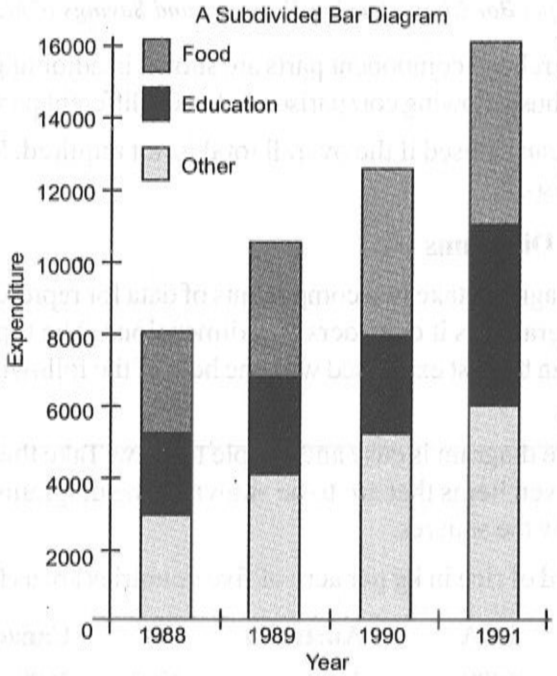
The bars drawn can be further subdivided into components depending upon the type of information to be shown in the diagram. This will be clear by the following example in which we present three components in a bar.

NOTES

Example 4.26: Construct a subdivided bar chart for the three types of expenditures in dollars for a family of four for the years 1988, 1989, 1990 and 1991 which is given as follows:

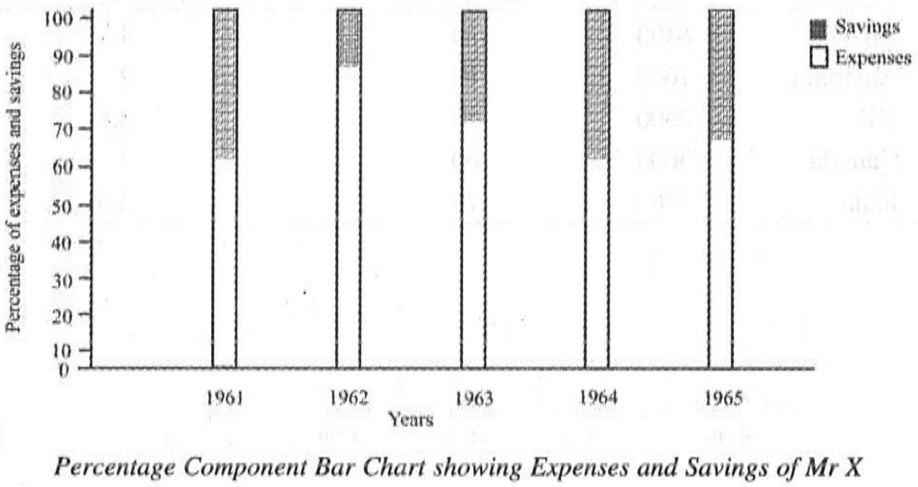
Year	*Food*	*Education*	*Other*	*Total*
1988	3000	2000	3000	8000
1989	3500	3000	4000	10500
1990	4000	3500	5000	12500
1991	5000	5000	6000	16000

Solution: The subdivided bar chart would be as follows:



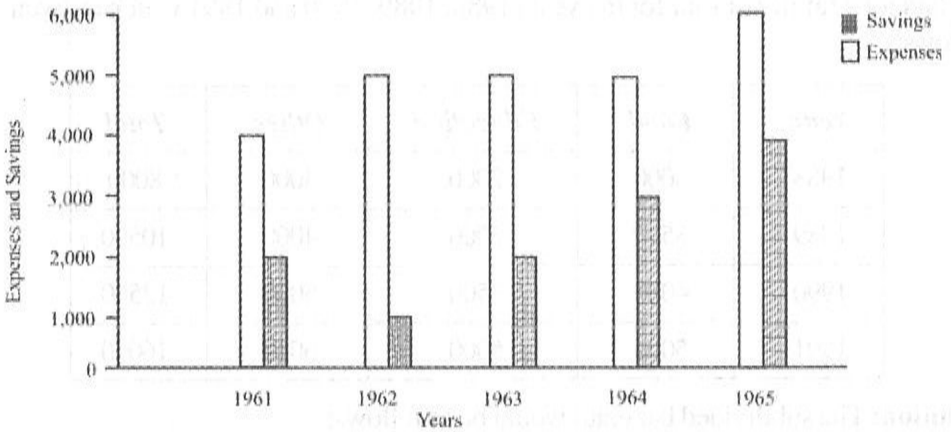
Percentage Component Bars or Divided Bar Charts

When in the previous case, the component lengths represent the percentages (instead of the actual amounts) of each component, we get percentage component bar charts. The heights of all the bars will be the same.



Percentage Component Bar Chart showing Expenses and Savings of Mr X

Multiple Bar Charts



Multiple Bar Chart showing Expenses and Savings of Mr X

Here, the interrelated component parts are shown in adjoining bars, coloured or marked differently, thus, allowing comparison between different parts.

These charts can be used if the overall total is not required. Some charts given earlier show totals also.

Two-Dimensional Diagrams

Two-dimensional diagrams take two components of data for representation. These are also called area diagrams as it considers two dimensions. The types are rectangles, squares and pie. It can be best explained with the help of the following square diagram example:

Squares: The square diagram is easy and simple to draw. Take the square root of the values of various given items that are to be shown in the diagrams and then select a suitable scale to draw the squares.

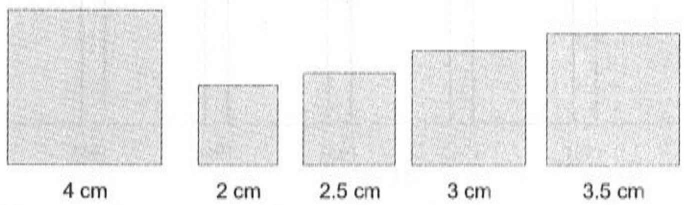
Example 4.27: Yield of rice in kg per acre of five countries are as follows:

Country	**USA**	**Australia**	**UK**	**Canada**	**India**
Yield of rice in kg per acre	6400	1600	2500	3600	4900

Represent this data using square diagram.

Solution: To draw the square diagrams, calculate as follows:

Country	*Yield*	*Square root*	*Side of the Square in cm*
USA	6400	80	4
Australia	1600	40	2
UK	2500	50	2.5
Canada	3600	60	3
India	4900	70	3.5



Pie Diagram

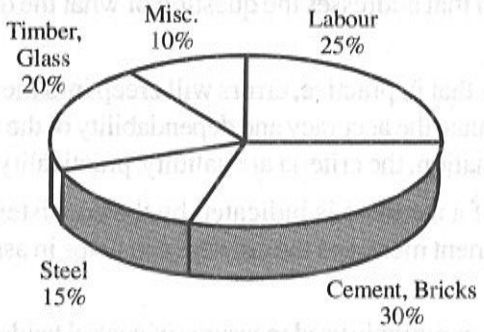
This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle, and it is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

Example 4.28: The following figures relate to the cost of the construction of a house. The various components of cost that go into it are represented as percentages of the total cost.

Item	*% Expenditure*
Labour	25
Cement, bricks	30
Steel	15
Timber, glass	20
Miscellaneous	10

Construct a pie chart for the above data.

Solution: The pie chart for this data is presented as follows:



Pie charts are very useful for comparison purposes, especially when there are only a few components. If there are too many components, it may become confusing to differentiate the relative values in the pie.

Three-Dimensional Diagrams

Three-dimensional diagrams are also termed as volume diagram, and consist of cubes, cylinders, spheres, and so on. In these diagrams, three dimensions namely length, width and height are taken into account. Cubes are used to represent where the side of a cube is drawn in proportion to the cube root of the magnitude of data.

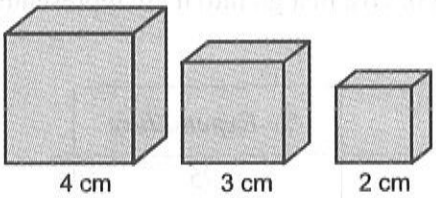
Example 4.29: Represent the following data using volume diagram.

Category	*Number of Students*
Undergraduate	64000
Postgraduate	27000
Professionals	8000

NOTES

Solution: The sides of cubes are calculated as follows:

Category	*Number of Students*	*Cube Root*	*Side of Cube*
Undergraduate	64000	40	4 cm
Postgraduate	27000	30	3 cm
Professional	8000	20	2 cm



4.14 SUMMARY

- Research problems are questions that indicate gaps in the scope or the certainty of our knowledge. They point either to the problematic phenomena, observed events that are puzzling in terms of the accepted ideas or to problematic theories and current ideas that are challenged by new hypothesis.
- The problem recognition process invariably starts with the decision-maker and some difficulty or decision dilemma that he/she might be facing. This is an action oriented problem that addresses the question of what the decision-maker should do.
- It is very obvious that in practice, errors will creep into measurements, making it necessary to evaluate the accuracy and dependability of the measuring instrument. For such an evaluation, the criteria are validity, practicality and reliability.
- The reliability of a measure is indicated by the consistency and stability with which the instrument measures the concept and helps in assessing the 'goodness' of a measure.
- There are several commonly used measures of central tendency, such as arithmetic mean, mode and median. These values are very useful not only in presenting the overall picture of the entire data but also for the purpose of making comparisons among two or more sets of data.
- Seasonal variation has been defined as predictable and repetitive movement around the trend line in a period of one year or less. For the measurement of seasonal variation, the time interval involved may be in terms of days, weeks, months or quarters.
- A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.
- The crudest measure of dispersion is the range of the distribution. The range of any series is the difference between the highest and the lowest values in the series.
- Correlation analysis is the statistical tool generally used to describe the degree to which, one variable is related to another. The relationship, if any, is usually assumed to be a linear one.

Check Your Progress

24. Define tabulation.
25. What are ogive curves? States its types.

- The term 'regression' was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move back or 'regress' toward the mean height of the population.
- Least squares method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line.
- The time series analysis method is quite accurate where future is expected to be similar to past. The underlying assumption in time series is that the same factors will continue to influence the future patterns of economic activity in a similar manner as in the past.
- When a time series shows an upward or downward long-term linear trend, then regression analysis can be used to estimate this trend and project the trends into forecasting the future values of the variables involved.
- Smoothing techniques improve the forecasts of future trends provided that the time series is fairly stable with no significant trend, cyclical or seasonal effect and the objective is to smooth out the irregular component of the time series through the averaging process.
- A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. Hypothesis refers to a provisional idea whose merit needs evaluation, but having no specific meaning.
- A statistical hypothesis test is a test in which a hypothesis is tested. It analyses gathered data to decide on hypothesis. In decision making such analysis is highly desired.
- SPSS is abbreviated term for Statistical Package for the Social Sciences, and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies.
- For the use of a chi-square test, data is required in the form of frequencies. The data expressed in percentages or proportion can also be used, provided it could be converted into frequencies.
- On completion of the research study and after obtaining the research results, the real skill of the researcher lies in terms of analysing and interpreting the findings and linking them with the propositions formulated in the form of research hypotheses at the beginning of the study.
- A research proposal means a written application that proposes to pursue or conduct a research study. It aims at presenting the idea around which the research study revolves.
- Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data. These important features are more simply presented in the form of graphs. Also, diagrams facilitate comparisons between two or more sets of data.

4.15 KEY TERMS

- **Measures of central tendency:** It is a measure used to describe data; the mean, median and mode are measures of central tendency.
- **Measure of dispersion:** The interquartile range is a measure of dispersion, which is calculated by subtracting the first quartile ( ) from the third quartile ( ). This gives the range of the middle half of the data set.
- **Correlation analysis:** It is the use of statistical correlation to evaluate the strength of the relations between variables.
- **Regression analysis:** In statistics, regression analysis is a statistical process for estimating the relationships among variables.
- **Time series analysis:** Time series analysis comprises methods for analysing time series data in order to extract meaningful statistics and other characteristics of the data.
- **SPSS:** SPSS is the acronym of Statistical Package for the Social Science. SPSS is one of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions.
- **Chi-square test:** It is a statistical method assessing the goodness of fit between a set of observed values and those expected theoretically.
- **Histogram:** It is a diagram consisting of rectangles whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

4.16 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Research problems are questions that indicate gaps in the scope or the certainty of our knowledge. They point either to the problematic phenomena, observed events that are puzzling in terms of the accepted ideas or to problematic theories, current ideas that are challenged by new hypothesis.
2. The research problem formulation involves the following interrelated steps:
   (a) Ascertaining the objectives of the decision-maker
   (b) Understanding the problem's background
   (c) Identifying and isolating the problem, rather than its symptoms
   (d) Determining the unit of analysis
   (e) Determining the relevant variables
   (f) Stating the research objectives and research questions (hypotheses)
3. Primary data is the data specially collected in a research by the researcher. These are products of experiments, surveys, interviews or observations conducted in the research. Primary data is generated and collected through specific tools of data collection, like questionnaires, by the researcher.
4. The reliability of a measure is indicated by the consistency and stability with which the instrument measures the concept and helps in assessing the 'goodness' of a measure. A measure is reliable to the degree that it supplies consistent results.

NOTES

5. Each measure of central tendency should meet the following requisites:
    (a) It should be easy to calculate and understand.
    (b) It should be rigidly defined. It should have only one interpretation so that the personal prejudice or bias of the investigator does not affect its usefulness.
    (c) It should be representative of the data. If it is calculated from a sample, then the sample should be random enough to be accurately representing the population.
    (d) It should have sampling stability. It should not be affected by sampling fluctuations. This means that if we pick 10 different groups of college students at random and compute the average of each group, then we should expect to get approximately the same value from each of these groups.
    (e) It should not be affected much by extreme values. If few very small or very large items are present in the data, they will unduly influence the value of the average by shifting it to one side or other, so that the average would not be really typical of the entire series. Hence, the average chosen should be such that it is not unduly affected by such extreme values.
6. The median is a measure of central tendency and it appears in the centre of an ordered data. It divides the list of ordered values in the data into two equal parts so that half of the data will have values less than the median and half will have values greater than the median.
7. Simple averages is the simplest method of isolating seasonal fluctuations in time series. It is based on the assumption that the series contain only the seasonal and irregular fluctuations.
8. The moving average (or the centred moving average) aims to eliminate seasonal and irregular fluctuations (S and I) from the original time series, so that this average represents the cyclical and trend components of the series.
9. A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.
10. The merits of mean deviation are as follows:
    (a) It is easy to understand.
    (b) As compared to standard deviation (discussed later), its computation is simple.
    (c) As compared to standard deviation, it is less affected by extreme values.
    (d) Since it is based on all values in the distribution, it is better than range or quartile deviation.
11. The crudest measure of dispersion is the range of the distribution. The range of any series is the difference between the highest and the lowest values in the series.
12. The theory by means of which quantitative connections between two sets of phenomena are determined is called the *Theory of Correlation.*
13. The following are the two methods used for estimating the regression equation:
    (a) Scatter diagram method
    (b) Least squares method
14. Some factors that cause seasonal variations are as follows:
    (a) Season and Climate
    (b) Customs and Festivals

NOTES

15. The concept of the moving averages is based on the idea that any large irregular component of time series at any point in time will have a less significant impact on the trend, if the observation at that point in time is averaged with such values immediately before and after the observation under consideration.

16. Formalized hypotheses have two variables, independent and dependent. The independent variable is the person, may be the scientist, who is going to put the hypothesis and the dependent variable is one that the person observes.

17. Testing a statistical hypothesis on the basis of a sample enables us to decide whether the hypothesis should be accepted or rejected. The sample data enable us to accept or reject the hypothesis. Since the sample data give incomplete information about the population, the result of the test need not be considered to be final or unchallengeable.

18. The advantage of command syntax programming language is that it helps in data reproducibility, simplifying repetitive tasks, performing complex data manipulations and analyses. In addition, the user can program specific syntax for some complex applications which are not available in the predefined menu structure. The command syntax can also be generated by pull-down menu interface and can be displayed in the output.

19. SPSS Statistics 17.0 is a comprehensive system for analysing data based on the Graphical User Interface (GUI). SPSS Statistics can acquire data from almost any form of file, and use them to create tabulated reports, charts, plots of distributions and trends, descriptive statistics and complex statistical analyses.

20. A goodness of fit test is a statistical test of how well the observed data supports the assumption about the distribution of a population. The test also examines that how well an assumed distribution fits the data.

21. A chi-square test can be used to test the independence of two variables each having at least two categories. The test makes a use of contingency tables, also referred to as cross-tabs, with the cells corresponding to a cross classification of attributes or events.

22. Some of the steps involved in the planning of writing a good report are as follows:
    (a) **Step 1:** The entire study should be analysed very logically in a thorough manner.
    (b) **Step 2:** A rough or preliminary draft of the outline of the proposed research, should be prepared with thought and care.
    (c) **Step 3:** The final outline should be developed.
    (d) **Step 4:** This can then be polished and refined, before being re-written.
    (e) **Step 5:** This can be followed by the compiling of the bibliography.
    (f) **Step 6:** Then the final draft can be written.

23. The characteristics of a good report can be classified under four heads, which are as follows:
    (a) Language and style of the report
    (b) Structure of the report
    (c) Presentation of the report
    (d) References in the report

24. Tabulation is the systematic arrangement of data in columns and rows. The analysis of the data is done so by arranging the columns and rows to facilitate analysis and comparisons.
25. Cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than ogive and the other one is greater than ogive.

4.17 QUESTIONS AND EXERCISES

Short-Answer Questions

1. Write a short note on problem identification and formulation.
2. What are the objectives of value of information with respect to research?
3. What type of scales are used for level measurement?
4. List the characteristics of arithmetic mean.
5. Define seasonal variation. How is it measured?
6. What do you mean by quartile deviation? List its characteristics.
7. Briefly describe the coefficient of determination.
8. Illustrate a scatter diagram method and state its importance.
9. What are the different types of variations for analysing time series?
10. List the characteristics of hypothesis.
11. List the enhanced features available in the new version of SPSS Statistics 17.0.
12. State the important properties of chi-square distribution.
13. Identify the different types of reports.
14. Write a short note on the language and style of reports.
15. What is frequency polygon?

Long-Answer Questions

1. Explain the concept of problem measurement in data analysis.
2. Discuss the importance of validity and reliability in measurement.
3. Describe some of the commonly used measures of central tendency. Support your answer with suitable examples.
4. Differentiate between simple average and moving average.
5. Discuss some of the common measures of dispersion in detail. Use suitable examples to support your answer.
6. Distinguish between correlation analysis and regression analysis.
7. Examine the various measures of trends.
8. 'For most purposes, SPSS is the widely used software.' Explain.
9. Discuss the various applications of a chi-square test.
10. Describe the different methods used in report writing.

4.18 FURTHER READING

NOTES

Kothari, C. R. 1995. *Research Methodology–Methods and Techniques*. New Delhi: Wiley Eastern Ltd.

Creswall, John W. 2008. *Research Designs: Quantitative, Qualitative and Mixed Methods Approaches.* London: Sage Publications.

Christenson, Larry B. *et al.* 2010. *Research Methods, Design and Analysis*, Eleventh edition. New Jersey: Allyn and Bacon.

Wilkinson, T. S. and P. L. Bhandarkar. 2003. *Methodology and Techniques of Social Research.* Mumbai: Himalaya Publishing House.

Chaudhary, C. M. 1991. *Research Methodology*. Rajasthan: RBSA Publishers.

Gupta, S. C. and V. K. Kapoor. 1996. *Fundamentals of Applied Statistics*. New Delhi: Sultan Chand & Sons.

Chiang, Alpha C. 1986. *Fundamental Methods of Mathematical Economics*. New York: McGraw-Hill.

தமிழ்நாடு திறந்தநிலைப் பல்கலைக்கழகம்
TAMILNADU OPEN UNIVERSITY

(மாநில திறந்தநிலைப் பல்கலைக்கழகம் தமிழக அரசால் நிறுவப்பட்டது.
UGC மற்றும் DEB அங்கீகாரம் பெற்றது. ஆசிய திறந்தநிலைப் பல்கலைக்கழகங்கள் மற்றும்
காமன்வெல்த் பல்கலைக்கழகங்களின் அமைப்பில் உறுப்பினராகப் பதிவு பெற்றது.)

577, அண்ணா சாலை, சைதாப்பேட்டை, சென்னை -600 015.
URL:www.tnou.ac.in, Email:tnouadmission@gmail.com, Admission Director Mobile:9345913378

UGC RECOGNISED PROGRAMMES

(அனைத்து இளநிலை/முதுநிலை பாடத்திட்டங்களும் UGC அங்கீகாரம் பெற்றவை)

UG PROGRAMMES

B.A. Tamil (NS)
B.A. Functional Tamil (NS)
B.Lit. Tamil (NS)
B.A. English
B.A. Englsih and Communication
B.A. History *
B.A. History and Heritage Management *
B.A. Tourism and Travel Studies
B.A. Political Studies *
B.A. Public Administration *
B.A. Human Rights *
B.A. Scociology *
B.A. Economics *
B.A. Business Economics
B.A. Urdu
B.A. Islamic Studies
B.A. Criminology and Criminal Justice Administration
B.Com.
B.Com. Accounting and Finance *
B.Com. Bank Management *
B.Com. Corporate Secretaryship *
B.Com. Computer Applications (NS)
B.B.A. *
B.B.A. Marketing Management * (NS)
B.B.A. Retail Management
B.B.A. Computer Applications
B.Sc. Mathematics
B.Sc. Mathematics with Computer Applications
B.Sc. Physics
B.Sc. Chemistry
B.Sc. Botany
B.Sc. Zoology
B.Sc. Psychology *
B.Sc. Geography *
B.Sc. Multimedia
B.Sc. Visual Communication
B.Sc. Apparel and Fashion Design / LE
B.Sc. Computer Science
B.C.A / LE
B.S.W (Social Work)*
B.P.A. Drama and Theatre Studies
B.Ed.(Spl.) (MR/VI/HI)
B.Ed.(Subject to Approval of UGC/NCTE)

PG PROGRAMMES

M.A. Tamil (NS)
M.A. English (NS)
M.A. Linguistics
M.A. Comparative Literature
M.A. Translation Studies
M.A. History *
M.A. Tourism and Travel Studies
M.A. Political Studies *
M.A. Public Administration *
M.A. Human Rights *
M.A. Police Administration
M.A. International Relations
M.A.Development Administration
M.A. Scociology * (NS)
M.A. Anthropology
M.A. Women Studies
M.A. Economics *
M.A. Islamic Studies
M.A. Criminology and Criminal Justice Administration (NS)
M.Com. *
M.B.A
M.B.A. Hospital Administration
M.B.A. Logistics Management
M.B.A. Shipping and Logistics Management
M.Sc. Mathematics
M.Sc. Physics
M.Sc. Chemistry
M.Sc. Botany
M.Sc. Zoology
M.Sc. Psychology
M.Sc. Counselling and Psychotherapy /NP/LE
M.Sc. Geography
M.Sc. Apparel and Fashion Design
M.Sc. Computer Science
M.C.A / LE
M.S.W (Social Work)*
M.L.I.Sc.

- * - Both English and Tamil Medium
- LE - Lateral Entry
- NP - Non-Psychology
- NS - Non-Semester

All 80 Programmes (UG-42, PG-38) offered by TNOU are recognized by the University Grants Commission (UGC) F.No.2-10/2018/(DEB-I), Dated 14th August 2018)

Interested Students from Government Colleges of Arts & Science, can undergo any of the skill Development Programmes through 16 short-term, 24 Certificates, 4 Diplomas, 14 Vocational Diplomas and 2 Advanced Diploma Programmes. (G.O.163 Dated: 22.05.2008)

The Students who could not get the admission in the interested subject in the Government Colleges of Arts & Science, can join in TNOU for the same subject under ODL Scheme in the same college. (G.O.150 Dated: 23.10.2020)

For online Admission visit:
http://tnouadmissions.in/onlineapp/

மண்டல மையங்கள்

சென்னை:7904509518 திருச்சி:9345913391 கோவை:9345913386 மதுரை:9345913388
தருமபுரி:9345913387 திருநெல்வேலி:9345913392 விழுப்புரம்:9345913393/9345913380 ஊட்டி:9345913390