



ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category-I University by MHRD-UGC]

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003



Directorate of Distance Education

M.A. [Sociology]

I - Semester

351 14

RESEARCH METHODS AND STATISTICS

Reviewer	
Dr. MA. Velusamy	Assistant Professor in Social Work, Directorate of Distance Education, Alagappa University, Karaikudi

Authors:

Dr (Mrs) Vasantha R Patri: *Chairperson, Indian Institute of Counselling, Delhi*
Units (1.2, 4, 14.2)

Dr Deepak Chawla: *Distinguished Professor, Dean, International Management Institute (IMI), New Delhi*

Dr Neena Sondhi: *Professor, International Management Institute (IMI), New Delhi*
Units (5.2, 8, 9, 10.2-10.3, 11.6, 12.2-12.2.1, 12.3-12.3.1)

Harish Kumar: *Associate Professor, AIE, Amity University, Noida*
Unit (7)

RP Hooda: *Emeritus Professor, Apeejay School of Management, Dwarka*
Units (10.4, 11.2-11.4, 13, 14.3)

CB Gupta: *Former Director and Former Professor of Finance and Accounting, Institute of Management Technology, Ghaziabad*

Vijay Gupta: *Distinguished Professor, Sharda University*
Unit (11.5)

Vikas® Publishing House: Units (1.0-1.1, 1.3- 1.7, 2, 3, 5.0-5.1, 5.3-5.7, 6, 10.0-10.1, 10.5-10.9, 11.0-11.1, 11.7-11.11, 12.0-12.1, 12.3.2, 12.4-12.8, 14.0-14.1, 14.4-14.8)

"The copyright shall be vested with Alagappa University"

All rights reserved. No part of this publication which is material protected by this copyright notice may be reproduced or transmitted or utilized or stored in any form or by any means now known or hereinafter invented, electronic, digital or mechanical, including photocopying, scanning, recording or by any information storage or retrieval system, without prior written permission from the Alagappa University, Karaikudi, Tamil Nadu.

Information contained in this book has been published by VIKAS® Publishing House Pvt. Ltd. and has been obtained by its Authors from sources believed to be reliable and are correct to the best of their knowledge. However, the Alagappa University, Publisher and its Authors shall in no event be liable for any errors, omissions or damages arising out of use of this information and specifically disclaim any implied warranties or merchantability or fitness for any particular use.



VIKAS®

Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.

E-28, Sector-8, Noida - 201301 (UP)

Phone: 0120-4078900 • Fax: 0120-4078999

Regd. Office: 7361, Ravindra Mansion, Ram Nagar, New Delhi 110 055

• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

Work Order No. AU/DDE/DE1-238/Preparation and Printing of Course Materials/2018 Dated 30.08.2018 Copies - 500

SYLLABI-BOOK MAPPING TABLE

Research Methods and Statistics

Syllabi	Mapping in Book
BLOCK I: INTRODUCTION TO RESEARCH, SCIENCE AND ITS CHARACTERISTICS, APPLICABILITY OF SCIENTIFIC CONDITION	
UNIT - I: Introduction to Research: Definition Scientific Research: Science and its Characteristics, Features.	Unit 1: Introduction to Research (Pages 1-7)
UNIT - II: Science and its Characteristics. Features, Purpose and Assumptions of Scientific Method. Steps in Scientific Method.	Unit 2: Science and its Characteristics (Pages 8-13)
UNIT - III: Applicability of Scientific Method to the Study of Social Phenomena. Theory and Research. Induction and Deduction.	Unit 3: Applicability of Scientific Method to the Study of Social Phenomena (Pages 14-20)
BLOCK II: RESEARCH PROBLEM, CONCEPTS AND REVIEW OF LITERATURE, HYPOTHESIS	
UNIT - IV: Research Problem: Formulation, Conditions and Considerations.	Unit 4: Research Problem (Pages 21-27)
UNIT - V: Concepts: Meaning, Categories, and Operationalization. Variables: Meaning, Types, and Measurement.	Unit 5: Concepts: Meaning, Categories and Operationalization (Pages 28-34)
UNIT - VI: Review of Literature: Scope and Purpose of Literature Review, Processes and Sources of Reviewing the Literature.	Unit 6: Review of Literature (Pages 35-42)
UNIT - VII: Hypothesis: Functions, Conditions for a Valid Hypothesis, Formulation of Hypothesis, Types and Forms of Hypothesis, Hypothesis Testing.	Unit 7: Hypothesis (Pages 43-66)
BLOCK III: RESEARCH DESIGN, SAMPLING, COLLECTION OF DATA	
UNIT - VIII: Research Design: Need for Research Design, Features. Types: Exploratory, Descriptive, Explanatory, Experimental and Evaluative.	Unit 8: Research Design (Pages 67-82)
UNIT - IX: Sampling: Census, Sample Survey, Characteristics and Implications of Sample Design, Sampling Criteria, Sampling Frame, Sampling Error.	Unit 9: Sampling (Pages 83-90)
UNIT - X: Types of Sampling: Probability and Non-Probability Sampling. Criteria for Selecting a Sampling Procedure.	Unit 10: Types of Sampling (Pages 91-114)
UNIT - XI: Collection of Data: Primary and Secondary Data, Sources of Secondary Data. Methods of Data Collection: Interview, Schedule, Questionnaire, Observation, Content Analysis and Case Study.	Unit 11: Collection of Data (Pages 115-144)

BLOCK IV: MEASUREMENT AND SCALING TECHNIQUES, MEASURE OF CENTRAL TENDENCY

UNIT - XII: Measurement and Scaling Techniques: Meaning, Need for Scales, Problems of Scaling, Methods of Scale Construction - Likert, Thurstone and Guttman Scales. Bogardus Scale. Reliability and Validity.

Unit 12: Measurement and Scaling Techniques
(Pages 145-175)

UNIT - XIII: Measures of Central Tendency: - Mean, Median, Mode- Measures of Dispersion: - Range, Quartile Deviation, Mean Deviation and Standard Deviation - Correlation Analysis: Karl Pearson's Coefficient of Correlation, Rank Correlation and Association of Attributes, Test of Significance.

Unit 13: Measures of Central Tendency, Dispersion and Correlation Analysis
(Pages 176-218)

BLOCK V: PREPARATION OF A RESEARCH REPORT

UNIT - XIV: Preparation of a Research Report: Format, Footnotes, Tables and Figures, Bibliography, Index, Editing and Evaluating the Final Report. Analysis of Data: Introduction, Importance, Scope, Function and Limitations.

Unit 14: Preparation of Research Report
(Pages 219-234)

CONTENTS

INTRODUCTION

BLOCK I: INTRODUCTION TO RESEARCH, SCIENCE AND ITS CHARACTERISTICS, APPLICABILITY OF SCIENTIFIC CONDITION

UNIT 1 INTRODUCTION TO RESEARCH 1-7

- 1.0 Introduction
- 1.1 Objectives
- 1.2 Meaning, Objectives and Significance
 - 1.2.1 Principles of Research
 - 1.2.2 Objectives of Research
 - 1.2.3 Research: Significance and Approach
 - 1.2.4 Methods versus Methodology
- 1.3 Answers to Check Your Progress Questions
- 1.4 Summary
- 1.5 Key Words
- 1.6 Self Assessment Questions and Exercises
- 1.7 Further Readings

UNIT 2 SCIENCE AND ITS CHARACTERISTICS 8-13

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Features, purpose and Assumptions
 - 2.2.1 Objectives of Scientific Inquiry
 - 2.2.2 Steps in Scientific Method
- 2.3 Answers to Check Your Progress Questions
- 2.4 Summary
- 2.5 Key Words
- 2.6 Self Assessment Questions and Exercises
- 2.7 Further Readings

UNIT 3 APPLICABILITY OF SCIENTIFIC METHOD TO THE STUDY OF SOCIAL PHENOMENA 14-20

- 3.0 Introduction
- 3.1 Objectives
- 3.2 Induction and Deduction
 - 3.2.1 Theory and Research
- 3.3 Answers to Check Your Progress Questions
- 3.4 Summary
- 3.5 Key Words
- 3.6 Self Assessment Questions and Exercises
- 3.7 Further Readings

**BLOCK II: RESEARCH PROBLEM, CONCEPTS AND REVIEW
OF LITERATURE, HYPOTHESIS**

UNIT 4 RESEARCH PROBLEM 21-27

- 4.0 Introduction
- 4.1 Objectives
- 4.2 Formulation, Conditions and Considerations
- 4.3 Answers to Check Your Progress Questions
- 4.4 Summary
- 4.5 Key Words
- 4.6 Self Assessment Questions and Exercises
- 4.7 Further Readings

**UNIT 5 CONCEPTS: MEANING, CATEGORIES AND
OPERATIONALIZATION 28-34**

- 5.0 Introduction
- 5.1 Objectives
- 5.2 Variables: Meaning, Types and Measurement
- 5.3 Answers to Check Your Progress Questions
- 5.4 Summary
- 5.5 Key Words
- 5.6 Self Assessment Questions and Exercises
- 5.7 Further Readings

UNIT 6 REVIEW OF LITERATURE 35-42

- 6.0 Introduction
- 6.1 Objectives
- 6.2 Scope and Purpose of Literature Review
 - 6.2.1 Processes and Sources of Reviewing the Literature
- 6.3 Answers to Check Your Progress Questions
- 6.4 Summary
- 6.5 Key Words
- 6.6 Self Assessment Questions and Exercises
- 6.7 Further Readings

UNIT 7 HYPOTHESIS 43-66

- 7.0 Introduction
- 7.1 Objectives
- 7.2 Functions, Conditions and Formulation
 - 7.2.1 Conditions for a Valid Hypothesis
 - 7.2.2 Functions of Hypotheses Formulation
- 7.3 Hypothesis Testing
 - 7.3.1 Types of Hypothesis Testing

- 7.4 Answers to Check Your Progress Questions
- 7.5 Summary
- 7.6 Key Words
- 7.7 Self Assessment Questions and Exercises
- 7.8 Further Readings

BLOCK III: RESEARCH DESIGN, SAMPLING COLLECTION OF DATA

UNIT 8 RESEARCH DESIGN 67-82

- 8.0 Introduction
- 8.1 Objectives
- 8.2 Need and Features
- 8.3 Types of Research Design
- 8.4 Answers to Check Your Progress Questions
- 8.5 Summary
- 8.6 Key Words
- 8.7 Self Assessment Questions and Exercises
- 8.8 Further Readings

UNIT 9 SAMPLING 83-90

- 9.0 Introduction
- 9.1 Objectives
- 9.2 Characteristics and Implications of Sample Design
 - 9.2.1 Sample vs Census
 - 9.2.2 Sampling vs Non-Sampling Error
- 9.3 Answers to Check Your Progress Questions
- 9.4 Summary
- 9.5 Key Words
- 9.6 Self Assessment Questions and Exercises
- 9.7 Further Readings

UNIT 10 TYPES OF SAMPLING 91-114

- 10.0 Introduction
- 10.1 Objectives
- 10.2 Probability Sampling
- 10.3 Non-Probability Sampling
- 10.4 Criteria for Selecting a Sampling Procedure
- 10.5 Answers to Check Your Progress Questions
- 10.6 Summary
- 10.7 Key Words
- 10.8 Self Assessment Questions and Exercises
- 10.9 Further Readings

UNIT 11 COLLECTION OF DATA

115-144

- 11.0 Introduction
- 11.1 Objectives
- 11.2 Primary and Secondary Data
- 11.3 Sources of Secondary Data
- 11.4 Methods of data Collection
- 11.5 Content Analysis
- 11.6 Case Study
- 11.7 Answers to Check Your Progress Questions
- 11.8 Summary
- 11.9 Key Words
- 11.10 Self Assessment Questions and Exercises
- 11.11 Further Readings

**BLOCK IV: MEASUREMENT AND SCALING TECHNIQUES MEASURE
OF CENTRAL TENDENCY**

UNIT 12 MEASUREMENT AND SCALING TECHNIQUES

145-175

- 12.0 Introduction
- 12.1 Objectives
- 12.2 Meaning, Problems and Methods
 - 12.2.1 Types of Measurement Scale
- 12.3 Methods of Scale Construction
 - 12.3.1 Criteria for Good Measurement: Reliability and Validity
 - 12.3.2 Thurstone, Guttman and Bogardus Scales
- 12.4 Answers to Check Your Progress Questions
- 12.5 Summary
- 12.6 Key Words
- 12.7 Self Assessment Questions and Exercises
- 12.8 Further Readings

**UNIT 13 MEASURES OF CENTRAL TENDENCY, DISPERSION
AND CORRELATION ANALYSIS**

176-218

- 13.0 Introduction
- 13.1 Objectives
- 13.2 Mean, Median Mode
- 13.3 Range, Quartile, Mean and Standard Deviation
- 13.4 Karl Pearson's Coefficient of Correlation
- 13.5 Rank Correlation and Attributes
- 13.6 Solved Problems
- 13.7 Answers to Check Your Progress Questions
- 13.8 Summary
- 13.9 Key Words
- 13.10 Self Assessment Questions and Exercises
- 13.11 Further Readings

BLOCK V: PREPARATION OF A RESEARCH REPORT

UNIT 14 PREPARATION OF RESEARCH REPORT

219-234

- 14.0 Introduction
- 14.1 Objectives
- 14.2 Research Report
 - 14.2.1 Types of Reports
 - 14.2.2 Characteristics of a Good Report
 - 14.2.3 Mechanics of Writing a Report
- 14.3 Analysis of Data
- 14.4 Answers to Check Your Progress Questions
- 14.5 Summary
- 14.6 Key Words
- 14.7 Self Assessment Questions and Exercises
- 14.8 Further Readings

INTRODUCTION

NOTES

Research is the quest for knowledge or a systematic investigation in order to establish facts. It helps to solve problems and 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.

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. This is followed by research design that may be experimental. 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, objective 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 and Statistics*, introduces to the students the meaning and basic concepts of research methodology. It deals with the techniques involved in defining a problem. It also explains the different kinds of research designs and the criteria for selecting a sampling procedure. In addition, the book also discusses the types of data and representation of data. Finally, the book deals with the interpretation and techniques of report writing.

The book has been written in keeping with the self-instructional mode or the SIM format wherein each Unit begins with an Introduction to the topic, followed by an outline of the Objectives. The detailed content is then presented in a simple and organized manner, interspersed with Check Your Progress questions to test the student's understanding of the topics covered. A Summary along with a list of Key Words and a set of Self-Assessment Questions and Exercises is provided at the end of each Unit for effective recapitulation.

BLOCK - I
INTRODUCTION TO RESEARCH, SCIENCE AND
ITS CHARACTERISTICS, APPLICABILITY OF
SCIENTIFIC CONDITION

NOTES

UNIT 1 INTRODUCTION TO
RESEARCH

Structure

- 1.0 Introduction
- 1.1 Objectives
- 1.2 Meaning, Objectives and Significance
 - 1.2.1 Principles of Research
 - 1.2.2 Objectives of Research
 - 1.2.3 Research: Significance and Approach
 - 1.2.4 Methods versus Methodology
- 1.3 Answers to Check Your Progress Questions
- 1.4 Summary
- 1.5 Key Words
- 1.6 Self Assessment Questions and Exercises
- 1.7 Further Readings

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.

Scientific research involves controlled observations, analysis of empirical data and interpretation of findings. This can further lead to the development of concepts, generalizations, etc., on the basis of which theories could be formulated. Such an investigation could help in determining cause and effect relationship. The ultimate aim of social science research is the control and prediction of behaviour.

NOTES

1.1 OBJECTIVES

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

- Differentiate between research methods and methodology
- Discuss the different approaches to research
- Describe the objectives of research
- Discuss the principles of research

1.2 MEANING, OBJECTIVES AND SIGNIFICANCE

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.' Redman and 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. 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. D. Slesinger and M. Stephenson, in the *Encyclopaedia of Social Sciences*, defined research as 'the manipulation 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.

1.2.1 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.

1.2.2 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.

1.2.3 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, etc. 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 like attitudes, interpersonal helping behaviour, environmental concerns like 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 problem is all require formal training in scientific methodology.

NOTES

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 the 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.

NOTES

- (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.

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.	<i>Library Research</i>	(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.	<i>Field Research</i>	(i) Non-participant direct observation (ii) Participant observation (iii) Mass observation	Observational behavioural scales, use of score cards, etc. Interactional recording, possible use of tape recorders, photographic techniques. Recording mass behaviour, interview using independent observers in public places. Identification of social and economic

3.	<i>Laboratory Research</i>	(iv) Mail questionnaire	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.
		Small group study of random behaviour, play and role analysis	Use of audio-visual recording devices, use of observers, etc.

NOTES**1.2.4 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, the 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 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 methodology also defines how the data should be evaluated to get the appropriate results.

Check Your Progress

1. On what type of data is a research based?
2. Name the emerging areas where HR research is being carried out.
3. Does the method of research change with the functional area?

1.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. A research is based on empirical data.
2. Critical success factor analysis and employer branding are some of the emerging areas where HR research is being carried out.
3. No, the method of research does not change with the functional area.

1.4 SUMMARY

- Research is done to find the solution to a problem, or to know more about something, or to know new things.
- Scientific research involves systematic, controlled, empirical and critical examination of a hypothesis or proposition about the relations in a phenomenon.
- The types of research are: descriptive, analytical, applied, fundamental, conceptual, empirical, quantitative and qualitative research.
- Research involves developing a scientific temper and logical thinking. The significance of research-based answers can never be underestimated.
- At the very beginning of research, the researcher must clearly define the research problem, i.e., the area of interest, the matter to be inquired into, etc.
- After conducting the research, the researcher has to prepare the report of what has been studied. Report must be written with great care.
- Interpretation of any research should be done keeping in mind the flaws in the procedural design and the extent to which it has an effect on the results.
- The validity and reliability of the data used in research should be double checked.
- The role of research is especially important in the fields of Economics, Business, Governance, etc. Here research helps in finding solutions to problems encountered in real life.

1.5 KEY WORDS

- **Quantitative approach:** In this approach, the data is in the form of quantities which is then subjected to mathematical and statistical approaches.
- **Qualitative approach:** It deals with data that cannot be strictly quantified, for example, opinions, tastes, and attitudes.

- **Social research:** It is conducted by social scientists in order to analyse a vast breadth of social phenomena.

1.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. Define research.
2. What are the objectives of research?
3. List the decision areas where research studies are carried out?

Long-Answer Questions

1. Give the objectives and significance of research. Also explain the significance of research in business decisions.
2. Write a detailed note on the approaches of research.

1.7 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

NOTES

UNIT 2 SCIENCE AND ITS CHARACTERISTICS

Structure

- 2.0 Introduction
- 2.1 Objectives
- 2.2 Features, purpose and Assumptions
 - 2.2.1 Objectives of Scientific Inquiry
 - 2.2.2 Steps in Scientific Method
- 2.3 Answers to Check Your Progress Questions
- 2.4 Summary
- 2.5 Key Words
- 2.6 Self Assessment Questions and Exercises
- 2.7 Further Readings

2.0 INTRODUCTION

Science is always a work in progress, and its conclusions are always tentative. But just as the word “theory” means something special to the scientist, so does the word “tentative.” Science’s conclusions are not tentative in the sense that they are temporary until the real answer comes along.

Scientific conclusions are well founded in their factual content and thinking and are tentative only in the sense that all ideas are open to scrutiny. In science, the tentativeness of ideas such as the nature of atoms, cells, stars or the history of the Earth refers to the willingness of scientists to modify their ideas as new evidence appears.

2.1 OBJECTIVES

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

- Describe the objectives of scientific inquiry
- Analyse the steps in scientific method
- Discuss the skills required in scientific enquiry
- Understand the characteristics of the scientific method

2.2 FEATURES, PURPOSE AND ASSUMPTIONS

Science refers to organized knowledge, but this knowledge and these facts are seldom conclusive. New experiences and additional information constantly change

previous findings and replace them with generalizations that confirm the latest bodies of findings.

A scientific enquiry is an investigation or experiment carried out to dispel or confirm various scientific theories. Most scientific enquiries are done practically in laboratories with specialized equipment.

The scientific method is based on techniques used to investigate phenomena, acquire new knowledge or correct and integrate previous knowledge. Any method is termed scientific when the inquiry is based on experiential and computable evidences subject to specific principles of reasoning. As per the *Oxford English Dictionary*, ‘*The scientific method is a method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses.*’

The key characteristic of the scientific method is that researchers can support a theory when the predictions given for any specific theory are confirmed and challenge a theory when its predictions prove false, even though procedures differ from one field of inquiry to another. Theories that include extensive domains of inquiry may combine many independently derived hypotheses together in a logical and supportive structure. Theories are developed on the basis of scientific inquiry and are normally intended to be objective so as to reduce biased interpretations of results. The overall process of theory development involves making assumptions by defining hypotheses and deriving predictions as logical consequences. The experiments are then carried out based on those defined predictions to establish whether the original assumption was correct. The scientific method steps are used to establish a theory.

2.2.1 Objectives of Scientific Inquiry

The objective of a scientific inquiry is to acquire knowledge in the form of testable explanations that can predict the results of future experiments. The more enhanced an explanation is at making predictions, the more beneficial it is in proving the predictions that it is correct. The most successful explanations that elucidate and formulate accurate predictions for broad range of conditions are termed as scientific theories. The power of a theory is related to how long it has persisted without distortion of its core principles.

Scientific Enquiry Skills

There are many scientific enquiry skills that must be observed in order to develop scientific theory. Some of which are as follows:

- Raising/asking questions
- Ways of enquiry
- Predicting and hypothesizing
- Making careful observations

NOTES

NOTES

- Using tools accurately and safely
- Making a record of evidence to present their findings
- Considering significant evidences
- Evaluating reliable evidences and findings accurate results
- Developing ideas from evidence

The same is the case with social sciences. The scientific method can also be applied to subjects in social sciences.

2.2.2 Steps in Scientific Method

The steps involved in scientific method are as follows:

- (i) Collection of data as per the problem at hand, according to some adequate plan and their systematic observation.
- (ii) Observations are made with a well defined purpose and they are recorded in definite terms.
- (iii) Classification and organization of data on the basis of similarities, variations, activities, causes and results.
- (iv) Generalization of data for the purpose of formulating principles and theories. The principles and theories must be specifically defined so that it can solve the problems in the related field.
- (v) Verification of generalizations through controlled experiments by tested prediction of results and by repetition of experiments. Correlation coefficient of original and verification of results is also calculated and probable errors are estimated. It is also determined whether the error lies in procedure or apparatus.
- (vi) Assumptions and limitations are noted down on the basis of verification of results.
- (vii) Reporting the research in detail.
- (viii) Announcement of the results before the general public for practical use.

Steps in Scientific Process

The steps involved in a scientific process are as follows:

- (i) **Purposeful Observation:** Observation should be accurate and extensive, and it must be done under various controlled conditions.
- (ii) **Analysis-Synthesis:** This include the following:
 - The essential elements in a problematic situation must be picked out by analysis.
 - Similarities as well as dissimilarities must be isolated.
 - Exceptions are to be given special attention.

- (iii) **Selective Recall:** A wide range of experiences is essential. These methods suffer from the normal error's caused due to poor memory of people (sample) and also from selective recall on the part of the individual. The memory follows a specific pattern to recall certain facts and may forget some other.
- (iv) **Hypothesis:** It is nothing but a tentative solution to the problem. There may be more than one solution depending on the nature of the problem.
- (v) **Verification by Inference and Experiment:** Here only one variable is manipulated and judgment is made on the adequacy and accuracy of data.

Redman and Mory define research as a 'Systematized effort to gain new knowledge'. According to Clifford Woody, *Research includes defining and redefining problems, formulating hypothesis or suggested solutions; collecting, organizing and evaluating data; and making deductions*'.

Check Your Progress

1. How is scientific method described in the oxford dictionary?
2. State the steps of analysis synthesis.

2.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. As per the *Oxford English Dictionary*, 'The scientific method is a method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses.'
2. Analysis synthesis includes the following:
 - The essential elements in a problematic situation must be picked out by analysis.
 - Similarities as well as dissimilarities must be isolated.
 - Exceptions are to be given special attention.

2.4 SUMMARY

- Science refers to organized knowledge, but this knowledge and these facts are seldom conclusive. New experiences and additional information constantly change previous findings and replace them with generalizations that confirm the latest bodies of findings.

NOTES

NOTES

- A scientific enquiry is an investigation or experiment carried out to dispel or confirm various scientific theories. Most scientific enquiries are done practically in laboratories with specialized equipment.
- The key characteristic of the scientific method is that researchers can support a theory when the predictions given for any specific theory are confirmed and challenge a theory when its predictions prove false, even though procedures differ from one field of inquiry to another.
- The overall process of theory development involves making assumptions by defining hypotheses and deriving predictions as logical consequences.
- The objective of a scientific inquiry is to acquire knowledge in the form of testable explanations that can predict the results of future experiments. The more enhanced an explanation is at making predictions, the more beneficial it is in proving the predictions that it is correct.
- There are many scientific enquiry skills that must be observed in order to develop scientific theory.

2.5 KEY WORDS

- **Selective recall:** The unconscious distortions involved in human recollection.
- **Scientific inquiry:** It refers to the diverse ways in which scientists study the natural world and propose explanations based on the evidence derived from their work.”

2.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. State the objectives of scientific enquiry.
2. What is the key characteristic if scientific method?

Long-Answer Questions

1. Analyse the skills required in scientific enquiry.
2. Describe the steps involved in scientific method and process.
3. Discuss the different steps in scientific process.

2.7 FURTHER READINGS

- Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.
- Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.
- Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.
- Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

NOTES

NOTES

UNIT 3 APPLICABILITY OF SCIENTIFIC METHOD TO THE STUDY OF SOCIAL PHENOMENA

Structure

- 3.0 Introduction
- 3.1 Objectives
- 3.2 Induction and Deduction
 - 3.2.1 Theory and Research
- 3.3 Answers to Check Your Progress Questions
- 3.4 Summary
- 3.5 Key Words
- 3.6 Self Assessment Questions and Exercises
- 3.7 Further Readings

3.0 INTRODUCTION

Scientific method is a systematic rational approach to seeking fact. It is objective, precise and arrives at conclusions on the basis of verifiable evidences. Hence research is systematic and logical study of an issue problem or phenomenon through scientific method.

This unit discusses the process of induction and deduction along with the theory and research that goes into scientific study.

3.1 OBJECTIVES

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

- Analyse the objectives of social science
- Differentiate between inductive and deductive method
- Discuss the concepts related to theory and research

3.2 INDUCTION AND DEDUCTION

The methods by which man from the earlier times sought answers to his problems can be classified under the following categories: (1) authority, (2) tradition, (3) experience, (4) deductive reasoning, (5) inductive reasoning, and (6) scientific method. These are also known as the objectives of social science.

Appeal to authority and seeking its advice was a well-established method of solving problems even in the earliest civilisations. We can find examples of reliance upon authority for truth, particularly during the ancient times, when floods, famines, or disease terrified man. He used to appeal to his elders and accepted their ancestral explanations for such incidents. At moments, he prayed to supernatural powers for help. During the Middle Ages, ancient scholars such as Plato and Aristotle, the early Fathers of Church, and others, were accepted as sources of truth than first-hand experience and analysis of facts. The modern man, sometime, also seeks advice from authorities for the solution of the problem faced by him. These authorities may be the persons who have had long experience with that problem and who have studied and thought much about it. In a court of law, for example, a judge may recognise a psychiatrist as an authority to testify the sanity of the defendant, or ask a handwriting specialist to compare signatures. When factual evidence cannot be obtained to solve a problem, one may have to rely upon authoritative opinion temporarily as the only possible method for solution. In such a situation, care must be employed in choosing authorities and evaluating their claims to knowledge. One should check not only the credentials of authorities but also the arguments and evidence upon which they base their judgements.

Closely related to authority is tradition, upon which man depends for solutions to many of his problems. He unquestioningly accepts many traditions of his forefathers or culture, such as the customary styles of dress, food, speech, and worship. In school settings, teachers often rely on tradition or past experiences as a dependable guide. Although automatic acceptance of tradition and custom is often necessary, one should not always assume that everything that has customarily been done is right and valid.

If we examine the historical records, we will find that many theories based upon tradition which prevailed for years were later found to be erroneous and had to be rejected. One should, therefore, evaluate custom and tradition carefully before he accepts them as truth.

Our own personal experience or that of others is the most primitive, and yet most familiar and fundamental, source of knowledge. In ancient times, nomads and various tribes from their personal experience probably remembered that certain wild fruits always made them ill, that grains ripened at particular times of the year, and that sudden floods in the rivers during the rainy season were due to the fact that water does not generally stay on hills. When confronted with a problem, modern man often tries to seek its answers from his own personal experience or from others who are familiar with the problem. Children often consult their teachers, parents or even their older siblings to derive answer to their questions.

NOTES

NOTES

Turning to personal experience or to that of others is a useful method to obtain knowledge, but its uncritical use may lead to incorrect conclusions. According to Van Dalen (1973, p. 5):

A person may make errors when observing or when reporting what he has seen or done. He may (1) omit evidence that does not agree with his opinion, (2) use measuring instruments that require many subjective estimates, (3) establish a belief on insufficient evidence, (4) fail to observe significant factors relating to a specific situation, or (5) draw improper conclusions or inferences owing to personal prejudices.

In the light of these remarks, one should cautiously and critically use experience as an avenue for obtaining knowledge.

A significant contribution towards the development of a systematic method for obtaining reliable knowledge was made by the ancient Greek philosophers like Aristotle and his followers. Aristotle developed the *sylogism*, which can be described as a thinking process in which one proceeds from general to specific statements by *deductive reasoning*. It provides a means of testing the validity of any given conclusion or idea by proceeding from the known to the unknown. The syllogistic reasoning consists of (1) a major premise based on a self-evident truth or previously established fact or relationship; (2) a minor premise concerning a particular case to which the truth, fact, or relationship invariably applies; and (3) a conclusion. If the major and minor premises can be shown to be true, the conclusion arrived at is necessarily true. To use a simple example, consider the following proposition:

1. All animals are mortal (Major Premise)
2. Dog is an animal (Minor Premise)
3. Therefore, dog will die (Conclusion)

The method of syllogism or deduction, however useful, has the following limitations:

1. The conclusion of a syllogism is always derived from the content of premises. Therefore, if the premises are unrelated or if one of the premises is erroneous, the conclusion arrived at will not be valid.
2. Another serious limitation of the deductive reasoning is its dependence upon verbal symbolism.
3. Deductive reasoning can systematise what is already known and can identify new relationships as one proceeds from known to unknown, but it cannot be relied upon as a self-sufficient method for securing reliable knowledge.

The conclusions derived from generalities and from statements of presumed authorities by deductive reasoning are true only if they are based upon true premises.

To determine whether the premises are true, Francis Bacon stressed the need for basing general conclusions upon specific facts gathered through direct observations. This is what is known as *inductive reasoning*, that is, going from the particular to the general. Rather than accepting premises laid down by authorities as absolute truths, Bacon advised man to observe nature closely, to experiment, to tabulate all the facts, to study these facts in order to reach minor generalizations, and then to proceed from minor generalizations to greater ones. He, however, cautioned against formulating any hypothesis or any probable solution to a problem until all the facts had been gathered.

In deductive reasoning, the premises or generalizations must be known before a conclusion can be reached. On the other hand, in inductive reasoning, a conclusion is reached by observing instances and generalizing from instances to the whole phenomenon. In order to be absolutely certain of an inductive conclusion, all instances must be observed. Under Baconian system of reasoning, it is known as perfect induction. In practical situations, however, it is not possible to examine every instance of a phenomenon to which a generalization refers. When examining all the instances of phenomenon under study is not practical, one may arrive at a generalization or theory by observing only some instances that make up the phenomenon. This is known as *imperfect induction*. Although imperfect induction does not help us to arrive at infallible conclusions, it can provide some knowledge upon which one can make reasonable decisions.

Both inductive and deductive methods when used independently of each other have limitations. If premises are true, deductive reasoning helps to arrive at absolutely true conclusions. These conclusions, however, do not probe beyond that, which is already known—already present, at least implicitly, in the premises. The use of inductive method solely does not help in providing a completely satisfactory way for the solution of problems. For instance, random collection of individual observations in the absence of a unifying concept is rarely helpful in drawing a generalization. Also, while studying a phenomenon the same set of observations can lead to different conclusions which may ultimately support different generalizations or theories. The conclusions reached by imperfect inductive reasoning do contain information that is not present, even implicitly, in one of the premises (the observed instances). If all the premises (observed instances) are true, the probability of conclusions arrived at may be of varying degrees.

The exclusive use of Bacon's inductive method resulted in the accumulation of isolated bits of information, and therefore, it made little contribution to the advancement of human knowledge. Moreover, many problems could not be solved by Aristotle's deductive method alone because in some situations, the acceptance of incomplete or false major premises, based on old dogmas or unreliable authority, could only lead to erroneous conclusions. In view of these limitations, it was superseded by the *deductive-inductive method*. This method, generally attributed

NOTES

NOTES

to Charles Darwin, integrates the most important aspects of the deductive and inductive methods which is now recognised as *scientific method*.

The scientific method is a back-and-forth movement of thought in which man first operates inductively from partially known or sometimes confused information learned from experience, previous knowledge, reflective thinking, observation and so on, towards a meaningful whole or hypothesis, and then deductively from suggested whole or hypothesis to the particular parts in order to connect these with one another in a meaningful pattern to find valid relationships. In the words of (Dewey 1933, p. 87):

While induction moves from fragmentary details (or particulars) to a connected view of situation (universal), deduction begins with the latter and works back again to particulars, connecting them and binding them together.

Although, in practice, scientific method involves a double movement of reasoning from induction to deduction, in its simplest form, it consists of working inductively from observations to hypotheses and then deductively from the hypotheses to the logical implications of the hypotheses in relation to what is already known.

3.2.1 Theory and Research

Theories are systematic statements that explain a particular segment of phenomenon by specifying certain relationship among variables.

Kerlinger has defined a theory as: ‘...*A set of interrelated constructs (concepts), definitions and propositions that present a systematic view of phenomena by specifying relationship among variables with the purpose of explaining and predicting the phenomena*’.

A theory can be explained on the following concepts:

- (i) Theory is a set of interrelated concepts, definitions and propositions.
- (ii) The interrelated concepts and definitions in a theory help us to understand the phenomena in a systematic manner.
- (iii) Theory establishes a relationship among various variables in a systematic manner. With the help of this relationship, we can predict the future nature of the phenomena.
- (iv) A theory helps us to formulate a hypothesis on the basis of which future research can be based.

Check Your Progress

1. State the methods by which man from the earlier times sought answers to his problems?
2. How has Kerlinger defined theory?

3.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The methods by which man from the earlier times sought answers to his problems can be classified under the following categories: (1) authority, (2) tradition, (3) experience, (4) deductive reasoning, (5) inductive reasoning, and (6) scientific method.
2. Kerlinger has defined a theory as: ‘...A set of interrelated constructs (concepts), definitions and propositions that present a systematic view of phenomena by specifying relationship among variables with the purpose of explaining and predicting the phenomena’.

NOTES

3.4 SUMMARY

- The methods by which man from the earlier times sought answers to his problems can be classified under the following categories: (1) authority, (2) tradition, (3) experience, (4) deductive reasoning, (5) inductive reasoning, and (6) scientific method.
- Appeal to authority and seeking its advice was a well-established method of solving problems even in the earliest civilisations.
- During the Middle Ages, ancient scholars such as Plato and Aristotle, the early Fathers of Church, and others, were accepted as sources of truth than first-hand experience and analysis of facts.
- Closely related to authority is tradition, upon which man depends for solutions to many of his problems. He unquestioningly accepts many traditions of his forefathers or culture, such as the customary styles of dress, food, speech, and worship.
- A significant contribution towards the development of a systematic method for obtaining reliable knowledge was made by the ancient Greek philosophers like Aristotle and his followers.
- In deductive reasoning, the premises or generalizations must be known before a conclusion can be reached. On the other hand, in inductive reasoning, a conclusion is reached by observing instances and generalizing from instances to the whole phenomenon.
- Theories are systematic statements that explain a particular segment of phenomenon by specifying certain relationship among variables.
- Kerlinger has defined a theory as: ‘...A set of interrelated constructs (concepts), definitions and propositions that present a systematic view of phenomena by specifying relationship among variables with the purpose of explaining and predicting the phenomena’.

NOTES

3.5 KEY WORDS

- **Imperfect induction:** The imperfect induction is the process of inferring from a sample of a group to what is characteristic of the whole group.
- **Deductive reasoning:** Deductive reasoning, also deductive logic, logical deduction is the process of reasoning from one or more statements (premises) to reach a logically certain conclusion.

3.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the objectives of social science?
2. State the limitations of the method of syllogism.

Long-Answer Questions

1. Discuss and differentiate between the methods of inductive and deductive reasoning.
2. Write a descriptive note on theory and research.

3.7 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

BLOCK - II
RESEARCH PROBLEM, CONCEPTS AND REVIEW
OF LITERATURE, HYPOTHESIS

Research Problem

NOTES

UNIT 4 RESEARCH PROBLEM

Structure

- 4.0 Introduction
- 4.1 Objectives
- 4.2 Formulation, Conditions and Considerations
- 4.3 Answers to Check Your Progress Questions
- 4.4 Summary
- 4.5 Key Words
- 4.6 Self Assessment Questions and Exercises
- 4.7 Further Readings

4.0 INTRODUCTION

A research problem is a statement about an area of concern, a condition to be improved, a difficulty to be eliminated, or a troubling question that exists in scholarly literature, in theory, or in practice that points to the need for meaningful understanding and deliberate investigation.

In this unit, we will describe the definition of research problems thereby analysing the various sources of it.

4.1 OBJECTIVES

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

- Define the concept of research problem
- Discuss the components of scientific thought
- Analyse the various sources of research problems
- Differentiate between decision problem and research problem

4.2 FORMULATION, CONDITIONS AND CONSIDERATIONS

All research in the field of behavioural sciences involves drawing inferences from a specified, identifiable group on the basis of a selected sample (formulation). The clearly identifiable and specified group is known as the population or universe.

*Self-Instructional
Material*

NOTES

The selected group of persons or objects is called the sample. The conclusions are drawn from the sample, which are deemed to be valid to the entire population. Such conclusions are known as the statistical inferences.

A population can be finite or infinite. A finite population is one where all the members can be counted. An infinite population is one where all the members cannot be counted (e.g., stars in the sky). A population can be imaginary or real.

A measure based upon the entire population is called a parameter. A measure based upon the sample is called a statistic.

A sample is a limited number or set of persons or elements that are chosen from a population, according to some plan. It is thought to represent the population. Samples are based on probabilities. Probability is a form of relative frequency. For example, the probability of seeing a head when coin is tossed once is $1/2$ or 0.5. Probability is expressed as a fraction or in decimal numbers.

Sources of Research Problems

A research problem is a discrepancy between what one knows and ought to know to solve a nursing problem. There are several sources of a research problem. The important ones are discussed as follows:

1. The 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. Any fault in this process can lead to a research problem. The 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:
 - o All well-executed projects have well-integrated teams. (Reason 1)
 - o The ABC project has many shortfalls. (Reason 2)
 - o 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?’ And 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 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 hypothesis and then collect more facts and reasons to deduct that the hypothesized conclusion is correct.

2. 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.

NOTES

NOTES

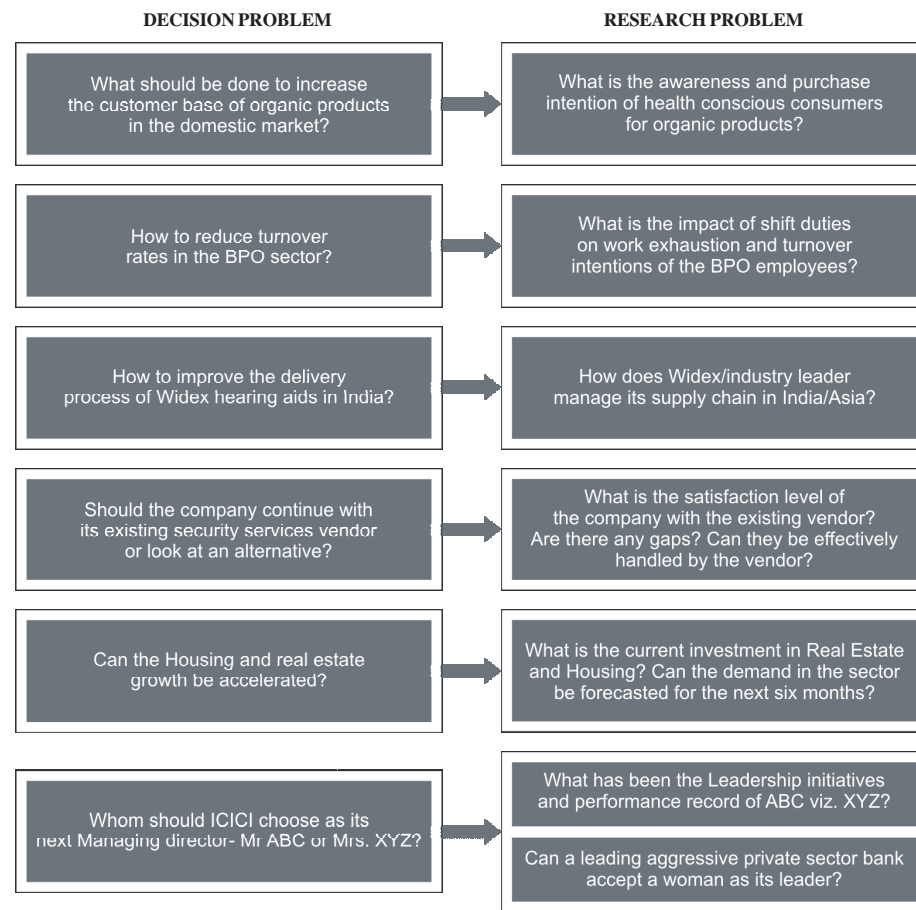


Fig. 4.1 Sources of a Research Problem

3. 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.

4. 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. These data however are 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.

5. 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.

Check Your Progress

1. State the process of review of existing literature.
2. Define research problem.

4.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. 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.
2. A research problem is a discrepancy between what one knows and ought to know to solve a nursing problem.

NOTES

NOTES

4.4 SUMMARY

- All research in the field of behavioural sciences involves drawing inferences from a specified, identifiable group on the basis of a selected sample. The clearly identifiable and specified group is known as the population or universe.
- A research problem is a discrepancy between what one knows and ought to know to solve a nursing problem. There are several sources of a research problem.
- 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.
- Inductive reasoning calls for generating a conclusion that is beyond the facts or information stated.
- In practice, scientific thought actually makes use of both inductive and deductive reasoning in a chronological order.
- 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.
- 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.

4.5 KEY WORDS

- **Applied research:** Applied research is a methodology used to solve a specific, practical problem of an individual or group.
- **Basic research:** Basic research, also called pure research or fundamental research, has the scientific research aim to improve scientific theories for improved understanding or prediction of natural or other phenomena.

4.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Differentiate between finite and infinite population.
2. State the sources of a research problem by the help of a chart.

Long-Answer Questions

1. Describe the concept of a research problem.
2. Analyse the various sources of research problems.

4.7 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

NOTES

NOTES

UNIT 5 CONCEPTS: MEANING, CATEGORIES AND OPERATIONALIZATION

Structure

- 5.0 Introduction
- 5.1 Objectives
- 5.2 Variables: Meaning, Types and Measurement
- 5.3 Answers to Check Your Progress Questions
- 5.4 Summary
- 5.5 Key Words
- 5.6 Self Assessment Questions and Exercises
- 5.7 Further Readings

5.0 INTRODUCTION

Operationalization is the process of strictly defining variables into measurable factors. For experimental research, where interval or ratio measurements are used, the scales are usually well defined and strict.

Operationalization sets down exact definitions of each variable, increasing the quality of the results, and improving the toughness of the design.

5.1 OBJECTIVES

After going through unit, you will be able to:

- Analyse the meaning of research variables
- Differentiate between dependent and independent variables
- Discuss the different categories of variables

5.2 VARIABLES: MEANING, TYPES AND MEASUREMENT

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, occupations can be: Teacher (1), Civil Servant (2), Private Sector Professional (3) and Self-employed (4). There are still others that possess an indefinite set, e.g., age, income and production data.

A dependent variable (DV) is measurable and quantifiable variable in nature. It is the most crucial variable to be analysed in a given research study.

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

NOTES

NOTES

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.

To understand the impact and role of the moderator variable let us represent the relationships graphically (Figure 5.1). Here a represents the effect of the independent variable (job satisfaction); b represents the effect of the second variable moderator variable (organizational commitment) and c represents the moderating effect, which is the combined effect of the moderating variable and the independent variable on the dependent variable. Thus, the effect of c has to be large enough and significant enough (statistically) to prove the moderation hypotheses.

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.

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.’

NOTES

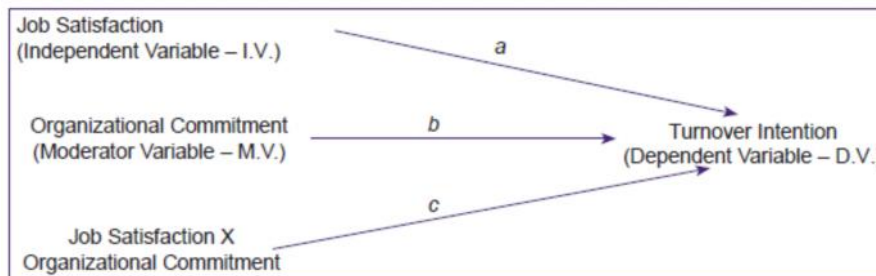
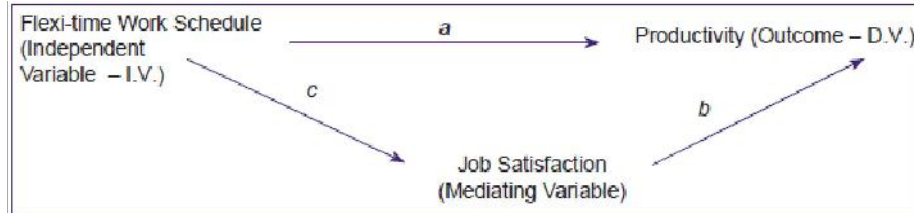


Fig. 5.1 Graphical Representation of Moderating Variable: Proposition 2

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) will result in increased brand awareness (IVV), which in turn will impact the first quarter sales (DV). This would be significantly higher amongst the younger female population (MV).



Note: b, c = indirect effect, a = direct effect

Fig. 5.2 Graphical Representation of Mediating Variable

In current research terminology, the intervening variable is also called a mediating variable, as it mediates the strength and direction of the relationship between the independent and dependent variable (Figure 5.2). For example in the above case, the direct effect of the predictor or the independent variable is measured by a ; and the mediating impact of the mediating variable is represented by b . However, the point to be noted is that the independent variable acts on the mediating variable as represented by c . Thus, to prove a mediating relationship, one would expect that the effect of b would be more than the effect of a and that this could be proven to be significantly significant. The best case of mediation would be if a was zero or the predictor had no direct effect on the outcome variable. The impact of the mediating variable is assessed by the method of structural equation modeling.

NOTES

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.

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.

Check Your Progress

1. What are discrete variables?
2. What is an intervening variable?

5.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Values that can only fit into prescribed number of categories are discrete variables.
2. An intervening variable (IVV) has a temporal connotation to it. It generally follows the occurrence of the independent variable and precedes the dependent variable.

5.4 SUMMARY

- 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 may be dichotomous in nature, that is, it can possess only two values such as male–female or customer–non-customer.

- A dependent variable (DV) is measurable and quantifiable variable in nature. It is the most crucial variable to be analysed in a given research study.
- 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).
- Moderating variables are the ones that have a strong contingent effect on the relationship between the independent and dependent 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.
- In current research terminology, the intervening variable is also called a mediating variable, as it mediates the strength and direction of the relationship between the independent and dependent variable.
- 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.

NOTES

5.5 KEY WORDS

- **Moderating variable:** A moderator variable, commonly denoted as just M, is a third variable that affects the strength of the relationship between a dependent and independent variable.
- **Mediating variable:** A mediator variable is the variable that causes mediation in the dependent and the independent variables. In other words, it explains the relationship between the dependent variable and the independent variable.

5.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. State certain situations wherein confusion might arise between a moderating variable and an independent variable.
2. Differentiate between the different kinds of variables in not more than a paragraph.

*Concepts: Meaning,
Categories and
Operationalization*

NOTES

Long-Answer Questions

1. Describe the concept of variables.
2. Discuss the different types of variables depending upon their role.

5.7 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

UNIT 6 REVIEW OF LITERATURE

Structure

- 6.0 Introduction
- 6.1 Objectives
- 6.2 Scope and Purpose of Literature Review
 - 6.2.1 Processes and Sources of Reviewing the Literature
- 6.3 Answers to Check Your Progress Questions
- 6.4 Summary
- 6.5 Key Words
- 6.6 Self Assessment Questions and Exercises
- 6.7 Further Readings

NOTES

6.0 INTRODUCTION

A literature review is an evaluative report of information found in the literature related to your selected area of study. The review should describe, summarise, evaluate and clarify this literature. It should give a theoretical base for the research and help determine the nature of your research.

This unit discusses the nature and scope of literature review. It also describes the various sources of reviewing literature.

6.1 OBJECTIVES

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

- Define the process of literature review
- Analyse the scope and purpose of literature review
- Describe the various sources of literature review

6.2 SCOPE AND PURPOSE OF LITERATURE REVIEW

Review of literature is the first task for a research, in order to decide on a specific problem for investigation. It also helps in formulating a theoretical framework for the entire study. It helps the researcher in formulating broader assumptions about the factors/variables involved in the problem, and the latter develops a hypothesis for the study. There are different sections in the report on review of literature. The first paragraph usually contains the reviewed areas of the related literature and outline of the presentation. While organizing a written report, the broad areas should be presented first followed by specific areas.

NOTES

In a thesis while presenting the review of research literature, a brief description of the research is given, including design, analysis and findings. The reviewer also emphasizes on what areas are explored and what is yet to be studied. It must keep in mind that the reviewed literature has to be critically analysed and summarized in terms of agreements and disagreements among authors and researchers. This is done in order to justify the necessity for conducting the investigation.

Sometimes non-research literature is also included in the review, but this mostly comprises theories and principles. The writer should acknowledge all sources of literature, using accepted form of presentation. At the end of the chapter, he may summarize the review presented.

A theoretical framework may follow the review of literature. It is important to bring out a comprehensive summary in the last paragraph of the literature. For an article, usually a paragraph or two are written on the related research review.

6.2.1 Processes and Sources of Reviewing the Literature

Methodology

This part of thesis-writing consists of the following:

- Description of research methodology
- Description of dependent and independent variables
- Sample defining the population and the sampling procedure, followed by selection of the sample
- Description of the tool and techniques used in the study
- Description of statistical techniques used in the analysis of data

The summary of the chapter is a necessity if the chapter is long. This is the third chapter of the thesis report. The organization of different section may vary on which comes first or second because there is no hard-and-fast rule about it. The writer uses his intuition and logic. At times, the use of figures becomes necessary to show the design or relation of variables.

If the researcher has used a standardized tool or an instrument prepared by some other researcher, he needs to take the author's (instrument) permission. There are tests that are confidential in nature, in which case the instrument is not given in the Appendix. This is also a major section in an article. A brief description of each of the sub-section is mentioned. Usually the content is presented under the heading of method or research design and covers at least 4–5 paragraphs in the article.

Data Analysis and Interpretation

This is one of the major chapters/sections which present: (a) Methods used for analysis, and (b) Findings of the study.

It is a logical development of analysis presented according to the objectives and hypothesis stated earlier. The most common methods of presentation of descriptive analysis are: (i) use of table and graphs to present data, (ii) statistical analysis for the test of significance by stating the null-hypothesis, and (iii) indicating the result of the test of significance.

Usually, interpretation of statistical analysis is done as the author presents each table or graph. Testing of hypothesis should indicate rejection or acceptance of null hypothesis and all its interpretation. At the end of analysis chapter, a summary of major findings are present following a discussion of findings. In a thesis or monograph, sometimes a chapter is written on discussion of the findings where the author compares the present result with findings of other studies indicating the similarities and differences. This may be presented in one or chapter 1 forming either the fourth chapter or the fourth and fifth chapters. Data analysis and discussion constitute the main body of the article. The article also includes essential tables and figures. Discussion is presented under a separate heading.

Summary, Conclusion, Implication and Recommendations

This part of the text should be a self contained summary of the whole report, containing a synopsis of essential background information, findings, conclusions and recommendations. Research steps, including a list of major findings, usually tables and figures, are not used.

All expected and unexpected findings and conclusions drawn from each of the findings are presented. Findings are statements of factual information, based upon the analysis of data. It also explains the extent to which generalization of results can be made. The researcher also mentions the reasons due to which the hypothesis tested is not found to be significant.

The implication indicates the author's reflective thinking, in terms of possible application of the result. For example, if the survey indicates that 70 per cent of pregnant women are anaemic, the implications may be written on the probable reasons for anaemia and what health care strategies can be adopted to improve the status. In other words, the implications suggest the values of these findings in terms of patient care for educational changes or the administrative strategies to be adopted.

Limitations of the present study are noted here. Limitations are those restrictions or a problem, which the researcher had not deliberately planned out but comes across while conducting the study.

The recommendations give direction to future research and suggestions for improving the present study. They should be specific and should not merely be vague statements. Recommendations indicate other aspects of the action suggested.

Besides a summary, an abstract is prepared (executive summary) which usually contains 500–1000 words.

NOTES

In an article, two or three paragraphs are written to discuss implications. A short summary is made, which usually works as a synopsis at the beginning of the article.

NOTES

Appendices, Bibliography, References

An *appendix* is the first of the terminal items presented at the end of the research report. A *bibliography* is a list of titles/books, research reports, articles, etc., that may or may not have been referred to in the text of the research report. *References* only include studies, books or papers that have actually been referred to, in the text of the research report. An approved style is to be adopted to write the references and bibliography.

Acknowledgement, Preface, Table of Content

Acknowledgement, table of content, list of tables and figures, are included in the first part of the research report. The monograph, in addition, also includes a page on preface. Articles do not require this section.

Outline of the Format of a Research Report

Research reports usually follow the structure given below or modified according to institution's specification.

Beginning

- Cover or title page
- Acknowledgement
- Table of contents
- List of tables
- List of figures and illustrations
- Glossary

The main body

- Introduction
- Review of literature
- Design of study
- Analysis and interpretation of data
- Major findings
- Conclusions and discussions
- Summary

Bibliography and references

References include all books/journals/reports, etc., referred or quoted by the author. A **bibliography** includes the entire literature source, surveyed and found relevant and useful, which may or may not have been quoted or referred to in the text.

Appendix

This section contains:

- Important correspondence, mainly with reference to permission for the study, subjects, willingness, request to experts, etc.
- Instrument; the final form of the tool and the key sheet for storing the master data sheet.
- Description of treatment variable.
- Any other important and relevant document that explains or brings clarity to the report.

Footnotes in-text references

Articles, papers, books, monographs, etc., quoted inside the text should always accompany relevant references, i.e., the author and the year of publication, e.g., (Kothari, 1988). If a few lines or sentences are actually quoted from a source, the page number too, should be noted, e.g., ‘Kothari, 1988: 120–124’. Besides, full reference should be placed in the ‘Reference’ section of the report. Usually, though traditional style of giving references is to place them as footnotes on the relevant page(s). The footnotes are serialized inside the text and in the footnotes of each chapter. These days, footnotes are usually avoided. However, they perform many functions. They provide ready reference on the page of the text itself, to avoid the tedious effort of consulting references at the end of the report, time and again. In certain cases, footnotes include explanatory statements, full form of abbreviations, extra justifications with reference to a portion of the text that may be read by a reader, if needed, i.e., if the text is not clearly understood. However, precision and necessity should be the main guidelines for these types of footnotes.

Precautions While Writing Research Thesis

A research thesis is a channel that communicates the results of a research to those who read it. A good research report performs this task in an efficient and effective manner. The points to be kept in mind while preparing a report are as follows:

- At the time of deciding the length of the report (since research reports vary greatly in length), it is important to bear in mind that it should be sufficiently long to cover the subject and equally short to capture interest.
- As far as possible, care should be taken to ensure that no research report is dull. It should have the ability to sustain the reader’s interest.

NOTES

NOTES

- A research report should not have non-figurative terms and technical jargon. The language of the report should be as simple as possible. This means that the style of the report should be objective and easy to understand, without expressions of uncertainty like ‘it seems’, ‘there may be’, etc.
- A reader would always prefer that the design of the report is such that he can get maximum information with minimum effort. This can be facilitated by using charts, graphs and statistical tables within the main report, in addition to the summary of important findings.
- The layout of the report should be planned well and should be suitable to the purpose of the research problem.
- There should be no grammatical errors in the report and it should follow the techniques of composition of report-writing, such as the use of quotations, footnotes, documentation, proper punctuation, abbreviations in footnotes, etc.
- The report must present a logical analysis of the subject matter. It must reflect a structure wherein the different pieces of analysis relating to the research problem fit well.
- A research report should be original and should essentially be directed at solving an intellectual problem. It must contribute to the solution of a problem and add to the store of knowledge.
- Towards the end, the report must also state the policy implications relating to the problem under consideration. It is usually considered desirable, if the report makes a forecast of the probable future of the subject concerned and indicates the kinds of research that still needs to be done in that particular field.
- Appendices should be enlisted with respect of all technical data in the report.
- Bibliography of sources consulted is a must for a good report and must necessarily be given.
- The index is also considered an essential part of a good report and as such must be prepared and appended at the end.
- A report must be attractive in appearance, neat and clean, whether typed or printed.
- Calculated confidence limits must be mentioned and various constraints experienced in conducting the research study may also be stated in the report.
- The objective of the study, nature of the problem, methods employed and the analysis techniques adopted must all be clearly stated in the beginning of the report, in the form of introduction.

Check Your Progress

1. What is considered to be the first task for a research?
2. What is a research thesis?

NOTES

6.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Review of literature is the first task for a research, in order to decide on a specific problem for investigation.
2. A research thesis is a channel that communicates the results of a research to those who read it. A good research report performs this task in an efficient and effective manner.

6.4 SUMMARY

- Review of literature is the first task for a research, in order to decide on a specific problem for investigation. It also helps in formulating a theoretical framework for the entire study.
- It helps the researcher in formulating broader assumptions about the factors/ variables involved in the problem, and the latter develops a hypothesis for the study.
- In a thesis while presenting the review of research literature, a brief description of the research is given, including design, analysis and findings. The reviewer also emphasizes on what areas are explored and what is yet to be studied.
- Sometimes non-research literature is also included in the review, but this mostly comprises theories and principles. The writer should acknowledge all sources of literature, using accepted form of presentation. At the end of the chapter, he may summarize the review presented.
- If the researcher has used a standardized tool or an instrument prepared by some other researcher, he needs to take the author's (instrument) permission.
- Usually, interpretation of statistical analysis is done as the author presents each table or graph. Testing of hypothesis should indicate rejection or acceptance of null hypothesis and all its interpretation.
- Acknowledgement, table of content, list of tables and figures, are included in the first part of the research report. The monograph, in addition, also includes a page on preface. Articles do not require this section.

NOTES

- Articles, papers, books, monographs, etc., quoted inside the text should always accompany relevant references, i.e., the author and the year of publication, e.g., (Kothari, 1988).
- A research thesis is a channel that communicates the results of a research to those who read it. A good research report performs this task in an efficient and effective manner.

6.5 KEY WORDS

- **Footnote:** An additional piece of information printed at the bottom of a page.
- **Appendix:** It is defined as the section at the end of a book that gives additional information on the topic explored in the contents of the text.

6.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. State the scope and purpose of literature review.
2. Briefly state the methodology of research writing.

Long-Answer Questions

1. Discuss the precautions to be taken while writing a research thesis.
2. Analyse the various sources and processes of reviewing literature.

6.7 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

UNIT 7 HYPOTHESIS

Structure

- 7.0 Introduction
- 7.1 Objectives
- 7.2 Functions, Conditions and Formulation
 - 7.2.1 Conditions for a Valid Hypothesis
 - 7.2.2 Functions of Hypotheses Formulation
- 7.3 Hypothesis Testing
 - 7.3.1 Types of Hypothesis Testing
- 7.4 Answers to Check Your Progress Questions
- 7.5 Summary
- 7.6 Key Words
- 7.7 Self Assessment Questions and Exercises
- 7.8 Further Readings

NOTES

7.0 INTRODUCTION

A working hypothesis is a tentative proposition that provides the solution, in the researcher's view, to the defined problem. After formulating a working hypothesis, the next step is to test the hypothesis. For testing the hypothesis, there are two kinds of tests, namely, parametric and non-parametric tests.

This unit will teach you how to choose the right type of test to test a hypothesis, depending upon the data.

7.1 OBJECTIVES

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

- Discuss the characteristics and conditions for a valid hypothesis
- Analyse the functions of hypothesis formulation
- Describe the different forms of hypothesis testing
- Discuss the advantages of hypothesis testing for comparing two related terms

7.2 FUNCTIONS, CONDITIONS AND FORMULATION

A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. It can contain either a suggested explanation for a phenomenon or a proposal having deductive reasoning to suggest a possible

interrelation between multiple phenomena. A deductive reasoning can be defined as a type of reasoning that can be derived from previously known facts.

7.2.1 Conditions for a Valid Hypothesis

NOTES

There are several condition of hypothesis, which are as follows:

- **Conceptually clear and accurate:** The hypothesis must be conceptually clear. The concepts and variables should be clearly defined operationally. The definition should use terms which are commonly accepted and communication is not hindered. 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 the different variables.
- **Testability:** A hypothesis should have empirical referents which mean that it should be testable through the empirical data. Hypothesis involving mystical or supernatural things are impossible to test. For example, the hypothesis 'education brings all-round development' is difficult to test because it is not easy to operationally isolate the other factors that might contribute towards all-round development. Since a hypothesis predicts the outcome of a study, it must relate variables that are capable of being measured. The hypothesis such as 'there is a positive relationship between the learning style and academic achievement of 8th grade students' can be tested since the variables in the hypothesis are operationally defined, and therefore can be measured.
- **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 a 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, 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 few more characteristics of a good hypothesis are as follows:

- It ensures that the sample is readily approachable.
- It maintains a very apparent distinction with what is called theory, law, facts, assumptions and postulates.
- It should have logical simplicity, a large number of consequences and be expressed in quantified form.

- It should have equal chances of confirmation and rejection.
- It permits the application of deduction reasoning.
- Tools and data should be easily available and effectively used.
- It should be based on study of previous literature and an existing theory, and should be verifiable.

As soon as a research question is formulated, it makes the hypothesis formulation imperative since a *hypothesis* is a tentative solution or an intelligent guess about a research question under study. It is an assumption or proposition whose tenability is to be tested on the basis of its implications with empirical evidence and previous knowledge. Modern investigators agree that, whenever possible, research should proceed from a hypothesis. In the words of Van Dalen (1973), '*a hypothesis serves as a powerful beacon that lights the way for the research worker*'.

7.2.2 Functions of Hypotheses Formulation

The reasons and functions for formulating a hypothesis are as follows:

1. A hypothesis directs, monitors and controls the research efforts. It provides tentative explanations of facts and phenomena and can be tested and validated. Such explanations, if held valid, lead to generalizations, which help significantly in understanding a problem. They thereby extend the existing knowledge in the area to which they pertain and thus help in theory building and facilitate the extension of knowledge in an area.
2. The hypothesis not only indicates what to look for in an investigation but also how to select a sample, choose the design of research, how to collect data and how to interpret the results to draw valid conclusions.
3. The hypothesis orients the researcher to be more sensitive to certain relevant aspects of the problem so as to focus on specific issues and pertinent facts. It helps the researcher to delimit his/her study in scope so that it does not become broad and unwieldy.
4. The hypothesis provides the researcher with rational statements, consisting of elements expressed in a logical order of relationships, which seek to describe or to explain conditions or events that have not yet been confirmed by facts. Some relationships between elements or variables in hypotheses are known facts, and others transcend the known facts to give reasonable explanations for known conditions. The hypothesis helps the researcher relate logically known facts to intelligent guesses about unknown conditions (Ary, *et al.*, 1972, pp. 73–74).
5. Hypothesis formulation and its testing add a scientific rigour to all type of researches. A well thought set of hypothesis places a clear and specific goal before the researcher and equips him/her with understanding. It provides the basis for reporting the conclusions of the study on the basis of these conclusions. The researcher can make the research report interesting and

NOTES

NOTES

meaningful to the reader. The importance of a hypothesis is generally recognized more in the studies which aim to make predictions about some outcome. In an experimental study, the researcher is interested in making predictions about the expected outcomes and, hence the hypothesis takes on a critical role. In the case of historical or descriptive studies, however, the researcher investigates the history of an event, or life of a man, or seeks facts in order to determine the *status quo* of a situation and hence may not have a basis for making a prediction of the results. In studies of this nature, where fact finding itself is the objective of the study, a hypothesis may not be required.

Most historical or descriptive studies involve fact finding as well as the interpretation of facts in order to draw generalizations. For all such major studies, a hypothesis is recommended so as to explain observed facts, conditions or behaviour and to serve as a guide in the research process. If a hypothesis is not formulated, a researcher may waste time and energy in gathering extensive empirical data, and then find that he/she cannot state facts clearly and detect relevant relationships between variables as there is no hypothesis to guide him/her.

7.3 HYPOTHESIS TESTING

Hypothesis testing means to determine whether or not the hypothesis is appropriate. This involves either accepting or rejecting a null hypothesis. The researcher has to pursue certain activities contained in the procedure of hypothesis.

In the formulation of hypothesis, the investigator looks for the statements where he/she relates one or more variables to make predictions about the relationships. The hypothesis tells the researcher what to do and why to do it in the context of the problem.

For example, the researcher is interested to study a problem, 'Why does a gifted child become a poor achiever in school'? The researcher then moves towards finding out the causes and factors that have been responsible for his/her poor achievement. He/She makes a conjecture that he/she might be suffering from some disease at the time of the examination. Conjecture is in the form of a hypothesis, and this now determines what the researcher should do to verify whether it is a fact or not. He/She shall go to the student's home, meet his/her parents and enquire about the student's health. All that the investigator is doing is guided by the hypothesis he/she had developed.

Thus, hypothesis refers to a conjecture statement about the solution to a problem, which the researcher goes on to verify on the basis of the relevant information collected by him/her. It is said to be a hunch, shrewd guess or supposition about what the answer to a problem may be. It is a statement which is tested in terms of the relationship or prediction, etc., which after testing is either accepted or rejected.

A hypothesis relates theory to observation and vice-versa. Hypotheses when tested are either rejected or accepted, and help to infer the conclusion, which helps in theory building. Being a specific statement of prediction, a hypothesis describes in concrete (rather than theoretical) terms what you expect will happen in your study. Not all studies have hypotheses. Sometimes a study is designed to be exploratory. In such researches, no formal hypothesis is established, and it may be the case that the actual objective of the study is to explore one or more specific areas more thoroughly in order to develop specific hypotheses or predictions that could be tested through research in the future. A single study could result in one or several hypotheses.

Some definitions of hypothesis are:

- According to Townsend, ‘Hypothesis is defined as suggested answer to a problem’.
- According to McGuigan, ‘A hypothesis is a testable statement of a potential relationship between two or more variables’.
- According to Uma Sekaran, ‘A hypothesis is defined as a logically conjectured relationship between two or more variables in the form of testable statement. These relationships are based on theoretical framework formulated for the research problem. The hypotheses are often statements about population parameters like expected value and variance, for example a hypothesis might be that the expected value of the height of 10-year-old boys in the Scottish population is not different from that of 10-year-old girls.’
- According to Kerlinger, ‘A good hypothesis is one which satisfies the following criteria:
 - o Hypothesis should state the relationship between variables.
 - o They must carry clear implications for testing the stated relations.’

This means that (a) statements contain two or more variables which can be measured, (b) they must state clearly how the two or more variables are related, and (c) it is important to note that facts and variables are not tested but relations between variables exist.

Sources of Hypothesis

Since the mind is fed by innumerable streams and sources, it is difficult to pinpoint how a particular good idea came to the researcher. The following are some of the popularly known sources of research hypothesis:

- **Scientific theories:** A systematic review and analysis of theories developed in the field of psychology, sociology, economics, political science and biological science may provide the researcher with potential clues for constructing a good and testable hypothesis.

NOTES

NOTES

- **Expert opinions:** Discussion with the experts in the field of research may further help the researcher obtain necessary insight and skill into the problem and in formulation of a hypothesis.
- **Method of related difference:** When we find that two phenomena differ constantly and the other circumstances remaining the same, we suspect a causal connection. For example, when we find more uncontrolled traffic in a locality, resulting in a greater number of road accidents, we suspect a causal connection between uncontrolled traffic and road accidents. This method also suggested a hypothesis.
- **Intellectual equipment of researcher:** Intellectual abilities of a researcher like creative thinking and problem solving techniques are very helpful in the formulation of a good hypothesis.
- **Related literature:** Related literature is the most important source of hypothesis formulation. A review of this literature may reveal to the researcher the variables that have been considered important in relation to his/her problem, which aspects have already been studied and which still remain to be studied, which theories have supported the relationships and which theories present a contradictory relationship. Familiarity with related literature may give the researcher a tremendous advantage in the construction of hypothesis.
- **Experience:** One's own experience may be a rich source of hypothesis generation. Personal experiences of an individual which has been gained through reading of biographies, autobiographies, newspaper readings or through informal talks among friends, etc., can be a potential source of generation of a hypothesis. For example, a researcher who is working on the effectiveness of guidance in teaching, can think of factors such as the teacher's polite behaviour, techniques of counselling, mastery over the subject, effective use of teaching skills, decision-making capability, perception of his/her competence, perception of student's capacity for better interaction, use of communication skills, etc.
- **Analogies:** Several hypotheses in a branch of knowledge may be made by using analogies from other sciences. Models and theories developed in a discipline may help, through extrapolation, in the formulation of hypothesis in another discipline. By comparing the two situations, analysing their similarities and differences, some rationale may emerge in the mind of the researcher which may take the form of a hypothesis for testing. For example, in a research problem like the studying the factors of unrest among college level students, the researcher insightfully thinks: 'Why was unrest found among school students? and What has changed them: quality of teaching or quality of leadership?'

Arguing analogically in this way may lead the investigator to some conclusions which may be used for identifying variables and relationships, which form

the basis of hypothesis construction. If a researcher knows from previous experience that the old situation is related to other factors Y and Z as well as to X, he/she may reason out that the new situation may also be related to Y and Z.

- **Methods of residues:** When the greater part of a complex phenomenon is explained by some causes already known, we try to explain the residual part of phenomenon according to the known law of operation. It also provides possible hypothesis.
- **Induction by simple enumeration:** Sometimes scientists take common experience as a starting point of their investigation. For example, after observing a large number of scarlet flowers that are devoid of fragrance, we frame a hypothesis that all scarlet flowers are devoid of fragrance. Thus induction by simple enumeration is a source of discovery.
- **Formulation of hypothesis:** It may also originate from the need and practice of present times.
- **Existing empirical uniformities:** In terms of common sense proposition, the existing empirical uniformities may form the basis for scientific examination.
- **A study of general culture:** It is also a good source of hypothesis.
- **Suggestions:** When given by other researchers in their reports, suggestions are quite helpful in establishment of hypothesis for future studies.

Procedure of Hypothesis Testing

The procedure for hypothesis testing is as follows:

- **Making formal statement:** In this step, the nature of a hypothesis is clearly stated, which could be either null hypothesis or alternate hypothesis. Stating a problem in hypothesis testing is of utmost importance, which should be done with proper care, keeping in mind the object and nature of the problem.
- **Choosing a significance level:** In this step, a hypothesis is tested on the basis of a present significance level, which has to be adequate in terms of nature and purpose of the problem.
- **Sampling distribution:** In this step, determination of an appropriate sampling distribution and making a choice between normal distribution and t -distribution is included.
- **Selection of a sample randomly:** In this step, a random sample is selected from the sample data for determining an apt value.
- **Probability calculation:** In this step, the probability regarding viability of the sample result is made dependent on the null hypothesis.
- **Comparison:** In this step, the calculated probability and the value of alpha in case of one-tailed test and alpha in case of two-tailed test is compared.

NOTES

NOTES

7.3.1 Types of Hypothesis Testing

Hypothesis is tested to identify the errors occurred in the statements and concepts used in hypothesis. Hypothesis testing can be broadly divided into two types, which are as follows:

- Parametric tests or standard tests of hypothesis
- Non-parametric tests or distribution-free tests of hypothesis

Parametric tests or standard tests of hypothesis

These kinds of tests assume certain properties of the population sample such as observations from a normal population, large sample size, population parameters like mean and variance. The various parametric tests of hypothesis are based on the assumption of normality. In other words, the source of data for them is normally distributed. They can be listed as follows:

- **Z-test:** This kind of test is based on normal probability distribution. It is mostly used to judge the significance of mean as a statistical measure. This is the most frequently used test in research studies. It is generally used to compare the mean of a sample with the hypothesized mean of the population. It is also used in case the population variance is known. It is helpful in judging the significance of difference between the means of two independent large samples, to compare the sample proportion to a theoretical value of population proportion and to judge the significance of median, mode and coefficient of correlation.
- **T-test:** This test is based on t -distribution and is aptly considered to judge the significance of a sample mean or the difference between the means of two small samples when population variance is not known.
- χ^2 : This test is based on a chi-square distribution and is used for comparing a sample variance to a theoretical population variance.
- **F-test:** This test is based on F-distribution and is also used to compare the variance of two independent samples. It is also used to compare the significance of multiple correlation coefficients.

Non-parametric tests or distribution-free tests of hypothesis

There are situations where assumptions cannot be made. In such situations, different statistical methods are used which are known as ‘non-parametric tests’. There are various types of non-parametric tests. The important non-parametric tests are as follows:

- **Sign test:** This is one of the easiest tests in practice based on the plus/minus sign of an observation in a sample. The sign may be one of the following two types:
 - o **One-sample sign test:** This is a very simple distribution-free test and is applied in case of a sample from a continuous symmetrical

population, wherein the probability of a sample to be either less or more than mean is half. Here, to test a null hypothesis, all those items which are greater than the alternate hypothesis are replaced by a plus sign and those which are less than the alternate hypothesis are replaced by a minus sign.

- o **Two-sample sign test:** In case of all the problems consisting of paired data, two-sample sign test is used. Here, each pair of values can be replaced with a plus sign in the first value of the first sample with the first value of the second sample. If the first value is less, minus sign is assigned.
- **Fisher-Irwin test:** This is applied where there is no difference between two sets of data. In other words, it is used where you can assume that two different treatments are supposedly different in terms of the results that they produce. It is applied in all those cases where result for each item in a sample can be divided into one of the two mutually exclusive categories.
- **McNamara test:** It is applied where the data is nominal in nature, and is related to two interrelated samples. By using this test, you can judge the significance of any observed changes in the same subject.
- **Wilcoxon matched-pairs test:** This test is applied in the case of a matched-pair such as output of two similar machines. Here, you can determine both the direction and the magnitude between the matched values. This test is also called Signed Rank Test.

Hypothesis Testing for Comparing Two Related Terms

Researchers often use hypothesis testing for comparing two population parameters based on the corresponding statistics from each population. For instance, researchers might want to check if the two populations have the same mean, which they can test with the help of hypothesis testing.

In this method two separate scores are to be obtained for each individual sample where the data in each sample set is related in some special way. For example, a group of patient's blood pressure is measured before and after a drug therapy. In this case, the same variable is measured two times for the same set of samples. Hypothesis testing uses t -statistic for comparing two related terms, which is described in the following sections:

t-Statistic for Comparing Two Related Terms

The t -statistic for comparing two related terms is based on the 'Difference Scores' and not on the 'Raw Scores'.

What is the Difference Scores?

Suppose a sample has $n = 4$ participants. Each individual's blood pressure is measured before and after medication.

NOTES

NOTES

Let,

X_1 = The first score for each person before medication.

X_2 = The second score for each person after medication.

Then, the difference scores are obtained by subtracting the first score from the second score for each participant.

Hence, difference score (d) = $X_2 - X_1$

The following figure illustrates the difference scores:

Subject	I	II	d
A	10	15	5
B	20	25	5
C	15	10	-5
D	25	30	5

The t-Statistic Formula for Comparing Two Related Terms

t -statistic formula for comparing two related terms is as follows:

$$t = \frac{M_d - \mu_d}{S_{M_d}}$$

Where,

M_d = The mean for the sample of difference scores or sample mean difference.

μ_d = The mean for the population of difference scores.

S_{M_d} = The standard error for M_d .

Decision

If the t value obtained falls in the critical region then reject the null hypothesis, otherwise do not reject it.

- M_d can be calculated as follows:

$$M_d = \frac{\sum d}{n}$$

Where, n = Sample size .

- Sample variance (S^2) of difference score (d) can be calculated as follows:

$$s^2 = \frac{SS}{n-1}$$

$$= \frac{SS}{df}$$

Where,

SS = It denotes sum of square of deviation.

df = It denoted degree of freedom.

Also, the formula to calculate SS is as follows:

$$SS = \sum d^2 - \frac{(\sum d)^2}{N}$$

Where, N = The size of the population.

- The sample standard deviation is then calculated as follows:

$$s = \sqrt{\frac{SS}{df}}$$

- Next, using sample variance, the estimated standard error is computed as follows:

$$S_{Md} = \sqrt{\frac{s^2}{n}} \text{ or } \frac{s}{\sqrt{n}}$$

Now, from the above computations, it is clear that all the calculations are done with the d scores, which is unique for each subject.

Degrees of Freedom

If there is a sample of n scores, then there will be total of nd scores. Degrees of freedom (df) describe the number of scores in a sample that are independent and can vary freely. The mean for the sample of difference scores places a restriction on the value of one sample, therefore the degrees of freedom will be $n - 1$ for n scores.

Basic Assumptions for t-Statistic

1. The sample size should not exceed 30.
2. Each observation within each treatment condition should be independent.

Example 7.1: For the following data set from a study of examining the effect of a treatment on college students by measuring a group of $n = 6$ subjects before and after they receive the treatment, find:

- The difference score.
- The sample means difference.
- Variance for difference scores.
- Standard error for the sample means difference.

NOTES

NOTES

<i>Subjects</i>	<i>Before Treatment</i>	<i>After Treatment</i>
I	7	8
II	2	9
III	4	6
IV	5	7
V	5	6
VI	3	8

Solution: (a) Let us denote,

X_2 = The second score for each participant after treatment

X_1 = The first score for each participant before treatment

Now, the difference score for each participant is given by: $d = X_2 - X_1$

The difference scores are calculated in the table given below:

<i>Subjects</i>	X_1	X_2	d
I	7	8	1
II	2	9	7
III	4	6	2
IV	5	7	2
V	5	6	1
VI	3	8	5

The required difference scores are: 1, 7, 2, 2, 2, 1 and 5.

(b) The sample mean difference of the score is given by:

$$\begin{aligned}
 M_D &= \frac{\sum d}{n} \\
 &= \frac{18}{6} \\
 &= 3
 \end{aligned}$$

The required sample means difference is 3.

(c) First, calculate SS by calculating d^2 , which is given in the table below:

<i>Subjects</i>	X_1	X_2	d	d^2
I	7	8	1	1
II	2	9	7	49
III	4	6	2	4
IV	5	7	2	4
V	5	6	1	1
VI	3	8	5	25

The formula of SS is given by:

$$\begin{aligned} SS &= \sum d^2 - \frac{(\sum d)^2}{N} \\ &= 84 - \frac{18^2}{6} \\ &= 30 \end{aligned}$$

The variance for the sample of difference scores is given by:

$$\begin{aligned} s^2 &= \frac{SS}{df} \\ &= \frac{30}{5} \\ &= 6 \end{aligned}$$

Hence, the required answer is 6.

(d) The standard error for the sample mean difference is given by:

$$\begin{aligned} S_{MD} &= \frac{s}{\sqrt{n}} \\ &= \frac{\sqrt{6}}{\sqrt{6}} \\ &= 1 \end{aligned}$$

Hence, the required answer is 1.

Every hypothesis test contains two opposite statements. One of the statements is null hypothesis and other is alternative hypothesis. These are the two types of hypothesis, which are described as follows:

Null Hypothesis

The null hypothesis states that the population parameter is equal to the claimed value and is denoted by H_0 . It is used for comparing statistics with the help of mean, μ . For example, if the average time taken by the student to complete his homework is 5 hours, then, $H_0: \mu = 5$.

Alternative Hypothesis

Before conducting the hypothesis the other possible hypothesis can also be treated vice versa. It is used for comparing statistics assuming that there is a difference between the two.

If the population parameter is *not equal* to the claimed value:

$$H_1: \mu \neq 5$$

NOTES

NOTES

If the population parameter is *greater than* the claimed value:

$$H_1: \mu > 5$$

If the population parameter is *less than* the claimed value:

$$H_1: \mu < 5$$

Example 7.2: Use the dataset of Example 7.1 to find whether there is any significant treatment effect. Use, $\alpha = 0.05$ for two tails.

Solution:

The null hypothesis is: $H_0: \mu_D = 0$

The mean difference is zero as there is no difference between the treatment conditions.

Against the alternative hypothesis 7, which is:

$$H_0: \mu_D \neq 0$$

There is a significant mean difference. Use, $\alpha = 0.05$ for two tails.

With a sample of $n = 6$, the t statistic has $df = 6 - 1 = 5$.

Hence, for a two tailed test with $\alpha = 0.05$ and $df = 5$, the critical t values are ± 2.571 .

Now, compute t -value using t -statistic formula:

$$\begin{aligned} t &= \frac{M_d - \mu_d}{S_{M_d}} \\ &= \frac{3 - 0}{1} \\ &= 3 \end{aligned}$$

The value $t = 3$ is greater than the critical value, so reject the null hypothesis H_0 .

Hence, it can be concluded that there is a significant mean difference.

Confidence Intervals (CI) for t-Statistics

Confidence interval gives an estimate about a range of values that are centered around the sample statistic. It is calculated by using sample mean so that it can be confidently estimated that the value of the parameter lies in the interval of the known population. The sample mean difference M_d is used to estimate the population mean difference μ_d .

Formula of CI for t-Statistics

For the t -test of comparing two related terms, $\mu_d = M_d \pm tS_{M_d}$

Where t stands for t -value and tS_{M_d} stands for standard error of mean differences.

Example 7.3: Use the dataset of Example 7.1 to construct a 95% CI.

Solution: Confidence level= 95%

Now, $\alpha = 0.05$ and $df = 5$, the critical t values are ± 2.571 .

Hence, the confidence interval is:

$$\begin{aligned}\mu_d &= M_d \pm tS_{M_d} \\ &= 3 \pm 2.571 \times 1 \\ &= 3 \pm 2.571\end{aligned}$$

The required confidence interval to estimate the sample mean difference is: (0.429, 5.571).

Advantages of Hypothesis Testing for Comparing Two Related Terms

The advantage of this study is that it removes individual differences, which lowers sample variability and increases the chances of obtaining significant results.

Hypothesis Testing of Proportions

Many a times, crucial decisions rely on the percentage or proportion of the population that meets certain predefined criteria. For example, a state's Chief Minister might be interested in knowing the percentage of females attending school in that state in order to come up with a policy decision to enhance female literacy. An economist might be interested in the proportion of the firms in an industry that make excessive profits and hence suggest the existence of an oligopolistic market structure, which is a situation where market is controlled by few sellers. A manager of a big enterprise may want to estimate the percentage of employees with an attendance rate of more than 90 per cent. The central point is that we might want to check if the population proportion exceeds or is less than some cut off value. Stated differently, hypothesis testing would allow us to check if the population proportion is significantly different from the hypothesized proportion, which is the one that we ideally desire for our data should possess.

Let us denote the population proportion by q in which we are interested in testing. Let r denote the sample proportion of observations that are considered as successes according to the defined rationale. So, if n is the sample size and X is the number of successes, then,

$$\rho = \frac{X}{n}$$

Let σ_ρ denote the standard error of the sampling proportion. It measures the tendency for the sample proportions, σ to deviate from the unknown population proportion, ρ . So,

$$\sigma_\rho = \sqrt{\frac{\rho(1-\rho)}{n}}$$

NOTES

Assuming that n is large, we can use the standard Z-test technique for hypothesis testing. Z is defined as,

$$Z = \frac{\rho - \theta}{\sigma_\rho}$$

NOTES

Note that this variable follows a standard normal distribution since it is in the form,

$$Z = \frac{\rho - E(\rho)}{\text{Standard error of } \rho}$$

Thus, we can use the standard normal tables to determine whether the calculated Z -value exceeds the Z -value at the given level of significance or not. Let us denote this critical Z -value by Z_α where α denotes the level of significance.

The general procedure for testing hypothesis can be thus outlined as follows:

- Set up the null hypothesis, that may be $\theta = \theta_H$ where θ_H is the hypothesized value of θ . This is what we want to test.
- Set up the alternative hypothesis which is complimentary to the null hypothesis. So, if the null is $\theta = \theta_H$, then the alternative is $\theta \neq \theta_H$.
- Choose the appropriate level of significance (α).
- Compute the relevant test statistic Z in this case.
- Find out the critical value (Z_α).
- Use the decision rule to accept or reject the null.

In this case, the decision rule is: If $|Z| > |Z_\alpha|$ then reject the null hypothesis, otherwise accept it.

Example 7.4: In a random sample of 500 people from a large population in a college, 200 are females. Is it correct to say that the sex ratio in this college is 1:1? Use level of significance as 1%.

Solution: We define success as the number of females in the sample. It is given that,

$$n = 500$$

$$X = 200$$

Since, the given ratio is 1:1, we define the null hypothesis as:

$$H_0: \theta = 0.5$$

$H_0: \theta = 0.5$, where θ is the proportion of females in the population.

The alternative hypothesis is therefore,

$$H_1: \theta \neq 0.5$$

The level of significance is given to be 0.01.

We compute,

$$\begin{aligned}\rho &= \frac{X}{n} \\ &= \frac{200}{500} \\ &= 0.4\end{aligned}$$

$$\begin{aligned}\sigma_{\rho} &= \sqrt{\frac{\rho(1-\rho)}{n}} \\ &= \sqrt{\frac{0.4 \times 0.6}{500}} \\ &= \sqrt{0.00048} \\ &= 0.02191\end{aligned}$$

So, the test statistic in this case is,

$$\begin{aligned}Z &= \frac{\rho - \theta}{\sigma_{\rho}} \\ &= \frac{0.4 - 0.5}{0.02191} \\ &= -4.564\end{aligned}$$

Note that it is a two tailed test. So,

$$Z_{0.01} = -2.58$$

Also,

$$\begin{aligned}|Z| &= 4.56 \\ |Z_{0.01}| &= 2.58\end{aligned}$$

Thus, $|Z| > |Z_{\alpha}|$. Hence, we reject the null hypothesis.

Therefore, we cannot claim that the sex ratio in the college is at level of significance 1%.

Hypothesis Testing for Differences between Proportions

In this method, one usually tests a claim made about two population proportions. The two estimated proportions may be different due to a difference in the populations. A hypothesis test helps in determining if there is a difference in the estimated proportions: $p_1 - p_2$ which reflects a difference in the population proportions.

In this section, we shall consider those tests only where the hypothesized difference between proportions is zero, since this is generally the case is.

NOTES

NOTES

Denoting our two population proportions as \hat{p}_1 and \hat{p}_2 , we can write the null hypothesis as

$H_0: p_1 - p_2 = 0$ which states that the two population proportions are equal.

Decision

If the Z-value obtained falls in the critical region then reject the null hypothesis, otherwise accept it.

Important Notations for the Two Population Proportions

Suppose, there are two populations: Population 1 and Population 2.

p_1 = Population Proportion

n_1 = size of the sample

x_1 = number of success in the sample

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ (sample proportion)}$$

$$\hat{q}_1 = 1 - \hat{p}_1$$

For Population 2, the corresponding notations are: p_2, n_2, x_2, \hat{p}_2 and \hat{q}_2 .

Formula for Pooled Sample Proportion

The notation of pooled sample proportion is \bar{p} and the formula is given by:

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\bar{q} = 1 - \bar{p}$$

Z-Statistic for Differences between Two Proportions

The Z-Statistic for differences between two proportions is based on the *Sample proportion* and the *Pooled sample proportion*.

The Z-Statistic formula for differences between two proportions is given by:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

Where,

$p_1 - p_2$ is assumed to be 0 (Discussed later under the null hypothesis)

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$\bar{q} = 1 - \bar{p}$$

$$\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}$$
 is the standard error.

P-Value

P-value is the probability of observing the sample statistic as extreme as the test statistic. Here the test statistic is a Z-statistic. The Standard Normal Distribution table is used to calculate the probability associated with the computed Z-statistic.

Basic Assumptions for Z-Statistic

- The two samples must be independent, that is, the two samples must be drawn from two different populations, so that the samples have no effect on each other.
- The samples must be large enough to use a normal sampling distribution and the difference of two population proportions should follow normal distribution approximately.
- The samples must be randomly selected.
- In both the samples, the number of successes as well as number of failures should be at least 5.

Computation of Z-Statistic Formula

Steps to calculate Z-Statistic formula are as follows:

Step 1: Calculate sample proportions: \hat{p}_1 and \hat{p}_2 .

Where,

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2}$$

Step 2: Calculate the difference between the two sample proportions: $\hat{p}_1 - \hat{p}_2$.

Step 3: Calculate the pooled sample proportion: $\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$.

Step 4: Calculate the standard error: $\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}$.

NOTES

Step 5: Divide the result of Step 2 by the result from Step 4. Then, the obtained test statistic is as follows:

NOTES

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}}$$

Example 7.5: Consider the following data:

	Sample 1	Sample 2
Rate	26.7%	29.0%
Total number	13200	13433

Find x_1 , x_2 , \bar{p} , \bar{q} .

Solution: Calculate x_1 in the following manner:

$$\begin{aligned} x_1 &= \frac{26.7 \times 13200}{100} \\ &= 3524 \text{ (Rounded Up)} \end{aligned}$$

Similarly,

$$\begin{aligned} x_2 &= \frac{29 \times 13433}{100} \\ &= 3896 \text{ (Rounded Up)} \end{aligned}$$

Now, calculate the pooled sample estimate \bar{p} as shown below:

$$\begin{aligned} \bar{p} &= \frac{x_1 + x_2}{n_1 + n_2} \\ &= \frac{3524 + 3896}{13200 + 13433} \\ &= 0.2786 \end{aligned}$$

And

$$\begin{aligned} \bar{q} &= 1 - 0.2786 \\ &= 0.7214 \end{aligned}$$

Example 7.6: Use the dataset of Example 7.1 to find whether there is any significant treatment effect. Use $\alpha = 0.01$ for one tail.

Solution: Now, the null hypothesis H_0 is,

$$H_0: p_1 = p_2 \text{ (Original claim of equality)}$$

The alternative hypothesis H_1 is,

$$H_1: p_1 > p_2$$

The significance level is $\alpha = 0.01$.

Now, calculate the value of 'Test Statistic' as follows:

$$\begin{aligned}
 z &= \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}\bar{q}}{n_1} + \frac{\bar{p}\bar{q}}{n_2}}} \\
 &= \frac{\left(\frac{3524}{13200} - \frac{3896}{13433}\right) - 0}{\sqrt{\frac{(0.2786)(0.7214)}{13200} + \frac{(0.2786)(0.7214)}{13433}}} \\
 &= -4.19
 \end{aligned}$$

The negative standard normal table gives p -value that is equal to 0.0001.

The p value is less than the level of significance 0.01, so, we reject the null hypothesis of $p_1 = p_2$.

Confidence Intervals (CI) for Z-Statistics

Confidence interval contains a range of values that are centered around the sample statistic. For this method, the confidence interval is computed to estimate the difference between two population proportions $p_1 - p_2$.

CI uses the standard deviation based on the estimated values of the population proportions, but the hypothesis testing method involves standard deviation based on the assumption that the two population proportions are equal.

Formula of CI for Z-Statistics

The margin of error is given by:

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

Where, $Z_{\alpha/2}$ is the critical value directly obtained from the standard normal table corresponding to confidence level.

The CI estimate is given by:

$$(\hat{p}_1 - \hat{p}_2) - E < (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E$$

Example 7.7: Consider the following data:

	Sample 1	Sample 2
Number of samples	126	205
Total number	331	331

Construct a 95% confidence level.

NOTES

Solution: With 95% confidence level, $Z_{\alpha/2} = 1.96$ from the standard normal table.

Calculate the margin of error E as follows:

NOTES

$$\begin{aligned}
 E &= z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} \\
 &= 1.96 \sqrt{\frac{\left(\frac{126}{331}\right)\left(\frac{205}{331}\right)}{331} + \frac{\left(\frac{205}{331}\right)\left(\frac{126}{331}\right)}{331}} \\
 &= 0.0739
 \end{aligned}$$

Hence, $E = 0.0739$

Where,

$$\hat{p}_1 = \frac{x_1}{n_1} \text{ and } \hat{p}_2 = \frac{x_2}{n_2} \text{ (Sample Proportion)}$$

$$\bar{q} = 1 - \bar{p}$$

p_i = Population Proportion

n_i = Size of the sample

x_i = Number of successes in the sample

Construct the 95% confidence Interval as follows:

$$\begin{aligned}
 (\hat{p}_1 - \hat{p}_2) - E &< (p_1 - p_2) < (\hat{p}_1 - \hat{p}_2) + E \\
 (0.3806 - 0.619) - 0.0739 &< (p_1 - p_2) < (0.3806 - 0.619) + 0.0739 \\
 -0.3123 &< (p_1 - p_2) < -0.1645
 \end{aligned}$$

Check Your Progress

1. What do you understand by a hypothesis testing?
2. Mention any three parametric tests for hypothesis.

7.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Hypothesis testing means to determine whether or not the hypothesis is appropriate. This involves either accepting or rejecting a null hypothesis.
2. Three parametric tests for hypothesis are Z-test, T-test and F-test.

7.5 SUMMARY

- A hypothesis is an approximate assumption that a researcher wants to test for its logical or empirical consequences. It can contain either a suggested explanation for a phenomenon or a proposal having deductive reasoning to suggest a possible interrelation between multiple phenomena.
- The hypothesis must be conceptually clear. The concepts and variables should be clearly defined operationally.
- A hypothesis, which is specific with limited scope, is easily testable than a hypothesis with limitless scope.
- As soon as a research question is formulated, it makes the hypothesis formulation imperative since a *hypothesis* is a tentative solution or an intelligent guess about a research question under study.
- A hypothesis directs, monitors and controls the research efforts. It provides tentative explanations of facts and phenomena and can be tested and validated.
- The hypothesis not only indicates what to look for in an investigation but also how to select a sample, choose the design of research, how to collect data and how to interpret the results to draw valid conclusions.
- Hypothesis testing means to determine whether or not the hypothesis is appropriate. This involves either accepting or rejecting a null hypothesis. The researcher has to pursue certain activities contained in the procedure of hypothesis.
- Hypothesis refers to a conjecture statement about the solution to a problem, which the researcher goes on to verify on the basis of the relevant information collected by him/her.
- ‘A hypothesis is defined as a logically conjectured relationship between two or more variables in the form of testable statement.’
- Researchers often use hypothesis testing for comparing two population parameters based on the corresponding statistics from each population. For instance, researchers might want to check if the two populations have the same mean, which they can test with the help of hypothesis testing.

NOTES

7.6 KEY WORDS

- **Hypothesis:** A hypothesis (plural hypotheses) is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it.

- **Probability:** Probability is the measure of the likelihood that an event will occur. Probability is quantified as a number between 0 and 1 (where 0 indicates impossibility and 1 indicates certainty).

NOTES

7.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Outline the general procedure for testing hypothesis,
2. Trace certain definitions of hypothesis.
3. State the procedure for hypothesis testing.

Long-Answer Questions

1. What are the major characteristics of a valid hypothesis?
2. Write a detailed note on the types of hypothesis testing.
3. Discuss the different sources of hypothesis.

7.8 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

BLOCK - III
RESEARCH DESIGN, SAMPLING
COLLECTION OF DATA

Research Design

NOTES

UNIT 8 RESEARCH DESIGN

Structure

- 8.0 Introduction
- 8.1 Objectives
- 8.2 Need and Features
- 8.3 Types of Research Design
- 8.4 Answers to Check Your Progress Questions
- 8.5 Summary
- 8.6 Key Words
- 8.7 Self Assessment Questions and Exercises
- 8.8 Further Readings

8.0 INTRODUCTION

It has been found by research scholars and managers alike that most research studies do not result in any significant findings because of a faulty research design. Most researchers feel that once the problem is defined and hypotheses are made, one can go ahead and collect the data on a specified group, or sample, and then analyse it using statistical tests. However, unless the formulated research problem and the study hypotheses is tested through a well-defined plan, answers are going to be based on hit and trial rather than any sound logic.

The design approaches available to the researcher are many and will depend on whether the study is of descriptive or conclusive nature. The designs range from very simple, loosely structured to highly scientific experimentation. In this unit, you will study the complete choice of designs, along with detailed reasoning on which design should be used under what conditions. Just as experiments in science, in business research also there are chances of error and this needs to be understood and controlled for more accurate results for the decision-maker.

8.1 OBJECTIVES

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

- Discuss the nature and classification of research designs
- Explain exploratory research designs
- Examine descriptive research designs
- Describe experimental designs

8.2 NEED AND FEATURES

NOTES

Once you have established the ‘what’ of the study, i.e., the research problem, the next step is the ‘how’ of the study, which specifies the method of achieving the research objectives. In other words, this is the research design.

Green *et al.* (2008) defines research design as ‘the specification of methods and procedures for acquiring the information needed. It is the overall operational pattern or framework of the project that stipulates what information is to be collected from which sources by what procedures. If it is a good design, it will ensure that the information obtained is relevant to the research questions and that it was collected by objective and economical procedures.’

Thyer (1993) states that, ‘A traditional research design is a blueprint or detailed plan for how a research study is to be completed—operationalizing variables so they can be measured, selecting a sample of interest to study, collecting data to be used as a basis for testing hypotheses, and analysing the results.’ Sellitz *et al.* (1962) state that, ‘A research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure.’

One of the most comprehensive and holistic definition has been given by Kerlinger (1995). He refers to a research design as, ‘..... a plan, structure and strategy of investigation so conceived as to obtain answers to research questions or problems. The plan is the complete scheme or programme of the research. It includes an outline of what the investigator will do from writing the hypotheses and their operational implications to the final analysis of data.’

Thus, the formulated design must ensure three basic principles:

- Convert the research question and the stated assumptions/hypotheses into variables that can be measured.
- Specify the process to complete the above task.
- Specify the ‘control mechanism(s)’ to follow so that the effect of other variables that could have an effect on the outcome of the study have been controlled.

At this stage, one needs to understand the difference between research design and research method. While the design is the specific framework that has been created to seek answers to the research question, the research method is the technique to collect the information required to answer the research problem, given the created framework. Thus, research designs have a critical and directive role to play in the research process. The execution details of the research question to be investigated are referred to as the research design.

The researcher has a number of designs available to him for investigating the research objectives. The classification that is universally followed is the one

based upon the objective or the purpose of the study. A simple classification that is based upon the research needs ranges from simple and loosely structured to the specific and more formally structured. The best way is to view the designs on a continuum as shown in Figure 8.1. Hence, in case the research objective is diffused and requires a refinement, one uses the exploratory design, and this might lead to the slightly more concrete descriptive design—here one describes all the aspects of the construct and concepts under study. This leads to a more structured and controlled experimental research design.

Figure 8.1 illustrates research designs as a continuous process.

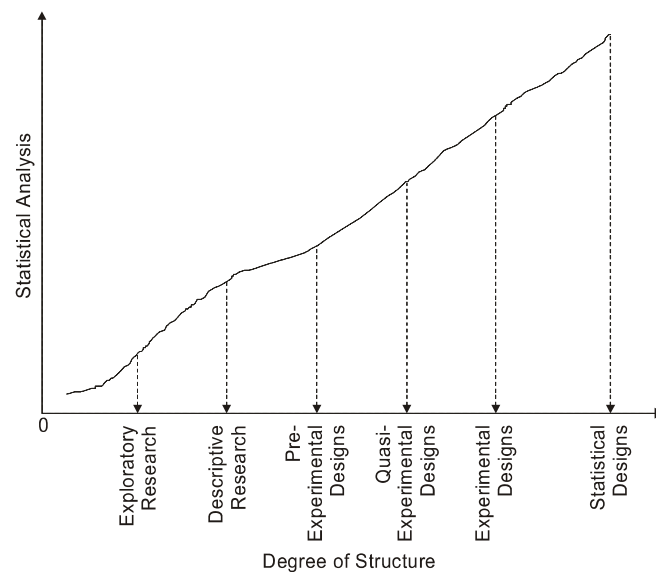


Fig. 8.1 *Research Designs—A Continuous Process*

Check Your Progress

1. List the three basic principles which should be covered while formulating a research design.
2. What is a research design?

8.3 TYPES OF RESEARCH DESIGN

Let us analyse the different types of research design.

Exploratory Research Designs

Exploratory designs, as stated earlier, are the simplest and most loosely structured designs. As the name suggests, the basic objective of the study is to explore and obtain clarity about the problem situation. It is flexible in its approach and mostly

NOTES

NOTES

involves a qualitative investigation. The sample size is not strictly representative and at times it might only involve unstructured interviews with a couple of subject experts. The essential purpose of the study is to assist in the following:

- Define and understand the research problem to be investigated.
- Explore and evaluate the diverse and multiple research opportunities.
- Assist in the development and formulation of the research hypotheses.
- Define the variables and constructs under study.
- Identify the possible nature of relationships that might exist between the variables under study.
- Explore the external factors and variables that might impact the research.

For example, a university professor might decide to do an exploratory analysis of the new channels of distribution that are being used by the marketers to promote and sell products and services. To do this, a structured and defined methodology might not be essential as the basic objective is to understand how to teach this to students of marketing. The researcher can make use of different methods and techniques in an exploratory research- like secondary data sources, unstructured or structured observations, expert interviews and focus group discussions with the concerned respondent group. Here, we will discuss them in brief in the light of their use in exploratory research.

Secondary Resource Analysis

Secondary sources of data, as the name suggests, are data in terms of the details of previously collected findings in facts and figures—which have been authenticated and published. It is a fast and inexpensive way of collecting information. The past details can sometimes point out to the researcher that his proposed research is redundant and has already been established earlier. Secondly, the researcher might find that a small but significant aspect of the concept has not been addressed and should be studied. For example, a marketer might have extensively studied the potential of the different channels of communication for promoting a ‘home maintenance service’ in Greater Mumbai. However, there is no impact of any mix that he has tested. An anthropologist research associate, on going through the findings, postulated the need for studying the potential of WOM (word of mouth) in a close-knit and predominantly Parsi colony where this might be the most effective culture-dependent technique that would work. Thus, such insights might provide leads for carrying out an experimental and conclusive research subsequently.

Another valuable secondary resource is the compiled and readily available databases of the entire industry, business or construct. These might be available on free and public domains or through a structured acquisition process and cost. These are both government and non-government publications. Based on the resources and the level of accuracy required, the researcher might decide to make use of them.

Case Study Method

Another way of conducting an exploratory research is the case study method. This requires an in-depth study and is focused on a single unit of analysis. This unit could be an employee or a customer; an organization or a complete country analysis. They are by their nature, generally, post-hoc studies and report those incidences which might have occurred earlier. The scenario is reproduced based upon the secondary information and a primary interview/discussion with those involved in the occurrence. Thus, there might be an element of bias as the data, in most cases, becomes a judgemental analysis rather than a simple recounting of events.

For example, BCA Corporation wants to implement a performance appraisal system in the organization and is debating between the merits of a traditional appraisal system and a 360° appraisal system. For a historical understanding of the two techniques, the HR director makes use of books on the subject. However, for better understanding, he should do an in-depth case accounting of Allied Association which had implemented traditional appraisal formats, and Surakhsha International which uses 360° appraisal systems. Thus, the two exploratory researches carried out were sufficient to arrive at a decision in terms of what would be best for the organization.

Expert Opinion Survey

At times, there might be a situation when the topic of a research is such that there is no previous information available on it. In these cases, it is advisable to seek help from experts who might be able to provide some valuable insights based upon their experience in the field or with the concept. This approach of collecting particulars from significant and knowledgeable people is referred to as the expert opinion survey. This methodology might be formal and structured and is useful when authenticated or supported by a secondary/primary research or it might be fluid and unstructured and might require an in-depth interviewing of the expert. For example, the evaluation of the merit of marketing organic food products in the domestic Indian market cannot be done with the help of secondary data as no such structured data sources exist. In this case the following can be contacted:

- Doctors and dieticians as experts would be able to provide information whether consumers would eat organic food products as a healthier alternative.
- Chefs who are experimental and would like to look at providing better value to their clients.
- Retailers who like to sell contemporary new products.

These could be useful in measuring the viability of the proposed plan. Discussions with knowledgeable people may reveal some information regarding who might be considered as potential consumers. Secondly, the question whether a healthy proposition or a lifestyle proposition would work better to capture the

NOTES

targeted consumers' needs to be examined. Thus, this method can play a directional role in shaping the research study.

Focus Group Discussions

NOTES

Another way to conduct an exploratory analysis is carry out discussions with individuals associated with the problem under study. This technique, though originally from sociology, is actively used in business research. In a typical focus group, there is a carefully selected small set of individuals' representative of the larger respondent population under study. It is called a focus group as the selected members discuss the concerned topic for the duration of 90 minutes to, sometimes, two hours. Usually the group is made up of six to ten individuals. The number thus stated is because less than six would not be able to throw enough perspectives for the discussion and there might emerge a one-sided discussion on the topic. On the other hand, more than ten might lead to more confusion rather than any fruitful discussion and that would be unwieldy to manage. Generally, these discussions are carried out in neutral settings by a trained observer, also referred to as the moderator. The moderator, in most cases, does not participate in the discussion. His prime objective is to manage a relatively non-structured and informal discussion. He initiates the process and then manoeuvres it to steer it only to the desired information needs. Sometimes, there is more than one observer to record the verbal and non-verbal content of the discussion. The conduction and recording of the dialogue requires considerable skill and behavioural understanding and the management of group dynamics. In the organic food product study, the focus group discussions were carried out with the typical consumers/buyers of grocery products. The objective was to establish the level of awareness about health hazards, environmental concerns and awareness of organic food products. A series of such focus group discussions carried out across four metros—Delhi, Mumbai, Bengaluru and Hyderabad—revealed that even though the new age consumer was concerned about health, the awareness about organic products varied from extremely low to non-existent. (*This study was carried out in the year 2004–05 by one of the authors for an NGO located in Delhi.*)

Descriptive Research Designs

As the name implies, the objective of descriptive research studies is to provide a comprehensive and detailed explanation of the phenomena under study. The intended objective might be to give a detailed sketch or profile of the respondent population being studied. For example, to design an advertising and sales promotion campaign for high-end watches, a marketer would require a holistic profile of the population that buys such luxury products. Thus a descriptive study, (which generates data on *who, what, when, where, why* and *how* of luxury accessory brand purchase) would be the design necessary to fulfil the research objectives.

Descriptive research, thus, are conclusive studies. However, they lack the precision and accuracy of experimental designs, yet it lends itself to a wide range of situations and is more frequently used in business research. Based on the time

period of the collection of the research information, descriptive research is further subdivided into two categories: cross-sectional studies and longitudinal studies.

Cross-sectional Studies

As the name suggests, cross-sectional studies involve a slice of the population. Just as in scientific experiments one takes a cross-section of the leaf or the cheek cells to study the cell structure under the microscope, similarly one takes a current subdivision of the population and studies the nature of the relevant variables being investigated.

There are two essential characteristics of cross-sectional studies:

- The cross-sectional study is carried out at a single moment in time and thus the applicability is most relevant for a specific period. For example, one cross-sectional study was conducted in 2002 to study the attitude of Americans towards Asian-Americans, after the 9/11 terrorist attack. This revealed the mistrust towards Asians. Another cross-sectional study conducted in 2012 to study the attitude of Americans towards Asian-Americans revealed more acceptance and less mistrust. Thus the cross-sectional studies cannot be used interchangeably. .
- Secondly, these studies are carried out on a section of respondents from the population units under study (e.g., organizational employees, voters, consumers, industry sectors). This sample is under consideration and under investigation only for the time coordinate of the study.

There are also situations in which the population being studied is not of a homogeneous nature but composed of different groups. Thus it becomes essential to study the sub-segments independently. This variation of the design is termed as *multiple cross-sectional studies*. Usually this multi-sample analysis is carried out at the same moment in time. However, there might be instances when the data is obtained from different samples at different time intervals and then they are compared. *Cohort analysis* is the name given to such cross-sectional surveys conducted on different sample groups at different time intervals. Cohorts are essentially groups of people who share a time zone or have experienced an event that took place at a particular time period. For example, in the post-9/11 cross-sectional study done in 2002, we study and compare the attitudes of middle-aged Americans versus teenaged Americans towards Asian-Americans. These two American groups are separate cohorts and this would be a cohort analysis. Thus the teenage American is one cohort and the middle-aged cohort is separate and thinks differently.

The technique is especially useful in predicting election results, cohorts of males–females, different religious sects, urban–rural or region-wise cohorts are studied by leading opinion poll experts like Nielsen, Gallup and others. Thus, Cross-sectionals studies are extremely useful to study current patterns of behaviour or opinion.

NOTES

NOTES

Longitudinal Studies

A single sample of the identified population that is studied over a longer period of time is termed as a longitudinal study design. A panel of consumers specifically chosen to study their grocery purchase pattern is an example of a longitudinal design. There are certain distinguishing features of the same:

- The study involves the selection of a representative panel, or a group of individuals that typically represent the population under study.
- The second feature involves the repeated measurement of the group over fixed intervals of time. This measurement is specifically made for the variables under study.
- A distinguishing and mandatory feature of the design is that once the sample is selected, it needs to stay constant over the period of the study. That means the number of panel members has to be the same. Thus, in case a panel member due to some reason leaves the panel, it is critical to replace him/her with a representative member from the population under study.

Longitudinal study using the same section of respondents thus provides more accurate data than one using a series of different samples. These kinds of panels are defined as true panels and the ones using a different group every time are called omnibus panels. The advantages of a true panel are that it has a more committed sample group that is likely to tolerate extended or long data collecting sessions. Secondly, the profile information is a one-time task and need not be collected every time. Thus, a useful respondent time can be spent on collecting some research-specific information.

However, the problem is getting a committed group of people for the entire study period. Secondly, there is an element of mortality and attrition where the members of the panel might leave midway and the replaced new recruits might be vastly different and could skew the results in an absolutely different direction. A third disadvantage is the highly structured study situation which might be responsible for a consistent and structured behaviour, which might not be the case in the real or field conditions.

Experimental Designs

Experimental designs are conducted to infer causality. In an experiment, a researcher actively manipulates one or more causal variables and measures their effects on the dependent variables of interest. Since any changes in the dependent variable may be caused by a number of other variables, the relationship between cause and effect often tends to be probabilistic in nature. It is virtually impossible to prove a causality. One can only infer a cause-and-effect relationship.

The necessary conditions for making causal inferences are: (i) concomitant variation, (ii) time order of occurrence of variables and (iii) absence of other possible causal factors. The first condition implies that cause and effect variables should

have a high correlation. The second condition means that causal variable must occur prior to or simultaneously with the effect variable. The third condition means that all other variable except the one whose influence we are trying to study should be absent or kept constant.

There are two conditions that should be satisfied while conducting an experiment. These are:

- (i) **Internal validity:** Internal validity tries to examine whether the observed effect on a dependent variable is actually caused by the treatments (independent variables) in question. For an experiment to be possessing internal validity, all the other causal factors except the one whose influence is being examined should be absent. Control of extraneous variables is a necessary condition for inferring causality. Without internal validity, the experiment gets confounded.
- (ii) **External validity:** External validity refers to the generalization of the results of an experiment. The concern is whether the result of an experiment can be generalized beyond the experimental situations. If it is possible to generalize the results, then to what population, settings, times, independent variables and the dependent variables can the results be projected. It is desired to have an experiment that is valid both internally and externally.

However, in reality, a researcher might have to make a trade-off between one type of validity for another. To remove the influence of an extraneous variable, a researcher may set up an experiment with artificial setting, thereby increasing its internal validity. However, in the process the external validity will be reduced.

There are four types of experimental designs. These are explained below:

1. **Pre-experimental designs:** There are three designs under this – one short case study where observation is taken after the application of treatment, one group pre test-post test design where one observation is taken prior to the application of treatment and the other one after the application of treatment, and static group comparison, where there are two groups – experimental group and control group. The experiment group is subjected to treatment and a post-test measurement is taken. In the control group measurement is taken at the time when it was done for experimental group. These do not make use of any randomization procedures to control the extraneous variables. Therefore, the internal validity of such designs is questionable.
2. **Quasi-experimental designs:** In these designs, the researcher can control when measurements are taken and on whom they are taken. However, this design lacks complete control of scheduling of treatment and also lacks the ability to randomize test units' exposure to treatments. As the experimental control is lacking, the possibility of getting confounded results is very high.

NOTES

NOTES

Therefore, the researchers should be aware of what variables are not controlled and the effects of such variables should be incorporated into the findings.

3. **True experimental designs:** In these designs, researchers can randomly assign test units and treatments to an experimental group. Here, the researcher is able to eliminate the effect of extraneous variables from both the experimental and control group. Randomization procedure allows the researcher the use of statistical techniques for analysing the experimental results.
4. **Statistical designs:** These designs allow for statistical control and analysis of external variables. The main advantages of statistical design are the following:
 - The effect of more than one level of independent variable on the dependent variable can be manipulated.
 - The effect of more than one independent variable can be examined.
 - The effect of specific extraneous variable can be controlled.

Statistical design includes the following designs:

- (i) *Completely randomized design:* This design is used when a researcher is investigating the effect of one independent variable on the dependent variable. The independent variable is required to be measured in nominal scale i.e. it should have a number of categories. Each of the categories of the independent variable is considered as the treatment. The basic assumption of this design is that there are no differences in the test units. All the test units are treated alike and randomly assigned to the test groups. This means that there are no extraneous variables that could influence the outcome.

Suppose we know that the sales of a product is influenced by the price level. In this case, sales are a dependent variable and the price is the independent variable. Let there be three levels of price, namely, low, medium and high. We wish to determine the most effective price level i.e. at which price level the sale is highest. Here, the test units are the stores which are randomly assigned to the three treatment level. The average sales for each price level is computed and examined to see whether there is any significant difference in the sale at various price levels. The statistical technique to test for such a difference is called analysis of variance (ANOVA).

The main limitation of completely randomized designs is that it does not take into account the effect of extraneous variables on the dependent variable. The possible extraneous variables in the present example could be the size of the store, the competitor's price and price of the substitute product in question. This design assumes that all the

extraneous factors have the same influence on all the test units which may not be true in reality. This design is very simple and inexpensive to conduct.

- (ii) *Randomized block design*: As discussed, the main limitation of the completely randomized design is that all extraneous variables were assumed to be constant over all the treatment groups. This may not be true. There may be extraneous variables influencing the dependent variable. In the randomized block design it is possible to separate the influence of one extraneous variable on a particular dependent variable, thereby providing a clear picture of the impact of treatment on test units.

In the example considered in the completely randomized design, the price level (low, medium and high) was considered as an independent variable and all the test units (stores) were assumed to be more or less equal. However, all stores may not be of the same size and, therefore, can be classified as small, medium and large size stores. In this design, the extraneous variable, like the size of the store could be treated as different blocks. Now the treatments are randomly assigned to the blocks in such a way so that each treatment appears in each block at least once. The purpose of forming these blocks is that it is hoped that the scores of the test units within each block would be more or less homogeneous when the treatment is absent. What is assumed here is that block (size of the store) is correlated with the dependent variable (sales). It may be noted that blocking is done prior to the application of the treatment.

In this experiment one might randomly assign 12 small-sized stores to three price levels in such a way that there are four stores for each of the three price levels. Similarly, 12 medium-sized stores and 12 large-sized stores may be randomly assigned to three price levels. Now the technique of analysis of variance could be employed to analyse the effect of treatment on the dependent variable and to separate out the influence of extraneous variable (size of store) from the experiment.

- (iii) *Factorial design*: A factorial design may be employed to measure the effect of two or more independent variables at various levels. The factorial designs allow for interaction between the variables. An interaction is said to take place when the simultaneous effect of two or more variables is different from the sum of their individual effects. An individual may have a high preference for mangoes and may also like ice-cream, which does not mean that he would like mango ice cream, leading to an interaction.

The sales of a product may be influenced by two factors, namely, price level and store size. There may be three levels of price—low

NOTES

NOTES

(A1), medium (A2) and high (A3). The store size could be categorized into small (B1) and big (B2). This could be conceptualized as a two-factor design with information reported in the form of a table. In the table, each level of one factor may be presented as a row and each level of another variable would be presented as a column. This example could be summarized in the form of a table having three rows and two columns. This would require $3 \times 2 = 6$ cells. Therefore, six different levels of treatment combinations would be produced each with a specific level of price and store size. The respondents would be randomly selected and randomly assigned to the six cells.

The tabular presentation of 3×2 factorial design is given in Table 8.1.

Table 8.1 3×2 Factorial Design for Price Level and Store Size

Price	Store	
	Small (B ₁)	Big (B ₂)
Low Level (A ₁)	A ₁ B ₁	A ₁ B ₂
Medium Level (A ₂)	A ₂ B ₁	A ₂ B ₂
High Level (A ₃)	A ₃ B ₁	A ₃ B ₂

Respondents in each cell receive a specified treatment combination. For example, respondents in the upper left hand corner cell would face small level of price and small store. Similarly, the respondents in the lower right hand corner cell will be subjected to both high price level and big store.

The main advantages of factorial design are as follows:

- It is possible to measure the main effects and interaction effect of two or more independent variables at various levels.
- It allows a saving of time and effort because all observations are employed to study the effects of each factor.
- The conclusion reached using factorial design has broader applications as each factor is studied with different combinations of other factors.

The limitation of this design is that the number of combinations (number of cells) increases with increased number of factors and levels. However, a fractional factorial design could be used if interest is in studying only a few of the interactions or main effects.

Errors Affecting Research Design

We have discussed three types of research designs, namely, exploratory, descriptive and experimental. All of these have some scope of error. There could be various sources of errors in research design.

Exploratory research is conducted using focus group discussion, secondary data, analysis of case study and expert opinion survey. It is quite likely that members

of the focus group have not been selected properly. Secondary data may not be free from errors (in fact, one needs to evaluate the methodology used in collecting such a data). Also, the experts chosen for the survey may not be experts in the field. As a matter of fact, getting an expert is very difficult task. All these factors could lead to errors in the exploratory design.

In the descriptive design, the purpose is to describe a phenomenon. For this one could use a structured questionnaire. It could always happen that the respondents do not give correct responses to some of the questions, thereby resulting in wrong information.

In the true experimental design and statistical design, the respondents are selected at random which may not be the case in real life. Many a times, in actual business situation, the value judgements play very important role in selecting the respondents. Further, there can always be errors in observations.

Check Your Progress

3. What is the main objective of conducting an exploratory research?
4. Mention the various methods of conducting an exploratory research.
5. What is the objective of conducting descriptive research studies?

8.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The three basic principles which should be covered while formulating a research design are as follows:
 - Convert the research question and the stated assumptions/hypotheses into variables that can be measured.
 - Specify the process to complete the above task.
 - Specify the ‘control mechanism(s)’ to follow so that the effect of other variables that could have an effect on the outcome of the study has been controlled.
2. Research design refers to the all-inclusive technique chosen to combine the different constituents of the study in a comprehensible and logical way, hence, effectively addressing the research problem.
3. The basic objective of conducting an exploratory research is to explore and obtain clarity about the problem situation. It is flexible in its approach and mostly involves a qualitative investigation.
4. The various methods of conducting an exploratory research are as follows:
 - Secondary Resource Analysis

NOTES

NOTES

- Case Study Method
- Expert Opinion Survey
- Focus Group Discussions

5. The objective of descriptive research studies is to provide a comprehensive and detailed explanation of the phenomena under study. The intended objective might be to give a detailed sketch or profile of the respondent population being studied.

8.5 SUMMARY

- It has been found by research scholars and managers alike that most research studies do not result in any significant findings because of a faulty research design.
- Green *et al.* (2008) defines research design as ‘the specification of methods and procedures for acquiring the information needed.
- Exploratory designs, as stated earlier, are the simplest and most loosely structured designs. As the name suggests, the basic objective of the study is to explore and obtain clarity about the problem situation.
- Secondary sources of data, as the name suggests, are data in terms of the details of previously collected findings in facts and figures—which have been authenticated and published.
- Another valuable secondary resource is the compiled and readily available databases of the entire industry, business or construct. These might be available on free and public domains or through a structured acquisition process and cost.
- Another way of conducting an exploratory research is the case study method. This requires an in-depth study and is focused on a single unit of analysis. This unit could be an employee or a customer; an organization or a complete country analysis.
- The objective of descriptive research studies is to provide a comprehensive and detailed explanation of the phenomena under study.
- Descriptive research, thus, are conclusive studies. However, they lack the precision and accuracy of experimental designs, yet it lends itself to a wide range of situations and is more frequently used in business research.
- A single sample of the identified population that is studied over a longer period of time is termed as a longitudinal study design.
- Longitudinal study using the same section of respondents thus provides more accurate data than one using a series of different samples. These

kinds of panels are defined as true panels and the ones using a different group every time are called omnibus panels.

- Experimental designs are conducted to infer causality. In an experiment, a researcher actively manipulates one or more causal variables and measures their effects on the dependent variables of interest.
- The necessary conditions for making causal inferences are: (i) concomitant variation, (ii) time order of occurrence of variables and (iii) absence of other possible causal factors. The first condition implies that cause and effect variables should have a high correlation.
- A factorial design may be employed to measure the effect of two or more independent variables at various levels. The factorial designs allow for interaction between the variables.
- We have discussed three types of research designs, namely, exploratory, descriptive and experimental. All of these have some scope of error. There could be various sources of errors in research design.

NOTES

8.6 KEY WORDS

- **Case study method:** This refers to an in-depth study of a single unit of analysis. This could be an employee, the owner, a customer, a company or even a country.
- **Cross-sectional designs:** It is a descriptive study done on a representative group of people at a single moment in time.
- **Focus group discussion:** It is a sociological method in which 6-10 people discuss the topic being researched.
- **Judgemental analysis:** It is the formation of a judgement based upon personal impressions rather than facts.
- **Longitudinal designs:** It is a single sample studied over a longer period of time. There are periodic measurements done of the study variable.

8.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are exploratory designs?
2. Briefly discuss the methods that can be used in an exploratory design.
3. What are descriptive designs?
4. Write a short note on the different kinds of descriptive designs available.

Long Answer Questions

1. Distinguish between internal and external validity of the experiments.
2. Explain the four types of experimental designs.
3. Discuss the various sources of errors.

NOTES

8.8 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

UNIT 9 SAMPLING

Structure

- 9.0 Introduction
- 9.1 Objectives
- 9.2 Characteristics and Implications of Sample Design
 - 9.2.1 Sample vs Census
 - 9.2.2 Sampling vs Non-Sampling Error
- 9.3 Answers to Check Your Progress Questions
- 9.4 Summary
- 9.5 Key Words
- 9.6 Self Assessment Questions and Exercises
- 9.7 Further Readings

NOTES

9.0 INTRODUCTION

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed but may include simple random sampling or systematic sampling.

In this unit, we will discuss the different characteristics and implications of sampling.

9.1 OBJECTIVES

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

- Define sampling and analyse the concept of sampling
- Define the terms—sampling frame, sampling unit, census and population
- Differentiate between sample versus census
- Differentiate between sampling and non-sampling error

9.2 CHARACTERISTICS AND IMPLICATIONS OF SAMPLE DESIGN

Research objectives are generally translated into research questions that enable the researchers to identify the information needs. Once the information needs are specified, the sources of collecting the information are sought. Some of the information may be collected through secondary sources (published material), whereas the rest may be obtained through primary sources. The primary methods of collecting information could be the observation method, personal interview with

NOTES

questionnaire, telephone surveys and mail surveys. Surveys are, therefore, useful in information collection, and their analysis plays a vital role in finding answers to research questions. Survey respondents should be selected using the appropriate procedures, otherwise the researchers may not be able to get the right information to solve the problem under investigation. The process of selecting the right individuals, objects or events for the study is known as sampling. Sampling involves the study of a small number of individuals, objects chosen from a larger group.

Sampling Concept

Before we get into the details of various issues pertaining to sampling, it would be appropriate to discuss some of the sampling concepts.

- **Population:** Population refers to any group of people or objects that form the subject of study in a particular survey and are similar in one or more ways. For example, the number of full-time MBA students in a business school could form one population. If there are 200 such students, the population size would be 200. We may be interested in understanding their perceptions about business education. If there are 200 class IV employees in an organization and we are interested in measuring their job satisfaction, all the 200 class IV employees would form the population of interest. If a TV manufacturing company produces 150 TVs per week and we are interested in estimating the proportion of defective TVs produced per week, all the 150 TVs would form our population. If, in an organization there are 1000 engineers, out of which 350 are mechanical engineers and we are interested in examining the proportion of mechanical engineers who intend to leave the organization within six months, all the 350 mechanical engineers would form the population of interest. If the interest is in studying how the patients in a hospital are looked after, then all the patients of the hospital would fall under the category of population.
- **Element:** An element comprises a single member of the population. Out of the 350 mechanical engineers mentioned above, each mechanical engineer would form an element of the population. In the example of MBA students whose perception about the management education is of interest to us, each of the 200 MBA students will be an element of the population. This means that there will be 200 elements of the population.
- **Sampling frame:** Sampling frame comprises all the elements of a population with proper identification that is available to us for selection at any stage of sampling. For example, the list of registered voters in a constituency could form a sampling frame; the telephone directory; the number of students registered with a university; the attendance sheet of a particular class and the payroll of an organization are examples of sampling frames. When the population size is very large, it becomes virtually impossible to form a sampling frame. We know that there is a large number of consumers of soft

drinks and, therefore, it becomes very difficult to form the sampling frame for the same.

- **Sample:** It is a subset of the population. It comprises only some elements of the population. If out of the 350 mechanical engineers employed in an organization, 30 are surveyed regarding their intention to leave the organization in the next six months, these 30 members would constitute the sample.
- **Sampling unit:** A sampling unit is a single member of the sample. If a sample of 50 students is taken from a population of 200 MBA students in a business school, then each of the 50 students is a sampling unit. Another example could be that if a sample of 50 patients is taken from a hospital to understand their perception about the services of the hospital, each of the 50 patients is a sampling unit.
- **Sampling:** It is a process of selecting an adequate number of elements from the population so that the study of the sample will not only help in understanding the characteristics of the population but will also enable us to generalize the results. We will see later that there are two types of sampling designs—probability sampling design and non-probability sampling design.
- **Census (or complete enumeration):** An examination of each and every element of the population is called census or complete enumeration. Census is an alternative to sampling. We will discuss the inherent advantages of sampling over a complete enumeration later.

Uses of Sampling in Real Life

In our day-to-day life we make use of the concept of sampling. There is hardly any person who has not made use of the concept in a real-life situation. Consider the following examples:

- Suppose you go to a grocery shop to purchase rice. You have been instructed by your mother to purchase good quality rice. On reaching the grocery shop you have the choice of buying the rice from any one of three bags. What is generally done is that you pick up a handful of rice from each bag, examine its quality and then decide about which bag's rice is to be bought. The concept of sampling is being used here as a handpick from each bag is a sample and examining the quality is a process by which you are trying to assess the quality of all the rice in the bag.
- Suppose you have a guest for dinner at your residence. Your mother prepares a number of dishes and before the guest arrives, she may give you a tablespoon of each of the dish to taste and tell her whether all the ingredients are in the right proportion or not. Again, a sample is being taken from each of the dish to know how each of them tastes.
- You go to a bookshop to buy a magazine. Before you decide to buy it, you may flip through its pages to know whether the contents of the magazines

NOTES

are of interest to you or not. Again, a sample of pages is taken from the magazine.

9.2.1 Sample vs Census

NOTES

In a research study, we are generally interested in studying the characteristics of a population. Suppose in a town there are 2 lakh households and we are interested in estimating the proportion of those households who spend their summer vacations in a hill station. This information can be obtained by asking every household in that town. If all the households in a population are asked to provide information, such a survey is called a census. There is an alternative way of obtaining the same information by choosing a subset of all the two lakh households and asking them for the same information. This subset is called a sample. Based upon the information obtained from the sample, a generalization about the population characteristic could be made. However, that sample has to be representative of the population. For a sample to be a representative of the population, the distribution of sampling units in the sample has to be in the same proportion as the elements in the population. For example, if in a town there are 50, 35 and 15 per cent households in lower, middle and upper income groups, then a sample taken from this population should have the same proportions in for it to be representative. There are several advantages of sample over census.

- Sample saves time and cost. Consider as an example that we are interested in estimating the monthly average household expenditure on food items by the people of Delhi. It is known that the population of Delhi is approximately 1.2 crore. Now, if we assume that there are five members per household, it would mean that the population comprises approximately 24 lakh households. Collecting data on the expenditure of each of the 24 lakh households on food items would be a very time-consuming and expensive exercise. This is because you will need to hire a number of investigators and train them before you conduct the survey on the 24 lakh households. Instead, if a sample of, say, 2000 households is chosen, the task would not only be finished faster but will be inexpensive, too.
- Many times a decision-maker may not have too much of time to wait till all the information is available. Therefore, a sample could come to his rescue.
- There are situations where a sample is the only option. When we want to estimate the average life of fluorescent bulbs, what is done is that they are burnt out completely. If we go for a complete enumeration there would not be anything left for use. Another example could be testing the quality of a photographic film. To test the quality, we need to expose it completely and the moment it is exposed it gets destroyed. Therefore, sample is the only choice.
- The study of a sample instead of complete enumeration may, at times, produce more reliable results. This is because by studying a sample, fatigue

is reduced and fewer errors occur while collecting the data, especially when a large number of elements are involved.

A census is appropriate when the population size is small, e.g., the number of public sector banks in the country. Suppose the researcher is interested in collecting information from the top management of a bank regarding their views on the monetary policy announced by the Reserve Bank of India (RBI), in this case, a complete enumeration may be possible as the population size is not very large. As another example, consider a business school having a few students from Europe, East Africa, South East Asia and the Middle East. These students would have their own problems in settling down in the Indian environment because of the differences in social, cultural and environmental factors. To understand their concerns, a survey of population may be more appropriate. Therefore, a survey of population could be used when there is a lot of heterogeneity in the variables of interest and the population size is small.

9.2.2 Sampling vs Non-Sampling Error

There are two types of error that may occur while we are trying to estimate the population parameters from the sample. These are called sampling and non-sampling errors.

Sampling error: This error arises when a sample is not representative of the population. For example, if our population comprises 200 MBA students in a business school and we want to estimate the average height of these 200 students by taking a sample of 10 (say). Let us assume for the sake of simplicity that the true value of population mean (parameter) is known. When we estimate the average height of the sampled students, we may find that the sample mean is far away from the population mean. The difference between the sample mean and the population mean is called sampling error, and this could arise because the sample of 10 students may not be representative of the entire population. Suppose now we increase the sample size from 10 to 15, we may find that the sampling error reduces. This way, if we keep doing so, we may note that the sampling error reduces with the increase in sample size as an increased sample may result in increasing the representativeness of the sample.

Non-sampling error: This error arises not because a sample is not a representative of the population but because of other reasons. Some of these reasons are listed below:

- The respondents when asked for information on a particular variable may not give the correct answers. If a person aged 48 is asked a question about his age, he may indicate the age to be 36, which may result in an error and in estimating the true value of the variable of interest.
- The error can arise while transferring the data from the questionnaire to the spreadsheet on the computer.

NOTES

NOTES

- There can be errors at the time of coding, tabulation and computation.
- If the population of the study is not properly defined, it could lead to errors.
- The chosen respondent may not be available to answer the questions or may refuse to be part of the study.
- There may be a sampling frame error. Suppose the population comprises households with low income, high income and middle class category. The researcher might decide to ignore the low-income category respondents and may take the sample only from the middle and the high-income category people.

Sampling Design

Sampling design refers to the process of selecting samples from a population. There are two types of sampling designs—probability sampling design and non-probability sampling design. Probability sampling designs are used in conclusive research. In a probability sampling design, each and every element of the population has a known chance of being selected in the sample. The known chance does not mean equal chance. Simple random sampling is a special case of probability sampling design where every element of the population has both known and equal chance of being selected in the sample. In case of non-probability sampling design, the elements of the population do not have any known chance of being selected in the sample. These sampling designs are used in exploratory research.

Check Your Progress

1. What is a sampling unit?
2. Why does the study of a sample instead of complete enumeration, at times produces reliable results?

9.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. A sampling unit is a single member of the sample. If a sample of 50 students is taken from a population of 200 MBA students in a business school, then each of the 50 students is a sampling unit.
2. The study of a sample instead of complete enumeration may, at times, produce more reliable results. This is because by studying a sample, fatigue is reduced and fewer errors occur while collecting the data, especially when a large number of elements are involved.

9.4 SUMMARY

- Research objectives are generally translated into research questions that enable the researchers to identify the information needs. Once the information needs are specified, the sources of collecting the information are sought.
- The primary methods of collecting information could be the observation method, personal interview with questionnaire, telephone surveys and mail surveys. Surveys are, therefore, useful in information collection, and their analysis plays a vital role in finding answers to research questions.
- Population refers to any group of people or objects that form the subject of study in a particular survey and are similar in one or more ways.
- An element comprises a single member of the population. Out of the 350 mechanical engineers mentioned above, each mechanical engineer would form an element of the population.
- An examination of each and every element of the population is called census or complete enumeration. Census is an alternative to sampling. We will discuss the inherent advantages of sampling over a complete enumeration later.
- In a research study, we are generally interested in studying the characteristics of a population. Suppose in a town there are 2 lakh households and we are interested in estimating the proportion of those households who spend their summer vacations in a hill station. This information can be obtained by asking every household in that town.
- There are two types of error that may occur while we are trying to estimate the population parameters from the sample. These are called sampling and non-sampling errors.
- Sampling design refers to the process of selecting samples from a population. There are two types of sampling designs—probability sampling design and non-probability sampling design. Probability sampling designs are used in conclusive research.

NOTES

9.5 KEY WORDS

- **Sampling design:** Sampling design refers to the process of selecting samples from a population. There are two types of sampling designs—probability sampling design and non-probability sampling design.
- **Sampling frame:** Sampling frame comprises all the elements of a population with proper identification that is available to us for selection at any stage of sampling.

9.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. Write a short note on sampling frame.
2. What are the uses of sampling in real life?
3. What are the reasons for the occurrence of non-sampling error?

Long-Answer Questions

1. Differentiate between sample and census.
2. Analyse the various concepts of sampling.
3. Discuss the differences between sampling and non-sampling error.

9.7 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

UNIT 10 TYPES OF SAMPLING

Structure

- 10.0 Introduction
- 10.1 Objectives
- 10.2 Probability Sampling
- 10.3 Non-Probability Sampling
- 10.4 Criteria for Selecting a Sampling Procedure
- 10.5 Answers to Check Your Progress Questions
- 10.6 Summary
- 10.7 Key Words
- 10.8 Self Assessment Questions and Exercises
- 10.9 Further Readings

NOTES

10.0 INTRODUCTION

Sampling means selecting a particular group or sample to represent the entire population. Sampling methods are majorly divided into two categories probability sampling and non-probability sampling. In the first case, each member has a fixed, known opportunity to belong to the sample, whereas in the second case, there is no specific probability of an individual to be a part of the sample.

Probability sampling is a sampling technique, in which the subjects of the population get an equal opportunity to be selected as a representative sample. Non-probability sampling is a method of sampling wherein, it is not known that which individual from the population will be selected as a sample.

10.1 OBJECTIVES

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

- Understand the basic concepts of sampling
- Distinguish between sample and census
- Differentiate between a sampling and non-sampling error
- Understand the meaning of sampling design
- Explain the different types of probability sampling designs
- Describe various types of non-probability sampling designs

10.2 PROBABILITY SAMPLING

Sampling design refers to the process of selecting samples from a population. There are two types of sampling designs—probability sampling design and non-

NOTES

probability sampling design. Probability sampling designs are used in conclusive research. In a probability sampling design, each and every element of the population has a known chance of being selected in the sample. The known chance does not mean equal chance. Simple random sampling is a special case of probability sampling design where every element of the population has both known and equal chance of being selected in the sample. In case of non-probability sampling design, the elements of the population do not have any known chance of being selected in the sample. These sampling designs are used in exploratory research.

Probability Sampling Design

Under this, the following sampling designs would be covered—simple random sampling with replacement (SRSWR), simple random sampling without replacement (SRSWOR), systematic sampling, stratified random sampling and cluster sampling.

Simple Random Sampling with Replacement

Under this scheme, a list of all the elements of the population from where the samples to be drawn is prepared. If there are 1000 elements in the population, we write the identification number or the name of all the 1000 elements on 1000 different slips. These are put in a box and shuffled properly. If there are 20 elements to be selected from the population, the simple random sampling procedure involves selecting a slip from the box and reading of the identification number. Once this is done, the chosen slip is put back to the box and again a slip is picked up and the identification number is read from that slip. This process continues till a sample of 20 is selected. Please note that the first element is chosen with a probability of $1/1000$, the second one is also selected with the same probability and so are all the subsequent elements of the population.

An alternative way of selecting the samples from the population is by using random number tables. Table 10.1 gives an illustrative example of random numbers.

Table 10.1 gives four-digit random numbers arranged in 20 rows and five columns. These random numbers can be generated by a computer programmed to scramble numbers. The logic for generating random number is that any number can be constructed from numbers 0 to 9. The probability that any one digit from 0 through 9 will appear is the same as that for any other digit and the appearance of the numbers is statistically independent. Further, the probability of one sequence of digits occurring is the same as that for any other sequence of the same length.

The use of random number table for selecting samples could be illustrated through an example. Suppose there are 75 students in a class and it is decided to select 15 out of the 75 students. These students can be numbered from 01 to 75. Now, to pick up 15 students using random numbers and following the scheme of simple random sampling with replacement, we proceed as follows:

Table 10.1 Select Four-Digit Random Numbers

I	II	III	IV	V
2807	0495	6183	7871	9559
8016	5732	3448	0164	2367
1322	4678	8034	1139	1474
0843	4625	7407	9987	5734
2364	1187	4565	2343	9786
4885	8755	4355	5465	0575
3406	4678	5950	7222	8494
5927	6010	7545	8979	1041
4447	3476	9140	0736	2332
4968	7553	1073	2493	4251
7489	1630	2330	4250	6170
4010	2707	3925	6007	8089
6531	9784	5520	7764	0008
7052	3861	7115	9521	2192
6573	2793	8710	2127	3846
8094	3205	2030	3035	5765
8615	6092	1900	4792	7684
9136	4016	3495	6549	9603
9656	5246	5090	8306	1522
2017	8323	1685	3006	3441

NOTES

- With eyes closed, we place our finger on a number on the random number table. Suppose it is on the first row and the first column of our table. Now, we go down the first two columns and choose two-digit random numbers running from 01 to 75. If any number greater than 75 appears, it gets rejected. This way, the first number to be selected would be 28. The second number is 80, which would be rejected as we are choosing numbers from 01 to 75. The next selected number would be 13, followed by 08, 23, 48, 34, 59, 44, 49, 74, 40, 65, 70 and 65. Note that 65 has appeared twice. Since we are using the scheme of simple random sampling with replacement, we would retain it. This way we have selected 14 samples. The 15th number selected would be 20. In brief, the scheme explained above states that any number greater than the population size (in this case 75) is rejected and only the numbers from 01 to 75 are selected. A number may get repeated because simple random sampling scheme is done with replacement.

Simple Random Sampling without Replacement

In the case of simple random sample without replacement, the procedure is identical to what was explained in the case of simple random sampling with replacement.

NOTES

The only difference here is that the chosen slip is not placed back in the box. This way, the first unit would be selected with the probability of $1/1000$, second unit with the probability of $1/999$, the third will be selected with a probability of $1/998$ and so on, till we select the required number of elements (in this case, 15) in our sample.

The simple random sampling (with or without replacement) is not used in a consumer research. This is because in a consumer research the population size is usually very large, which creates problems in the preparation of a sampling frame. For example, there is a large number of consumers of soft drinks, pizza, shampoo, soap, chocolate, etc. However, these (SRSWR and SRSWOR) designs could be useful when the population size is very small, for example, the number of steel/aluminum-producing companies in India and the number of banks in India. Since the population size is quite small, the preparation of a sampling frame does not create any problem.

Another problem with these (SRSWR and SRSWOR) designs is that we may not get a representative sample using such a scheme. Consider an example of a locality having 10,000 households, out of which 5,000 belong to low-income group, 3,500 belong to middle income group and the remaining 1,500 belong to high-income group. Suppose it is decided to take a sample of 100 households using the simple random sampling. The selected sample may not contain even a single household belonging to the high- and middle-income group and only the low-income households may get selected, thus, resulting in a non-representative sample.

Systematic Sampling

Systematic sampling takes care of the limitation of the simple random sampling that the sample may not be a representative one. In this design, the entire population is arranged in a particular order. The order could be the calendar dates or the elements of a population arranged in an ascending or a descending order of the magnitude which may be assumed as random. List of subjects arranged in the alphabetical order could also be used and they are usually assumed to be random in order. Once this is done, the steps followed in the systematic sampling design are as follows:

- First of all, a sampling interval given by $K = N/n$ is calculated, where N = the size of the population and n = the size of the sample. It is seen that the sampling interval K should be an integer. If it is not, it is rounded off to make it an integer.
- A random number is selected from 1 to K . Let us call it C .
- The first element to be selected from the ordered population would be C , the next element would be $C + K$ and the subsequent one would be $C + 2K$ and so on till a sample of size n is selected.

This way we can get representation from all the classes in the population and overcome the limitations of the simple random sampling. To take an example, assume that there are 1,000 grocery shops in a small town. These shops could be arranged in an ascending order of their sales, with the first shop having the smallest sales and the last shop having the highest sales. If it is decided to take a sample of 50 shops, then our sampling interval K will be equal to $1000 \div 50 = 20$. Now we select a random number from 1 to 20. Suppose the chosen number is 10. This means that the shop number 10 will be selected first and then shop number $10 + 20 = 30$ and the next one would be $10 + 2 \times 20 = 50$ and so on till all the 50 shops are selected. This way we can get a representative sample in the sense that it will contain small, medium and large shops.

It may be noted that in a systematic sampling the first unit of the sample is selected at random (probability sampling design) and having chosen this, we have no control over the subsequent units of sample (non-probability sampling). Because of this, this design at times is called mixed sampling.

The main advantage of systematic sampling design is its simplicity. When sampling from a list of population arranged in a particular order, one can easily choose a random start as described earlier. After having chosen a random start, every K^{th} item can be selected instead of going for a simple random selection. This design is statistically more efficient than a simple random sampling, provided the condition of ordering of the population is satisfied.

The use of systematic sampling is quite common as it is easy and cheap to select a systematic sample. In systematic sampling one does not have to jump back and forth all over the sampling frame wherever random number leads, and neither does one have to check for duplication of elements as compared to simple random sampling. Another advantage of a systematic sampling over simple random sampling is that one does not require a complete sampling frame to draw a systematic sample. The investigator may be instructed to interview every 10th customer entering a mall without a list of all customers.

There may be situations where it may not be possible to get a representative sample. The design can create problems if the sampling interval is a whole number multiple of some cycle related to the problem. On this design there may be a problem that there is a high probability of systematic bias creeping into the sample resulting in a non-representative sample. Consider, for example, the case of a certain PVR cinema hall where there may be a couple of snack bars. We may be interested in estimating the average daily sales of a particular snack bar in that PVR. Now, using the daily data with the population and sample size known, we compute a sampling interval which may be a multiple of seven. Using this, we may select our first element which would reflect one of the seven days of the week, say Friday. The next element would also be Friday, as our sampling interval is a multiple of seven and so the subsequent elements of the population. Therefore, our sample

NOTES

NOTES

would comprise only Fridays and the sample would not reflect day of the week variation in the sales data, which could result in a non-representative sample. Therefore, while using daily data, care should be taken that our sampling interval is not a multiple of seven.

Stratified Random Sampling

Under this sampling design, the entire population (universe) is divided into strata (groups), which are mutually exclusive and collectively exhaustive. By mutually exclusive, it is meant that if an element belongs to one stratum, it cannot belong to any other stratum. Strata are collectively exhaustive if all the elements of various strata put together completely cover all the elements of the population. The elements are selected using a simple random sampling independently from each group.

There are two reasons for using a stratified random sampling rather than simple random sampling. One is that the researchers are often interested in obtaining data about the component parts of a universe. For example, the researcher may be interested in knowing the average monthly sales of cell phones in 'large', 'medium' and 'small' stores. In such a case, separate sampling from within each stratum would be called for. The second reason for using a stratified random sampling is that it is more efficient as compared to a simple random sampling. This is because dividing the population into various strata increases the representativeness of the sampling as the elements of each stratum are homogeneous to each other.

There are certain issues that may be of interest while setting up a stratified random sample. These are:

What criteria should be used for stratifying the universe (population)?

The criteria for stratification should be related to the objectives of the study. The entire population should be stratified in such a way that the elements are homogeneous within the strata, whereas there should be heterogeneity between strata. As an example, if the interest is to estimate the expenditure of households on entertainment, the appropriate criteria for stratification would be the household income. This is because the expenditure on entertainment and household income are highly correlated. As another example, if the objective of the study is to estimate the amount of money spent on cosmetics, then, gender could be used as an appropriate criteria for stratification. This is because it is known that though both men and women use cosmetics, the expenditure by women is much more than that of their male counterparts. Someone may argue out that gender may no longer remain the appropriate criteria if it is not backed by income. Therefore, the researcher might have to use two or more criteria for stratification depending upon the problem in hand. This would only increase the number of strata thereby making the sampling difficult.

Generally stratification is done on the basis of demographic variables like age, income, education and gender. Customers are usually stratified on the basis of life stages and income levels to study their buying patterns. Companies may be stratified according to size, industry, profits for analysing the stock market reactions.

How many strata should be constructed?

Going by common sense, as many strata as possible should be used so that the elements of each stratum will be as homogeneous as possible. However, it may not be practical to increase the number of strata and, therefore, the number may have to be limited. Too many strata may complicate the survey and make preparation and tabulation difficult. Costs of adding more strata may be more than the benefit obtained. Further, the researcher may end up the practical difficulty of preparing a separate sampling frame as the simple random samples are to be drawn from each stratum.

What should be appropriate number of samples size to be taken in each stratum?

This question pertains to the number of observations to be taken out from each stratum. At the outset, one needs to determine the total sample size for the universe and then allocate it between each stratum. This may be explained as follows:

Let there be a population of size N . Let this population be divided into three strata based on a certain criterion. Let N_1 , N_2 and N_3 denote the size of strata 1, 2 and 3 respectively, such that $N = N_1 + N_2 + N_3$. These strata are mutually exclusive and collectively exhaustive. Each of these three strata could be treated as three populations. Now, if a total sample of size n is to be taken from the population, the question arises that how much of the sample should be taken from strata 1, 2 and 3 respectively, so that the sum total of sample sizes from each strata adds up to n .

Let the size of the sample from first, second and third strata be n_1 , n_2 , and n_3 respectively such that $n = n_1 + n_2 + n_3$. Then, there are two schemes that may be used to determine the values of n_i , ($i = 1, 2, 3$) from each strata. These are proportionate and disproportionate allocation schemes.

Proportionate allocation scheme: In this scheme, the size of the sample in each stratum is proportional to the size of the population of the strata. As an example, if a bank wants to conduct a survey to understand the problems that its customers are facing, it may be appropriate to divide them into three strata based upon the size of their deposits with the bank. If we have 10,000 customers of a bank in such a way that 1,500 of them are big account holders (having deposits more than ₹ 10 lakh), 3,500 of them are medium sized account holders (having deposits of more than ₹ 2 lakh but less than ₹ 10 lakh), the remaining 5,000 are

NOTES

NOTES

small account holders (having deposits of less than ₹ 2 lakh). Suppose the total budget for sampling is fixed at ₹ 20,000 and the cost of sampling a unit (customer) is ₹ 20. If a sample of 100 is to be chosen from all the three strata, the size of the sample from strata 1 would be:

$$n_1 = n \times \frac{N_1}{N} = 100 \times \frac{1500}{10000} = 15$$

The size of sample from strata 2 would be:

$$n_2 = n \times \frac{N_2}{N} = 100 \times \frac{3500}{10000} = 35$$

The size of sample from strata 3 would be:

$$n_3 = n \times \frac{N_3}{N} = 100 \times \frac{5000}{10000} = 50$$

This way the size of the sample chosen from each stratum is proportional to the size of the stratum. Once we have determined the sample size from each stratum, one may use the simple random sampling or the systematic sampling or any other sampling design to take out samples from each of the strata.

Disproportionate allocation: As per the proportionate allocation explained above, the sizes of the samples from strata 1, 2 and 3 are 15, 35 and 50 respectively. As it is known that the cost of sampling of a unit is ₹ 20 irrespective of the strata from where the sample is drawn, the bank would naturally be more interested in drawing a large sample from stratum 1, which has the big customers, as it gets most of its business from strata 1. In other words, the bank may follow a disproportionate allocation of sample as the importance of each stratum is not the same from the point of view of the bank. The bank may like to take a sample of 45 from strata 1 and 40 and 15 from strata 2 and 3 respectively. Also, a large sample may be desired from the strata having more variability.

Cluster Sampling

In the cluster sampling, the entire population is divided into various clusters in such a way that the elements within the clusters are heterogeneous. However, there is homogeneity between the clusters. This design, therefore, is just the opposite of the stratified sampling design, where there was homogeneity within the strata and heterogeneity between the strata. To illustrate the example of a cluster sampling, one may assume that there is a company having its corporate office in a multi-storey building. In the first floor, we may assume that there is a marketing department where the offices of the president (marketing), vice president (marketing) and so on to the level of management trainee (marketing) are there. Naturally, there would be a lot of variation (heterogeneity) in the amount of salaries they draw and hence a high amount of variation in the amount of money spent on entertainment. Similarly, if the finance department is housed on the second floor, we may find almost a

similar pattern. Same could be assumed for third, fourth and other floors. Now, if each of the floors could be treated as a cluster, we find that there is homogeneity between the clusters but there is a lot of heterogeneity within the clusters. Now, a sample of, say, 2 to 3 clusters is chosen at random and once having done so, each of the cluster is enumerated completely to be able to make an estimate of the amount of money the entire population spends on entertainment.

Examples of cluster sampling could include ad-hoc organizational committees drawn from various departments to advise the CEO of a company on product development, new product ideas, evaluating alternative advertising programmes, budget allocations and marketing strategies. Each of the clusters comprises a heterogeneous collection of members with different interests, background, experience, value system and philosophy. The CEO of the company may be able to take strategic decisions based upon their combined advice.

Although the per unit costs of cluster sampling are much lower than those of other probability sampling, the applicability of cluster sampling to an organizational context may be questioned as a cluster may not contain heterogeneous elements. The condition of heterogeneity within the cluster and homogeneity between the clusters may not be met. As another example, the households in a block are to be similar rather than dissimilar and as a result, it may be difficult to form heterogeneous clusters.

Cluster sampling is useful when populations under a survey are widely dispersed and drawing a simple random sample may be impractical.

Multi-stage Sampling

It is sometimes convenient and economical to collect certain items of information from the whole of the units of a sample and other items of usually more detailed information from a sub-sample of the units constituting the original sample. This may be termed two-phase sampling, e.g. if the collection of information concerning variate, y , is relatively expensive, and there exists some other variate, x , correlated with it, which is relatively cheap to investigate, it may be profitable to carry out sampling in two phases.

At the first phase, x is investigated, and the information thus obtained is used either (a) to stratify the population at the second phase, when y is investigated, or (b) as supplementary information at the second phase, a ratio or regression estimate being used. Two-phase sampling is sometimes called “double sampling”.

Check Your Progress

1. What is the alternate way of selecting samples from the population?
2. Why is simple random sampling not used in consumer research?
3. How is cluster sampling different from stratified sampling design?

NOTES

10.3 NON-PROBABILITY SAMPLING

NOTES

Under the non-probability sampling, the following designs would be considered—convenience sampling, purposive sampling, snowball sampling and quota sampling.

Convenience Sampling

Convenience sampling is used to obtain information quickly and inexpensively. The only criterion for selecting sampling units in this scheme is the convenience of the researcher or the investigator. Mostly, the convenience samples used are neighbours, friends, family members, colleagues and ‘passers-by’. This sampling design is often used in the pre-test phase of a research study such as the pre-testing of a questionnaire. Some of the examples of convenience sampling are:

- People interviewed in a shopping centre for their political opinion for a TV programme.
- Monitoring the price level in a grocery shop with the objective of inferring the trends in inflation in the economy.
- Requesting people to volunteer to test products.
- Using students or employees of an organization for conducting an experiment.
- Interviews conducted by a TV channel of people coming out of a cinema hall, to seek their opinion about the movie.
- A researcher visiting a few shops near his residence to observe which brand of a particular product people are buying, so as to draw a rough estimate of the market share of the brand.

In all the above situations, the sampling unit may either be self-selected or selected because of ease of availability. No effort is made to choose a representative sample. Therefore, in this design the difference between the population value (parameters) of interest and the sample value (statistic) is unknown both in terms of the magnitude and direction. Therefore, it is not possible to make an estimate of the sampling error and researchers won’t be able to make a conclusive statement about the results from such a sample. It is because of this, convenience sampling should not be used in conclusive research (descriptive and causal research).

Convenience sampling is commonly used in exploratory research. This is because the purpose of an exploratory research is to gain an insight into the problem and generate a set of hypotheses which could be tested with the help of a conclusive research. When very little is known about a subject, a small-scale convenience sampling can be of use in the exploratory work to help understand the range of variability of responses in a subject area.

Purposive or Judgemental Sampling

Under judgemental sampling, experts in a particular field choose what they believe to be the best sample for the study in question. The judgement sampling calls for

special efforts to locate and gain access to the individuals who have the required information. Here, the judgement of an expert is used to identify a representative sample. For example, the shoppers at a shopping centre may serve to represent the residents of a city or some of the cities may be selected to represent a country. Judgemental sampling design is used when the required information is possessed by a limited number/category of people. This approach may not empirically produce satisfactory results and, may, therefore, curtail generalizability of the findings due to the fact that we are using a sample of experts (respondents) that are usually conveniently available to us. Further, there is no objective way to evaluate the precision of the results. A company wanting to launch a new product may use judgemental sampling for selecting 'experts' who have prior knowledge or experience of similar products. A focus group of such experts may be conducted to get valuable insights. Opinion leaders who are knowledgeable are included in the organizational context. Enlightened opinions (views and knowledge) constitute a rich data source. A very special effort is needed to locate and have access to individuals who possess the required information.

The most common application of judgemental sampling is in business-to-business (B to B) marketing. Here, a very small sample of lead users, key accounts or technologically sophisticated firms or individuals is regularly used to test new product concepts, producing programmes, etc.

Snowball Sampling

Snowball sampling is generally used when it is difficult to identify the members of the desired population, e.g., deep-sea divers, families with triplets, people using walking sticks, doctors specializing in a particular ailment, etc. Under this design each respondent, after being interviewed, is asked to identify one or more in the field. This could result in a very useful sample. The main problem is in making the initial contact. Once this is done, these cases identify more members of the population, who then identify further members and so on. It may be difficult to get a representative sample. One plausible reason for this could be that the initial respondents may identify other potential respondents who are similar to themselves. The next problem is to identify new cases.

Quota Sampling

In quota sampling, the sample includes a minimum number from each specified subgroup in the population. The sample is selected on the basis of certain demographic characteristics such as age, gender, occupation, education, income, etc. The investigator is asked to choose a sample that conforms to these parameters. Field workers are assigned quotas of the sample to be selected satisfying these characteristics.

A researcher wants to measure the job satisfaction level among the employees of a large organization and believes that the job satisfaction level varies across different types of employees. The organization is having 10 per cent, 15 per cent,

NOTES

NOTES

35 per cent and 40 per cent, class I, class II, class III and class IV, employees, respectively. If a sample of 200 employees is to be selected from the organization, then 20, 30, 70 and 80 employees from class I, class II, class III and class IV respectively should be selected from the population. Now, various investigators may be assigned quotas from each class in such a way that a sample of 200 employees is selected from various classes in the same proportion as mentioned in the population. For example, the first field worker may be assigned a quota of 10 employees from class I, 15 from class II, 20 from class III and 30 from class IV. Similarly, a second investigator may be assigned a different quota such that a total sample of 200 is selected in the same proportion as the population is distributed. Please note that the investigators may choose the employees from each class as conveniently available to them. Therefore, the sample may not be totally representative of the population, hence the findings of the research cannot be generalized. However, the reason for choosing this sampling design is the convenience it offers in terms of effort, cost and time.

In the example given above, it may be argued that job satisfaction is also influenced by education level, categorized as higher secondary or below, graduation, and postgraduation and above. By incorporating this variable, the distribution of population may look as given in Table 10.2. From the table, we may note that there are 8 per cent class I employees who are postgraduate and above, there are 35 per cent class IV employees with a higher secondary education and below and so on. Now, suppose a sample of size 200 is again proposed. In this case, the distribution of sample satisfying these two conditions in the same proportion in the population is given in Table 10.3.

Table 10.2 *Distribution of Population (Percentage)*

Education	Category of Employees				
	Class I	Class II	Class III	Class IV	Total
Postgraduation and above	8	5	5	0	18
Graduation	2	10	20	5	37
Higher Secondary and below	0	0	10	35	45
Total	10	15	35	40	100

Table 10.2 *Distribution of Sample (Numbers)*

Education	Category of Employees				
	Class I	Class II	Class III	Class IV	Total
Postgraduation and above	16	10	10	0	36
Graduation	4	20	40	10	74
Higher Secondary and below	0	0	20	70	90
Total	20	30	70	80	200

Table 10.3 indicates that a sample of 20 class II employees who are graduates should be selected. Likewise, a sample of 10 employees who possess

postgraduate and above education should be selected. In the above table, the sample to be taken from each of the 12 cells has been specified. Having done so, each of the investigators is assigned a quota to collect information from the employees conforming to the above norms so that a sample of 200 is selected.

Quota sampling design may look similar to the stratified random sampling design. However, there are differences between the two. In the stratified sampling design, the selection of sample from each stratum is random but in the quota sampling, the respondents may be chosen at the convenience or judgement of the researchers. Further, as already stated, the results of stratified random sampling could be generalized, whereas it may not be possible in the case of quota sampling. Quota sampling has some advantages over the probabilistic techniques. This design is very economical and it does not take too much time to set it up. Also, the use of this design does not require a sampling frame.

However, quota sampling also has certain weaknesses like:

- The total number of cells depends upon the number of control characteristics associated with the objectives of the study. If the control characteristics are large, the total number of cells increases, which may result in making the task of the investigator difficult.
- The chosen control characteristics should be related to the objectives of the study. The findings of the study could be misleading if any relevant parameter is omitted for one reason or the other.
- The investigator may visit those places where the chances of getting the respondents with the required control characteristics are high. The investigator could also avoid some responses that appear to be unfriendly. All this could result in making the findings of the study less reliable.

Determination of Sample Size

The size of a sample depends upon the basic characteristics of the population, the type of information required from the survey and the cost involved. Therefore, a sample may vary in size for several reasons. The size of the population does not influence the size of the sample as will be shown later on.

There are various methods of determining the sample size in practice:

- Researchers may arbitrarily decide the size of sample without giving any explicit consideration to the accuracy of the sample results or the cost of sampling. This arbitrary approach should be avoided.
- For some of the projects, the total budget for the field survey (usually mentioned) in a project proposal is allocated. If the cost of sampling per sample unit is known, one can easily obtain the sample size by dividing the total budget allocation by the cost of sampling per unit. This method concentrates only on the cost aspect of sampling, rather than the value of information obtained from such a sample.

NOTES

NOTES

- There are other researchers who decide on the sample size based on what was done by the other researchers in similar studies. Again, this approach cannot be a substitute for the formal scientific approach.
- The most commonly used approach for determining the size of sample is the confidence interval approach covered under inferential statistics. Below will be discussed this approach while determining the size of a sample for estimating population mean and population proportion. In a confidence interval approach, the following points are taken into account for determining the sample size in estimation of problems involving means:
 - (a) **The variability of the population:** It would be seen that the higher the variability as measured by the population standard deviation, larger will be the size of the sample. If the standard deviation of the population is unknown, a researcher may use the estimates of the standard deviation from previous studies. Alternatively, the estimates of the population standard deviation can be computed from the sample data.
 - (b) **The confidence attached to the estimate:** It is a matter of judgement, how much confidence you want to attach to your estimate. Assuming a normal distribution, the higher the confidence the researcher wants for the estimate, larger will be sample size. This is because the value of the standard normal ordinate 'Z' will vary accordingly. For a 90 per cent confidence, the value of 'Z' would be 1.645 and for a 95 per cent confidence, the corresponding 'Z' value would be 1.96 and so on (see Appendix 1 at the end of the book). It would be seen later that a higher confidence would lead to a larger 'Z' value.
 - (c) **The allowable error or margin of error:** How accurate do we want our estimate to be is again a matter of judgement of the researcher. It will of course depend upon the objectives of the study and the consequence resulting from the higher inaccuracy. If the researcher seeks greater precision, the resulting sample size would be large.

Sample Size for Estimating Population Mean

We have learnt in the central limit theorem that the sampling distribution of the sample mean (\bar{X}) follows a normal distribution with a mean μ and a standard error $\sigma_{\bar{x}}$ irrespective of the shape of population distribution whenever the sample size is large. Symbolically, it may be written as:

$$\bar{X} \sim N(\mu, \sigma_{\bar{x}})$$

$$n \rightarrow 30$$

The above also holds true whenever samples are drawn from normal population. However, in that case, the requirement of a large sample is not there. The various notations are explained as under:

- \bar{X} = Sample mean
 μ = Population mean
 $\sigma_{\bar{x}}$ = Standard error of mean
 n = Sample size
 N = Population size
 σ = Population standard deviation

The value of:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ (when samples are drawn from an infinite population)}$$

$$= \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \text{ (when samples are drawn from a finite population)}$$

The expression $\sqrt{\frac{N-n}{N-1}}$ is called the finite population multiplier and need not be used while sampling from a finite population provided $\frac{n}{N} < 0.05$.

The standard normal variate Z may be written as:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{e\sqrt{n}}{\sigma}$$

Where $\bar{X} - \mu = e = \text{Margin of error}$

$$\therefore n = \frac{Z^2 \sigma^2}{e^2}$$

It may be noted from above that the size of the sample is directly proportional to the variability in the population and the value of Z for a confidence interval. It varies inversely with the size of the error. It may also be noted that the size of a sample does not depend upon the size of population. Below are given some worked out examples for the determination of a sample size.

Example 10.1: An economist is interested in estimating the average monthly household expenditure on food items by the households of a town. Based on past data, it is estimated that the standard deviation of the population on the monthly expenditure on food item is ₹ 30. With allowable error set at ₹ 7, estimate the sample size required at a 90 per cent confidence.

NOTES

NOTES

Solution:

90 per cent confidence $\Rightarrow Z = 1.645$

$$e = ₹ 7$$

$$\sigma = ₹ 30$$

$$n = \frac{Z^2 \sigma^2}{e^2}$$

$$= \frac{(1.645)^2 (30)^2}{(7)^2}$$

$$= 49.7025$$

$$= 50 \text{ (approx.)}$$

Example 10.2: You are given a population with a standard deviation of 8.6. Determine the sample size needed to estimate the mean of the population within ± 0.5 with a 99 per cent confidence.

Solution:

99 per cent confidence $\Rightarrow Z = 2.575$

$$e = \pm 0.5$$

$$\sigma = 8.6$$

$$n = \frac{Z^2 \sigma^2}{e^2}$$

$$= \frac{(2.575)^2 (8.6)^2}{(0.5)^2}$$

$$= 1961.60$$

$$= 1962 \text{ (approx.)}$$

Example 10.3: It is desired to estimate the mean life time of a certain kind of vacuum cleaner. Given that the population standard deviation $\sigma = 320$ days, how large a sample is needed to be able to assert with a confidence level of 96 per cent that the mean of the sample will differ from the population mean by less than 45 days?

Solution:

96 per cent confidence $\Rightarrow Z = 2.055$

$$e = 45$$

$$\sigma = 320$$

$$n = \frac{Z^2 \sigma^2}{e^2}$$

$$= \frac{(2.055)^2 (320)^2}{(45)^2}$$

$$= 213.55$$

$$= 214 \text{ (approx.)}$$

Determination of sample size for estimating the population proportion

Types of Sampling

If the sample proportion \bar{p} is used to estimate the population proportion p , the standard error of \bar{p} ($\sigma_{\bar{p}}$) would be $\sqrt{\frac{pq}{n}}$, where $q = 1 - p$. Now assuming normal distribution, we have

Therefore, $\bar{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$

$$Z = \frac{\bar{p} - p}{\sqrt{\frac{pq}{n}}}$$

Therefore, margin of error $e = \bar{p} - p = Z \sqrt{\frac{pq}{n}}$

$$Z = \frac{e}{\sqrt{\frac{pq}{n}}}$$

$$Z = \frac{e\sqrt{n}}{\sqrt{pq}}$$

$$n = \frac{Z^2 pq}{e^2}$$

The above formula will be used if the value of population proportion p is known. If, however, p is unknown, we substitute the maximum value of pq in the above formula. It can be shown that the maximum value of pq is $1/4$ when $p = 1/2$ and $q = 1/2$.

This is shown in Figure 10.1.

Therefore,

$$n = \frac{1}{4} \frac{Z^2}{e^2}$$

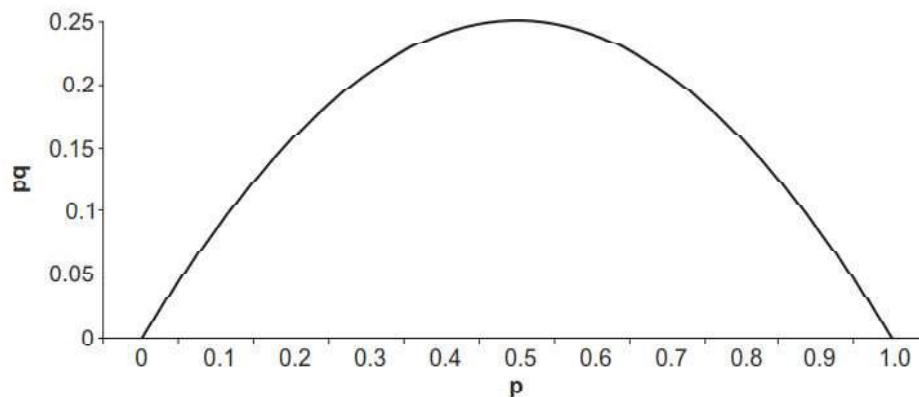


Fig. 10.1 Graph of pq Corresponding to the Values of p

NOTES

NOTES

Let us consider a few examples for determining a sample size while estimating the population proportion.

Example 10.4: A market researcher for a consumer electronics company would like to study the television viewing habits of the residents of a particular, small city. What sample size is needed if he wishes to be 95 per cent confident of being within ± 0.035 of the true proportion who watch the evening news on at least three weeknights if no previous estimate is available?

Solution:

$$95 \text{ per cent confidence} \Rightarrow Z = 1.96$$

$$e = \pm .035$$

$$n = \frac{1}{4} \frac{Z^2}{e^2}$$

$$= \frac{1}{4} \frac{(1.96)^2}{(.035)^2}$$

$$= 784$$

Example 10.5: A manager of a department store would like to study women's spending per year on cosmetics. He is interested in knowing the population proportion of women who purchase their cosmetics primarily from his store. If he wants to have a 90 per cent confidence of estimating the true proportion to be within ± 0.045 , what sample size is needed?

Solution:

$$90 \text{ per cent confidence} \Rightarrow Z = 1.645$$

$$e = \pm .045$$

$$n = \frac{1}{4} \frac{Z^2}{e^2}$$

$$= \frac{1}{4} \frac{(1.645)^2}{(.045)^2}$$

$$= 334.0772$$

$$= 335 \text{ (approx.)}$$

Example 10.6: A consumer electronics company wants to determine the job satisfaction levels of its employees. For this, they ask a simple question, 'Are you satisfied with your job?' It was estimated that no more than 30 per cent of the employees would answer yes. What should be the sample size for this company to estimate the population proportion to ensure a 95 per cent confidence in result, and to be within 0.04 of the true population proportion?

Solution:

95 per cent confidence $\Rightarrow Z = 1.96$

$$e = 0.04$$

$$p = 0.3$$

$$q = 0.7$$

$$n = \frac{Z^2 pq}{e^2}$$

$$= \frac{(1.96)^2 \times 0.3 \times 0.7}{(0.04)^2}$$

$$= 504.21$$

$$= 505 \text{ (approx.)}$$

NOTES**Points to be noted for sample size determination**

There are certain issues to be kept in mind before applying the formulas for the determination of sample size in this chapter. First of all, these formulas are applicable for simple random sampling only. Further, they relate to the sample size needed for the estimation of a particular characteristic of interest. In a survey, a researcher needs to estimate several characteristics of interests and each one of them may require a different sample size. In case the universe is divided into different strata, the accuracy required for determining the sample size for each strata may be different. However, the present method will not be able to serve the requirement. Lastly, the formulas for sample size must be based upon adequate information about the universe.

Check Your Progress

4. State any two examples of convenience sampling.
5. What is the most common application of judgemental sampling?

10.4 CRITERIA FOR SELECTING A SAMPLING PROCEDURE

Depending on how the buyer and the seller (or supplier) reach an agreement, the decision to accept or reject a lot is taken on the basis of the result of a single sample or more. This suggests the need for different acceptance sampling plans.

A sampling plan is specified by the sample size n and the acceptance number C . Accordingly, acceptance sampling may be a single, double, or sequential sampling plan, depending on the number of samples used for deciding the lot.

NOTES

Under each of these sampling plans, the decision to accept or reject a lot is taken by evaluating in terms of the laws of probability the risk of committing the two types of errors. That is, the risks of

- (i) accepting a lot as of satisfactory quality when it ought to have been re-jected for being below the desired quality level, and
- (ii) rejecting a lot as of unsatisfactory quality when it ought to have been accepted for being of satisfactory quality level.

Single Sampling Plan

Under the single sampling plan, a random sample of size n is taken from a randomly operating process (from the assembly line) or an isolated lot offered for sale. The sample is inspected for the number of defective items d contained in it. If d does not exceed a specific pre-determined number of defective items C , known as the accept-ance number, the lot is accepted, otherwise it is rejected. That is,

if $d > C$, the lot is rejected,

and

if $d \leq C$, the lot is accepted.

The requirements of a single sampling plan are met as soon as n and C are stated. *For example*, for a single sampling plan defined by $n = 100$ and $C = 2$, the lot is rejected if $d > 2$, and accepted if $d \leq 2$.

Double Sampling Plan

The double sampling plan provides for a decision to accept or reject a lot on the basis of a second sample, if the first sample fails to offer conclusive evidence taking a final decision either way. First, a sample of size n is selected from the lot. The lot is accepted if the number of defective items d in the sample is C_1 or less, and rejected if d is more than C_2 .

When the number of defective items d in the sample is between C_1 and C_2 , the lot is placed in a doubtful category and a second sample of the same size is drawn from the lot. The lot is accepted if the total number of defective items in both the samples is C_2 or less. The lot is rejected if the total number of defective items in both the samples is more than C_2 .

Sequential Sampling Plan

Sequential or multiple sampling plan allows selection of as many samples as are needed to reach a final decision. After inspecting each sample, the lot is either accepted, rejected, or placed in a doubtful category. It means sampling continues as long as the lot remains in a doubtful category. Theoretically, a sequential sampling plan permits a continuous process of sample selection and inspection. In actual practice, final decision to accept or reject the lot is usually taken after a maximum of 8 or 9 samples have been drawn.

The double sampling plan is used more often than either of the other two sampling plans. If a given lot has a small percentage of defective items, the chances that it is accepted on the basis of the result of the first sample are naturally high. This reduces the cost of sampling as the second sample need not be taken and inspected thereafter. When a lot comes to be put into a doubtful category on the basis of the results of the first sample, it is quite appealing to give the lot a second chance of evaluation before being finally rejected.

NOTES

10.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. An alternative way of selecting the samples from the population is by using random number tables.
2. The simple random sampling (with or without replacement) is not used in a consumer research. This is because in a consumer research the population size is usually very large, which creates problems in the preparation of a sampling frame.
3. Cluster sampling design is just the opposite of the stratified sampling design, where there was homogeneity within the strata and heterogeneity between the strata.
4. Some of the examples of convenience sampling are:
 - People interviewed in a shopping centre for their political opinion for a TV programme.
 - Monitoring the price level in a grocery shop with the objective of inferring the trends in inflation in the economy.
5. The most common application of judgemental sampling is in business-to-business (B to B) marketing. Here, a very small sample of lead users, key accounts or technologically sophisticated firms or individuals is regularly used to test new product concepts, producing programmes, etc.

10.6 SUMMARY

- Sampling design refers to the process of selecting samples from a population. There are two types of sampling designs—probability sampling design and non-probability sampling design.
- In the case of simple random sample without replacement, the procedure is identical to what was explained in the case of simple random sampling with replacement. The only difference here is that the chosen slip is not placed back in the box.

NOTES

- The simple random sampling (with or without replacement) is not used in a consumer research. This is because in a consumer research the population size is usually very large, which creates problems in the preparation of a sampling frame.
- Systematic sampling takes care of the limitation of the simple random sampling that the sample may not be a representative one. In this design, the entire population is arranged in a particular order.
- The main advantage of systematic sampling design is its simplicity. When sampling from a list of population arranged in a particular order, one can easily choose a random start as described earlier.
- Under stratified random sampling design, the entire population (universe) is divided into strata (groups), which are mutually exclusive and collectively exhaustive. By mutually exclusive, it is meant that if an element belongs to one stratum, it cannot belong to any other stratum. Strata are collectively exhaustive if all the elements of various strata put together completely cover all the elements of the population.
- The criteria for stratification should be related to the objectives of the study. The entire population should be stratified in such a way that the elements are homogeneous within the strata, whereas there should be heterogeneity between strata.
- Generally stratification is done on the basis of demographic variables like age, income, education and gender. Customers are usually stratified on the basis of life stages and income levels to study their buying patterns.
- In the cluster sampling, the entire population is divided into various clusters in such a way that the elements within the clusters are heterogeneous. However, there is homogeneity between the clusters.
- Under the non-probability sampling, the following designs are considered—convenience sampling, purposive sampling, snowball sampling and quota sampling.
- Convenience sampling is used to obtain information quickly and inexpensively. The only criterion for selecting sampling units in this scheme is the convenience of the researcher or the investigator.
- Under judgemental sampling, experts in a particular field choose what they believe to be the best sample for the study in question. The judgement sampling calls for special efforts to locate and gain access to the individuals who have the required information.
- Snowball sampling is generally used when it is difficult to identify the members of the desired population, e.g., deep-sea divers, families with triplets, people using walking sticks, doctors specializing in a particular ailment, etc.

- In quota sampling, the sample includes a minimum number from each specified subgroup in the population. The sample is selected on the basis of certain demographic characteristics such as age, gender, occupation, education, income, etc.
- Quota sampling design may look similar to the stratified random sampling design. However, there are differences between the two. In the stratified sampling design, the selection of sample from each stratum is random but in the quota sampling, the respondents may be chosen at the convenience or judgement of the researchers.

NOTES

10.7 KEY WORDS

- **Quota sampling:** Quota sampling is a non-probability sampling technique wherein the assembled sample has the same proportions of individuals as the entire population with respect to known characteristics, traits or focused phenomenon.
- **Sample size:** Sample size determination is the act of choosing the number of observations or replicates to include in a statistical sample. The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample.
- **Snowball sampling:** Snowball sampling is generally used when it is difficult to identify the members of the desired population, e.g., deep-sea divers, families with triplets, people using walking sticks, doctors specializing in a particular ailment, etc.

10.8 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. How do you distinguish between probability sampling and non-probability sampling?
2. Differentiate between the stratified random sampling and systematic sampling.
3. List the similarities and differences between the quota sampling and stratified sampling.
4. What is the main difference between a stratified sampling and cluster sampling?

Long-Answer Questions

1. What is the need of sampling? Discuss various probability sample techniques by giving their merits and demerits.

NOTES

2. Explain the meanings of sample and sample design. Briefly discuss some most of the popular sample designs used in research.
3. What is the significance of sample selection in research? Explain the factors which should be considered while selecting a sample for research.
4. What is sampling? Discuss different sampling methods.
5. What is a research design? Discuss the basis of stratification to be employed in sampling a public opinion on inflation.

10.9 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

UNIT 11 COLLECTION OF DATA

Structure

- 11.0 Introduction
- 11.1 Objectives
- 11.2 Primary and Secondary Data
- 11.3 Sources of Secondary Data
- 11.4 Methods of data Collection
- 11.5 Content Analysis
- 11.6 Case Study
- 11.7 Answers to Check Your Progress Questions
- 11.8 Summary
- 11.9 Key Words
- 11.10 Self Assessment Questions and Exercises
- 11.11 Further Readings

NOTES

11.0 INTRODUCTION

Correlation or content 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.

Case study method enables a researcher to closely examine the data within a specific context. In most cases, a case study method selects a small geographical area or a very limited number of individuals as the subjects of study. Case studies, in their true essence, explore and investigate contemporary real-life phenomenon through detailed contextual analysis of a limited number of events or conditions, and their relationships.

11.1 OBJECTIVES

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

- Discuss the different sources of secondary data
- Describe the methods of data collection with examples
- Differentiate between scatter diagram method and least square method
- Analyse the different steps of the case study method

NOTES

11.2 PRIMARY AND SECONDARY DATA

Though a reference to these two data types may appear a little out of context here, yet it is pertinent to distinguish between secondary and primary data even at this juncture. The distinction between the two needs to be understood in the light of a well-defined research investigation for which specific data are needed for analysis. As the required data may be drawn, compiled or collected from different sources, a correct identification of the source(s) becomes important.

Data sources could be seen as of two types, viz., secondary and primary. A data source is secondary if it already contains the needed data in one form or the other. A primary data source is invariably a sample or a population survey, undertaken solely with a view to collecting the needed data when the same are not available from an existing data source.

Thus, depending on the broad classification of possible data sources as being primary or secondary, statistical data may be distinguished as secondary data and primary data. Thus, the two can be defined as under:

- i. **Secondary data** are secondary in the sense that they already exist in some form—published or unpublished—in an identifiable secondary source. They are, generally, available from published source(s), though not necessarily in the form actually required.
- ii. **Primary data**, on the contrary, are those that do not already exist in any form, and thus have to be collected for the first time from the primary source(s). By their very nature, these data require fresh and first-time collection covering the whole population or a sample drawn from it.

11.3 SOURCES OF SECONDARY DATA

Once the data requirements of a given research study have been clearly identified, it becomes important to locate and reach the relevant data source(s). As the data needed are available from many sources, these may be categorized as i) external vs internal and ii) primary vs secondary sources.

- i. **External vs Internal Sources:** The distinction between external and internal data sources is based on who compiles the data and who mainly uses them. It is essentially from the point of view of the user that a data source is regarded as external or internal.

External sources consist largely of regular data publications of the central and state governments, including the various individual departments, who compile and bring out vast and varied data for meeting the needs of planning, monitoring, and evaluating different government activities. The compilers of these data (publications) are often not the users. Since the users are, generally,

different from the compilers, all such data publications may be said to constitute the external sources of data.

Internal sources mainly include all those data publications of one kind or the other which are brought out both by public and private organizations, corporations, trade bodies, companies, and other statutory agencies/institutions created under law. Most of these organizations and companies are required to compile and publish variety of data on their operations, mostly in the form of Annual Reports. The data presented in these reports are primarily meant for internal use, for monitoring and evaluating performance. The compilers of these data are the main users as well. Thus, all sources that provide data primarily for own use, to plan and control organizational operations, may be termed as internal sources of data.

- ii. **Primary vs Secondary Sources:** In another approach of looking at data sources, these can well be classified as primary and secondary sources. The difference between the two is delicate and slightly catchy, but functionally very pertinent and useful.

Primary sources are described as such from the point of view of an outside user, and include all data sources that are perceived as external and internal. Visualizing these sources as primary sources makes sense when the outside data users are understood as those not directly associated with the compilation and publication of any data. Among others, these data users include individual researchers, universities, and research organizations. Government and semi-government agencies, bodies, and institutions also importantly figure in the category of outside users.

They are the ones who are directly or indirectly engaged in using data for specific purposes and tasks. The Planning Commission working at the national level, and State Planning Boards at the state level, are the classic examples of such statutorily created institutional users in the Indian scenario. However, this does not altogether exclude the possibility of primary data being used by those who are mainly responsible for their compilation. But even as they do, they do it in a very limited way.

A hypothetical example will clarify the argument. A university teacher may use the data published in the RBI Bulletin in his research paper on a specific theme. The teacher-user of the required data in this case is an outside user, and the source from where data are taken is the primary source. Similarly, if the Institute of Economic Growth in New Delhi is working on a research project, say, on current demographic trends, the researchers engaged in the task are outside users and the sources from where the relevant data are drawn are the primary sources.

Secondary sources comprise all specific-purpose research papers and reports, whether published, mimeographed, or unpublished. Research papers and reports often provide useful data well organized and properly presented

NOTES

NOTES

in the light of the specific objectives of a research investigation on any particular theme. Significantly, the data used in these research papers and reports are normally those drawn from the primary sources. But when the needed data are taken from any such paper(s) and/or report(s), these are considered as having come from secondary sources.

An example will clarify the argument. A researcher may take the per capita income data from the Statistical Abstract of India, which is an external primary data source. He may draw data on the sales of a given product from the Annual Report of the company manufacturing that product, which is an internal primary data source. The research findings may get published somewhere, say, in the Economic and Political Weekly. If a subsequent user draws per capita income and sales data from this source, the Weekly is then regarded as a secondary source.

Interestingly, the above classification of data sources helps visualize these sources as forming two other categories. All data whether drawn from external primary or internal primary sources are largely general-purpose data sources. Those drawn from secondary sources, on the other hand, are invariably specific-purpose data sources.

11.4 METHODS OF DATA COLLECTION

A technique of data collection refers to the method by which we actually go about collecting the desired information in a survey. As information is always elicited from the respondents, three alternative techniques of data collection have come to be adopted. These are the same whether it is a sample survey or a population survey. As only one is chosen for the task, it is necessary to briefly describe the mechanics of each.

1. **Personal Interviews:** A personal interview can be held with the respondents comprising a population or a sample using a structured or unstructured questionnaire. A questionnaire, prepared invariably in advance, is essentially a list of questions through which the interviewer seeks information from the respondents at personal level.

As one of the two choices, each question may have adequate space provided there under to record the answer. A questionnaire so designed and prepared is called a **schedule**. Alternatively, a separate sheet may be used to record the responses. A list of questions that allows for collecting answers in this manner is called a questionnaire proper. The distinction between the two is, however, not much of a consequence. The two are frequently used interchangeably, since even a schedule is also basically a questionnaire, and vice-versa.

A questionnaire is structured when it consists of questions under each of which are recorded alternate possible answers. While interviewing, the

respondent is required to tick one of the suggested answers that best describes his position. Such questions do not give freedom to answer beyond the alternatives suggested. On the contrary, a questionnaire is unstructured when the listed questions do not carry suggested answers. Each question is a *free-response* question.

Personal interviewing is an often used technique with its own merits and demerits. One of the chief merits is that it facilitates better response. Quality of answers is also likely to be better. It is all because the interviewer gets an opportunity to make clarifications and do necessary explaining to the respondents, wherever needed. This also permits collection of additional information over and above what may have been intended through the pre-stated questions. Such additional information at times is extremely useful during the course of data analysis and interpretation.

The only potent demerit of personal interviewing is that it may bias the answers if the interviewer tends to be zealous and over-enthusiastic, and does unnecessary and uncalled for explaining. This happens particularly when the interviewer tries to unduly speed up the interviewing process for one reason or the other, and thus fails to allow sufficient time to the respondents to think and make up their mind in deciding answers. This becomes a serious demerit when interviewing is done through an unstructured questionnaire.

Imparting proper training to the interviewers before they are allowed to undertake the assigned task can greatly reduce the possibility of such dangers. Necessary training goes a long way in minimizing the element of interviewers' bias. It definitely improves the quality of responses in terms of reliability.

2. **Interviewing through Telephones:** When interviewing is done by approaching the respondents through telephone, it is called telephone interviewing or a telephone survey. This approach is extremely simple, and is conducted normally through unstructured question listing. One may resort to telephone interviewing where exactness in responses is not of much concern and the purpose is to obtain quick results.

The telephone survey technique is frequently used especially where one wishes to know the reach of TV programs introduced for the first time. It can also be fruitfully employed where interest is limited to collecting broad reactions of viewers to a specifically targeted televised program. This technique pre-supposes that the target audiences own telephones and have watched the concerned television program.

3. **Mailed Questionnaire:** In this technique, the questionnaire is mailed out to the respondents. They are allowed adequate time to fill the questionnaire and send it back by post. It is more convenient to adopt this method when the number of respondents to be reached is relatively small. It also greatly

NOTES

NOTES

suits when the respondents are widely scattered so that mail is the only convenient means to contact them. Respondents are also supposed to be educated enough to read and write, and can understand the issues involved. The questionnaire often used is of structured type.

A serious disadvantage of this method is that it fails to elicit good response. It has been discovered in practice that the response is normally well below 50 per cent. The quality of data obtained also leaves much to be desired. Indifference to requests for early response is the greatest challenge to overcome. Moreover, biases and prejudices of all sorts tend to enter into replies filed through the questionnaire. At times, their influence becomes serious enough to adversely affect the very reliability of answers. This occurs particularly where the questionnaire is designed to dwell upon and bring out social and cultural issues.

Preparing a Questionnaire

The quality and quantity of response data collected through a questionnaire largely depends on how best the questionnaire is prepared and designed. Well thought-over and objectively stated questions are more effective in fetching quick response. Experience in conducting research investigations helps a lot in this regard. There are, however, a few important points that do deserve careful attention while framing questions. The more important among these pertain to the following:

- (i) **Nature of questions:** Deciding the nature of questions to be included in a questionnaire needs utmost attention. A question may run a two-way answer form when the expected response is one of the two alternatives provided. *For example*, a given question may be answered in yes or no, correct or incorrect, effective or ineffective, and so on.

In another situation, one may choose to have *multiple-answer* choice questions. Each of such questions offers a number of relevant alternative answers to the respondent. He is called upon to freely indicate his position by ticking one of the answer choices that are provided. *For example*, consider the question: *Why have you joined the MBA course?* The more relevant and realistic multiple-answer choices may be as follows:

- (a) that the course offers better placement opportunities,
- (b) that the course offers a long-term career,
- (c) that the course is geared to overall personality development,
- (d) that the course has highly motivated and committed faculty, or
- (e) any other.

In yet another choice, a question may be stated so as to give unrestricted freedom of answer to the respondents. Such questions are called *free-choice* or *open-ended* questions. *For example*, it could be a question like: *What are your reactions to the new taxes proposed in the current budget?*

In answering such questions the respondent enjoys full freedom to offer an answer any way he feels. As a variety of answers could be expected to free-choice questions, subsequent classification and data presentation in a meaningful way become difficult to handle.

- (ii) **Language and wording of questions:** Questions should be framed and stated in a language which the respondents are adequately conversant with, and in which they feel quite at ease while answering. As far as possible, technical terms having uncommon meaning should be avoided. Else, specific terms and concepts may be properly explained, or better replaced by readily understandable expressions. Expressions that are likely to be misunderstood, or not understood properly, should be avoided. Otherwise, the same question may mean different things to different respondents. No question should be framed and put in a manner that makes the respondents inconvenient, or otherwise puts them in an embarrassing position. Posing *leading questions* is sternly barred as a fair rule of a good game. A leading question is one which is suggestive of the possible answer.
- (iii) **Ordering of questions:** Questions to follow one after the other should be systematically and objectively ordered. Those pertaining to a given topic or a particular theme should be at one place. They should not be allowed to mix with those on a different topic or another theme. *For example*, questions seeking answers on one's leisure habits be put in one block, and those on consumption habits in the other block.

It is also advisable to order questions in a manner that one is sequel to other. That is, where relevant, the response to a given question should follow from the previous one in a natural way. This requires questions, as well as possible answers to questions, to be put in a sequential and relational order. *For example*, consider the question: Are you married? This should precede the question: How many children do you have? Putting these two questions in a reverse order will only be absurd.

Pre-Testing and Editing

It is always worth the effort involved if a questionnaire or a schedule is tested with the prospective respondents as a trial before it is put to final use. This is called *pre-testing* of the questionnaire. Pre-testing is an important exercise which enables the researcher to have the last opportunity of testing the questionnaire for its suitability and worthiness. Pre-testing also helps check the questionnaire for mistakes, inconsistencies, and inadequacies.

It allows the much-needed last chance to ensure that the questions are correctly understood and are capable of fetching adequate response. On being put on trial, quite a few things get noticed and are attended to. This improves the overall quality of the questionnaire as an effective instrument for eliciting the desired information.

NOTES

NOTES

Once the survey work is over, all the completed schedules require thorough editing to look for possible mistakes in entering responses, gaps that may have been left out, information that would have remained incomplete, and other flaws with which the questionnaire may have been received back. This means last minute pruning of the filed-up questionnaires to decide which among these should be retained for gathering, organizing, and tabulating the information required for the purpose in hand.

Finally, it may be appreciated that no survey-based primary data offer adequately reliable results unless the research investigation at hand is properly conceived, effectively designed, and meticulously implemented. Preparing the schedule and getting them filled require much serious thought and imagination. Without that, the data collected may lack the needed reliability to an extent that seriously impairs the worth of the inferences that may be drawn.

Check Your Progress

1. What is a questionnaire?
2. What are free choice questions?

11.5 CONTENT ANALYSIS

In the field of content analysis, there are several methods that aim to compare texts and analyze them to determine whether they are equivalent. Within the content analysis, we found techniques used to compare and correlate texts.

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, 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 relationship 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 (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. But, 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, etc. are examples of correlation, since in all such cases the change in one phenomenon is explained by a change in 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. But 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.

Table 11.1 *Nature of Correlation*

<i>Changes in Independent Variable</i>	<i>Changes in Dependent Variable</i>	<i>Nature of Correlation</i>
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.

NOTES

We now explain, illustrate and interpret the said two coefficients concerning the relationship between two variables as under:

The Coefficient of Determination

NOTES

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., $\sum (Y - \hat{Y})^2$, technically known as the unexplained variation.
- (ii) The variation of the Y values around their own mean viz., $\sum (Y - \bar{Y})^2$, 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)

$$\begin{aligned} &= \sum (Y - \bar{Y})^2 - \sum (Y - \hat{Y})^2 \\ &= \sum (\hat{Y} - \bar{Y})^2 \end{aligned}$$

The Total and Explained as well as Unexplained variations can be shown as given in Figure 11.1.

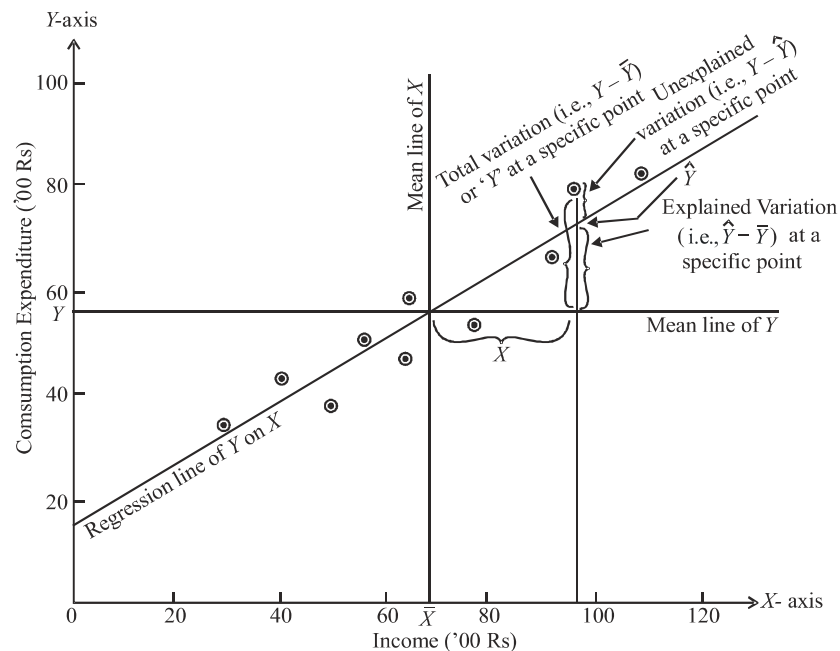


Fig. 11.1 Diagram Showing Total, Explained and Unexplained Variations

Coefficient 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:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

Alternatively, r^2 can also be stated as under:

$$r^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

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, $\sum(Y - \bar{Y})^2 = \sum(Y - \hat{Y})^2$, 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. But 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.

NOTES

Example 11.1: Calculate the coefficient of determination (r^2) using data given below. Calculate and analyse the result.

NOTES

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:

$$\text{Since, } r^2 = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}} = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum(Y - \bar{Y})^2}$$

As, $\sum(Y - \bar{Y})^2 = \sum Y^2 - n\bar{Y}^2$, we can write,

$$r^2 = 1 - \frac{\sum(Y - \hat{Y})^2}{\sum Y^2 - n\bar{Y}^2}$$

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

Regression Equations

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. 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.

NOTES

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 experimenter.
- (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.

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.

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).

NOTES

(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 11.2) to get the following information:

<i>Income</i>	<i>Consumption Expenditure</i>
X	Y
<i>(Hundreds of Rupees)</i>	<i>(Hundreds of Rupees)</i>
41	44
65	60
50	39
57	51
96	80
94	68
110	84
30	34
79	55
65	48

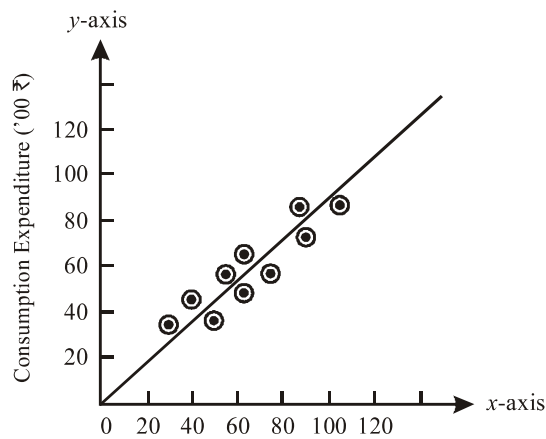


Fig. 11.2 Scatter Diagram

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 . But 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.

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 11.3 drawn below, where Figure 11.2 in scatter diagram has been reproduced along with a line which represents the least squares line fit to the data.

NOTES

NOTES

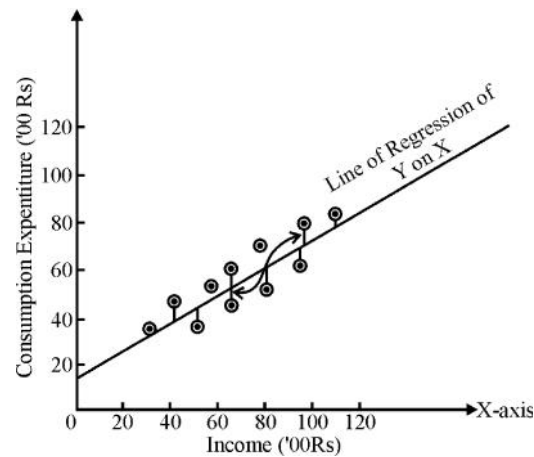


Fig. 11.3 Scatter Diagram, Regression Line and Short Vertical Lines Representing 'e'

Figure 11.2, 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 $\sum(Y - \hat{Y})$ or $\sum_{i=1}^n e_i$, to be as small as possible. However, mere examining $\sum(Y - \hat{Y})$ 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:

$$\begin{aligned}\sum Y &= na + b\sum X \\ \sum XY &= a\sum X + b\sum X^2\end{aligned}$$

In the above two equations, 'a' and 'b' are unknowns and all other values viz., $\sum X$, $\sum Y$, $\sum X^2$, $\sum XY$, 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 11.2: 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 $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ values from the given sample information table on Summations for Regression Equation.

<i>Summations for Regression Equation</i>					
<i>Observations</i>	<i>Income X ('00 ₹)</i>	<i>Consumption Expenditure Y ('00 ₹)</i>	<i>XY</i>	<i>X²</i>	<i>Y²</i>
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
<i>n = 10</i>	$\sum X = 687$	$\sum Y = 563$	$\sum XY = 42358$	$\sum X^2 = 53173$	$\sum Y^2 = 34223$

Putting the values in the required normal equations we have,

$$563 = 10a + 687b$$

$$42358 = 687a + 53173b$$

NOTES

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.

NOTES

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 $\sum(Y - \hat{Y})$ 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 11.2, we worked out the regression equation for the income and consumption data as,

$$\hat{Y} = 14.000 + 0.616X_i$$

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. But 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 and 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.

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 $\hat{Y} \pm 1 SE_e$ limits

95.5% of all points within $\hat{Y} \pm 2 SE_e$ limits

99.7% of all points within $\hat{Y} \pm 3 SE_e$ limits

NOTES

This can be stated as,

(i) The observed values of Y are normally distributed around each estimated value of \hat{Y} .

NOTES

(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:

$$\text{Upper limit} = \hat{Y} + 't' (SE_e)$$

$$\text{Lower limit} = \hat{Y} - 't' (SE_e)$$

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:

$$(\hat{Y} - \bar{Y}) = r \frac{\sigma_Y}{\sigma_X} (X_i - \bar{X})$$

or,

$$\hat{Y} = r \frac{\sigma_Y}{\sigma_X} (X_i - \bar{X}) + \bar{Y}$$

Where, r = Coefficient of simple correlation between X and Y .

σ_Y = Standard deviation of Y .

σ_X = Standard deviation of X .

\bar{X} = Mean of X .

\bar{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} .

$$= \frac{\sum XY - n\bar{X}\bar{Y} \times \sqrt{\sum Y^2 - n\bar{Y}^2}}{\sqrt{\sum Y^2 - n\bar{Y}^2} \sqrt{\sum X^2 - n\bar{X}^2} \sqrt{\sum X^2 - n\bar{X}^2}}$$

$$= \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

And,

$$= -r \frac{\sigma_Y}{\sigma_X} \bar{X} + \bar{Y}$$

$$= \bar{Y} - b\bar{X} \quad \left(\text{since } b = r \frac{\sigma_Y}{\sigma_X} \right)$$

Similarly, the estimating equation of X , also known as the regression equation of X on Y , can be stated as:

$$(\hat{X} - \bar{X}) = r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y})$$

or,

$$\hat{X} = r \frac{\sigma_X}{\sigma_Y} (Y - \bar{Y}) + \bar{X}$$

And the Regression coefficient of X on Y (or b_{XY}) = $r \frac{\sigma_X}{\sigma_Y} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$

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., \bar{X}) and the mean value of Y (i.e., \bar{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:

$$r = \sqrt{b_{YX} \cdot b_{XY}}$$

$$= \sqrt{r \frac{\sigma_Y}{\sigma_X} \cdot r \frac{\sigma_X}{\sigma_Y}}$$

$$= \sqrt{r \cdot r} = r$$

The (\pm) 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 11.3: Given is the following information:

	\bar{X}	\bar{Y}
Mean	39.5	47.5
Standard Deviation	10.8	17.8

NOTES

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,

$$\therefore (\hat{Y} - \bar{Y}) = r \frac{\sigma_Y}{\sigma_X} (X_i - \bar{X})$$

$$\begin{aligned} \text{or, } \hat{Y} &= r \frac{\sigma_Y}{\sigma_X} (X_i - \bar{X}) + \bar{Y} \\ &= 0.42 \frac{17.8}{10.8} (X_i - 39.5) + 47.5 \\ &= 0.69X_i - 27.25 + 47.5 \\ &= 0.69X_i + 20.25 \end{aligned}$$

Similarly, the estimating equation of X can be worked out as under:

$$\therefore (\hat{X} - \bar{X}) = r \frac{\sigma_X}{\sigma_Y} (Y_i - \bar{Y})$$

$$\begin{aligned} \text{or, } \hat{X} &= r \frac{\sigma_X}{\sigma_Y} (Y_i - \bar{Y}) + \bar{X} \\ \text{or, } &= 0.42 \frac{10.8}{17.8} (Y_i - 47.5) + 39.5 \\ &= 0.26Y_i - 12.35 + 39.5 \\ &= 0.26Y_i + 27.15 \end{aligned}$$

Example 11.4: 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:

$$4X - 5Y + 33 = 0 \quad (1)$$

$$20X - 9Y - 107 = 0 \quad (2)$$

If we multiply Equation (1) by 5, we have the following equations:

$$20X - 25Y = -165 \quad (3)$$

$$20X - 9Y = 107 \quad (2)$$

$$\begin{array}{r} - \quad + \quad - \\ \hline -16Y = -272 \end{array}$$

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, $\bar{X} = 13$ and $\bar{Y} = 17$

(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,

$$r = \sqrt{5/4 \times 20/9} = \sqrt{25/9} = 5/3 = 1.6$$

NOTES

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:

NOTES

$$\hat{X} = \frac{9Y_i}{20} + \frac{107}{20}$$

And,

$$\hat{Y} = \frac{4X_i}{5} + \frac{33}{5}$$

Hence,

$$\begin{aligned} r &= \sqrt{9/20 \times 4/5} \\ &= \sqrt{9/25} \\ &= 3/5 \\ &= 0.6 \end{aligned}$$

Since, regression coefficients have plus signs, we take $r = + 0.6$

(iii) Standard deviation of Y can be calculated as follows:

$$\therefore \text{Variance of } X = 9$$

$$\therefore \text{Standard deviation of } X = 3$$

$$\therefore b_{YX} = r \frac{\sigma_Y}{\sigma_X} = \frac{4}{5} = 0.6 \frac{\sigma_Y}{3} = 0.2\sigma_Y$$

$$\text{Hence, } \sigma_Y = 4$$

Alternatively, we can work it out as under:

$$\therefore b_{XY} = r \frac{\sigma_X}{\sigma_Y} = \frac{9}{20} = 0.6 \frac{\sigma_X}{\sigma_Y} = \frac{1.8}{\sigma_Y}$$

$$\text{Hence, } \sigma_Y = 4$$

11.6 CASE STUDY

Case studies are discussions of individual cases under topics of discussion which help researchers to corroborate known facts proved previously through research. Social scientists, in particular, used the case study method to conduct research for many years. A variety of disciplines used this method of research to corroborate their findings in real life situations. Researcher Robert K. Yin defines the case study research method as an empirical inquiry that investigates a contemporary phenomenon within its real-life context; when the boundaries between phenomenon and context are not clearly evident; and in which multiple sources of evidence are used (Yin, 1984, p. 23).

However, critics feel that the case study method is not reliable enough for establishing a rule or principle as it portrays only a minuscule population which forms not even a part of the entire population. Some feel that this method is only a reliable exploratory tool. Literature supports reports of carefully planned and crafted studies of the case study method. Robert E. Stake, Helen Simmons, and

Robert Yin are renowned researchers who have written about the utility of case studies in social sciences. They have prescribed six steps that should be used when utilizing the case study method. These are:

- Determine and define the research questions
- Select the cases and determine data gathering and analysis techniques
- Prepare to collect the data
- Collect data in the field
- Evaluate and analyse the data
- Prepare the report

1. Determine and define the research questions

Before a case study research is undertaken, cementing a research focus is important so that the researcher can refer to it during the course of study. The research object is often a person, an organizational policy, a group of people, etc. A number of data gathering methods are used by the researcher who studies every case study in depth. The researcher reads the available literature to understand where the topic stands in terms of prior research and undertakes a thorough planning before embarking on the actual case study. Literature and previous studies help him to decide where to look for evidence to corroborate his findings on the concerned topic. These help in designing the blueprint for the current study.

2. Select the cases and determine data gathering and analysis techniques

While designing the study, researchers finalize the approaches, methods of data extraction and data gathering for real-life cases that they need to study. While using multiple cases, each case is treated as a single case. The conclusions of these cases can then be utilized for underlining various facets of their study. The researchers need to discriminate positively for the case study that they want to utilize for corroborating their findings. Researchers should decide whether they want to study cases that are conventional or extraordinary while conducting the study. In case they are hesitant, they may go back to the purpose of the study that they had enumerated before beginning the research. The decision to choose a single or multiple case studies is an important one, while a single case study may be examined for analysing more than one inherent principle. These types of case studies involve two different levels of analysis which increases the complexity of data collected. Multiple sources and techniques in the data collecting process is a key strength of the case study method. Researchers need to determine what data they would wish to gather by examining a case and how to analyse the data collection. The tools they may use are interviews, surveys, documentation review, observation and collection of physical artifacts. During the design phase of the research, researchers should make sure that the study ensures construct validity, external validity, internal validity and reliability. Researchers need to use the correct measures for ensuring construct validity. Internal validity is ensured when the conditions may be used over and over again to prove validity of the case. External validity is

NOTES

NOTES

ensured when the findings may be generalized beyond the case or cases. A case study is said to be more externally valid when it can withstand more people, places and procedures. Techniques known as within-case examination and cross-case examination and literature review help ensure the validity of the case.

3. Prepare to collect the data

Researchers using the case study method generally gather a large amount of data from a number of sources. Organizing this data in a systematic manner is a challenge in itself. Researchers should plan ahead to prevent getting overwhelmed by this data. They might even lose sight of the original purpose of gathering the data. Researchers sort, categorize, store and retrieve data for analysis with the help of databases. Extraordinary cases help researchers by providing an efficient training programme, establishing proper protocols and conducting a pilot study before entering fieldwork. The training programme covers the concept to be studied, terminology, processes, methods, etc. The researchers also learn the application of techniques used in the study. In order to gather data from the interviewed population, researchers have to be skilled enough to retain or record the interviews without the gadget coming in the interviewee's way. Researchers should know how to steer conversation towards the questions they intend to ask next. They should be trained in analysing body language and interpret answers not expected by them. Researchers need to read between the lines and in case the topic is sensitive, understand a respondent's hesitation and silence. Researchers should not feel threatened by missed appointments and lack of space for holding the interview or unexpected turns of events during the interview; for example, a respondent may break down while answering a sensitive question. Researchers should be humane, understanding and flexible in approach. They should revisit the research design that they had created before starting the case studies and make changes as and when required.

4. Collect data in the field

Researchers should be trained to collect and store multiple sources of evidence in various formats while going about studying the case. Though case study research is flexible, any change that comes up needs to be documented carefully. The multiple storing of data is required so that converging lines of enquiry and patterns may be discovered. Field notes may be used for recording intuitions, hunches, feelings, and also for documenting the work in progress. Illustrations, anecdotes and special records may be written in the field notes so that the researcher may refer to it when making case study reports. The data and the field notes should be kept separately for analysis. The researcher needs to document, classify and cross-refer all evidence so that these could be efficiently recalled for examination and sorting as and when required.

5. Evaluate and analyse the data

The raw data gathered by the researchers need to be interpreted at different levels to find linkages between the objectives of the research and the outcome of studying

the case. Researchers must remain open to new insights and opportunities throughout the evaluation and analysis process. They can triangulate data with the help of different techniques and collection methods inherent to the case study method. Researchers will be provided with new insights and conflicting data by case studies which are extraordinary. They would need to categorize, tabulate and combine data to address the purpose of the study. In order to cross-check data collected, short, repeated interviews need to be conducted. Placing information into arrays, creating matrices of categories, making flow charts or other displays, etc., may be used by the researcher as specific techniques. The quantitative data collected may be used to corroborate the qualitative data collected during interviews. Many research organizations may also use multiple researchers to verify the data collected. When these multiple observations converge, researchers may become more confident of their findings. Conflicting observations need in-depth study of the findings. The cross-case search technique requires that researchers look at data from different angles and do not reach a premature conclusion. Across all cases investigated, the cross-case search divides data by type. When a pattern from one data is vouched for by another data, the finding is stronger. When these evidences do not form a data, a further probe is essential.

NOTES

6. Prepare the report

An exemplary case study report transforms the manner in which a complex issue is presented. Case study reports are often published so that readers may apply the experience in their real-life situations. Case studies mostly display evidences to gain the confidence of the readers. Researchers also underline the boundaries of the case and draw the attention of the readers to conflicting propositions. Many researchers present case study reports in the form of a chronological account. Some may treat a case as a fresh chapter. Once a report is completed, the researcher should always edit and examine it for loopholes. Representative audience group is used for comments and criticisms and the valid criticisms are incorporated in the next draft. Since case studies involve multiple sources of data, or may include more than one case within a study, they often become complex. The case study method is generally used by researchers from various disciplines to build upon a theory, to produce a new theory, to challenge or dispute a theory, to explore new horizons, to apply solutions to situations, to describe a phenomenon, etc. There are a number of advantages of the case-study method. These are: applicability to real life situations, to contemporary social situations and easy accessibility to its published reports. Case studies help common man understand a complex theory through easy, real-life situations that are used to exemplify the principle being discussed.

Check Your Progress

3. What are the six steps of the case study method?
4. What is a least square method?

11.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. A questionnaire, prepared invariably in advance, is essentially a list of questions through which the interviewer seeks information from the respondents at personal level.
2. A question may be stated so as to give unrestricted freedom of answer to the respondents. Such questions are called *free-choice* or *open-ended* questions.
3. The six steps that should be used when utilizing the case study method are:
 - Determine and define the research questions
 - Select the cases and determine data gathering and analysis techniques
 - Prepare to collect the data
 - Collect data in the field
 - Evaluate and analyse the data
 - Prepare the report
4. 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.

11.8 SUMMARY

- The distinction between the two needs to be understood in the light of a well-defined research investigation for which specific data are needed for analysis. As the required data may be drawn, compiled or collected from different sources, a correct identification of the source(s) becomes important.
- Data sources could be seen as of two types, viz., secondary and primary.
- Once the data requirements of a given research study have been clearly identified, it becomes important to locate and reach the relevant data source(s). As the data needed are available from many sources, these may be categorized as i) external vs internal and ii) primary vs secondary sources.
- A technique of data collection refers to the method by which we actually go about collecting the desired information in a survey. As information is always elicited from the respondents, three alternative techniques of data collection have come to be adopted.
- A questionnaire is structured when it consists of questions under each of which are recorded alternate possible answers. While interviewing, the respondent is required to tick one of the suggested answers that best describes his position.

- The quality and quantity of response data collected through a questionnaire largely depends on how best the questionnaire is prepared and designed. Well thought-over and objectively stated questions are more effective in fetching quick response.
- In the field of content analysis, there are several methods that aim to compare texts and analyse them to determine whether they are equivalent.
- Statisticians have developed *two measures for describing the correlation* between two variables, viz., the coefficient of determination and the coefficient of correlation.
- 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.
- 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.
- Case studies are discussions of individual cases under topics of discussion which help researchers to corroborate known facts proved previously through research.

NOTES

11.9 KEY WORDS

- **Theory of correlation:** The theory by means of which quantitative connections between two sets of phenomena are determined is called the ‘*Theory of Correlation*’.
- **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.

11.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Write a short note on primary and secondary data.
2. Write short notes on:
 - Interview

- Questionnaire
- Schedule

3. How is the coefficient of determination interpreted?

4. State the assumptions of regression analysis.

NOTES

Long-Answer Questions

1. Analyse the different sources of secondary data.
2. Describe the methods of data collection with examples.
3. Differentiate between scatter diagram method and least square method.
4. Analyse the different steps of the case study method.

11.11 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

BLOCK - IV
MEASUREMENT AND SCALING TECHNIQUES
MEASURE OF CENTRAL TENDENCY

*Measurement and
Scaling Techniques*

NOTES

**UNIT 12 MEASUREMENT AND
SCALING TECHNIQUES**

Structure

- 12.0 Introduction
- 12.1 Objectives
- 12.2 Meaning, Problems and Methods
 - 12.2.1 Types of Measurement Scale
- 12.3 Methods of Scale Construction
 - 12.3.1 Criteria for Good Measurement: Reliability and Validity
 - 12.3.2 Thurstone, Guttman and Bogardus Scales
- 12.4 Answers to Check Your Progress Questions
- 12.5 Summary
- 12.6 Key Words
- 12.7 Self Assessment Questions and Exercises
- 12.8 Further Readings

12.0 INTRODUCTION

There are four levels of measurements: nominal, ordinal, interval, and ratio. The measurement scales, commonly used in marketing research, can be divided into two types; comparative and non-comparative scales.

A number of scaling techniques are available for measurement of attitudes which have been discussed in detail in this unit.

12.1 OBJECTIVES

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

- Discuss the various techniques of scaling
- Describe the different methods of scale construction
- Discuss the criteria for evaluating measurements

12.2 MEANING, PROBLEMS AND METHODS

The term 'measurement' means assigning numbers or some other symbols to the characteristics of certain objects. When numbers are used, the researcher must

NOTES

have a rule for assigning a number to an observation in a way that provides an accurate description. We do not measure the object but some characteristics of it. Therefore, in research, people/consumers are not measured; what is measured only are their perceptions, attitude or any other relevant characteristics. There are two reasons for which numbers are usually assigned. First of all, numbers permit statistical analysis of the resulting data and secondly, they facilitate the communication of measurement results.

As mentioned earlier, the numbering is done based on certain rules. Therefore, the assignment of numbers to the characteristics must be isomorphic, i.e., there must be a one-to-one correspondence between the numbers and the characteristics being measured.

For example, same rupee figures should be assigned to a household with identical annual income. Only then numbers can be associated with specific characteristics of the measured object and vice versa. Further, they must not change over the objects or time. This means that the rules for a given assignment must be invariant over time or the object being measured.

Scaling is an extension of measurement. Scaling involves creating a continuum on which measurements on objects are located. Suppose you want to measure the satisfaction level towards Jet-Airways Airlines and a scale of 1 to 11 is used for the said purpose. This scale indicates the degree of dissatisfaction, with 1 = extremely dissatisfied and 11 = extremely satisfied. Measurement is the actual assignment of a number from 1 to 11 to each respondent whereas the scaling is the process of placing the respondent on a continuum with respect to their satisfaction towards Jet Airways.

12.2.1 Types of Measurement Scale

There are four types of measurement scales—nominal, ordinal, interval and ratio scales. We will discuss each one of them in detail. The choice of the measurement scale has implications for the statistical technique to be used for data analysis.

Nominal scale: This is the lowest level of measurement. Here, numbers are assigned for the purpose of identification of the objects. Any object which is assigned a higher number is in no way superior to the one which is assigned a lower number. In the nominal scale there is a strict one-to-one correspondence between the numbers and the objects. Each number is assigned to only one object and each object has only one number assigned to it. It may be noted that the objects are divided into mutually exclusive and collectively exhaustive categories.

Examples of nominal scale:

- What is your religion?
 - (a) Hinduism
 - (b) Sikhism
 - (c) Christianity

- (d) Islam
- (e) Any other, (please specify)

A Hindu may be assigned a number 1, a Sikh may be assigned a number 2, a Christian may be assigned a number 3 and so on. Any religion which is assigned a higher number is in no way superior to the one which is assigned a lower number. The assignment of numbers is only for the purpose of identification. We also note that all respondents have been divided into mutually exclusive and collectively exhaustive categories. For example:

- Are you married?
 - (a) Yes
 - (b) No

If a person is married, he or she may be assigned a number 101 and an unmarried person may be assigned a number 102.

- In which of the following departments do you work?
 - (a) Marketing
 - (b) HR
 - (c) Information Technology
 - (d) Operations
 - (e) Finance and Accounting
 - (f) Any other, (please specify)

Here also, a person working for the marketing department may be assigned a number 1, the one working for HR may be assigned a number 2 and so on.

Nominal scale measurements are used for identifying food habits (vegetarian or non-vegetarian), gender (male/female), caste, respondents, brands, attributes, stores, the players of a hockey team and so on.

The assigned numbers cannot be added, subtracted, multiplied or divided. The only arithmetic operations that can be carried out are the count of each category. Therefore, a frequency distribution table can be prepared for the nominal scale variables and mode of the distribution can be worked out. One can also use chi-square test and compute contingency coefficient using nominal scale variables.

Ordinal scale: This is the next higher level of measurement than the nominal scale measurement. One of the limitations of the nominal scale measurements is that we cannot say whether the assigned number to an object is higher or lower than the one assigned to another option. The ordinal scale measurement takes care of this limitation. An ordinal scale measurement tells whether an object has more or less of characteristics than some other objects. However, it cannot answer how much more or how much less. An ordinal scale tells us the relative positions of the objects and not the difference between the magnitudes of the objects. Suppose Shashi scores the highest marks in marketing and is ranked no. 1; Mohan scores the

NOTES

NOTES

second highest marks and is ranked no. 2; and Krishna scores third highest marks and is ranked no. 3. However, from this statement we cannot say whether the difference in the marks scored by Shashi and Mohan is the same as between Mohan and Krishna. The only statement which can be made under ordinal scale is that Shashi has scored higher than Mohan and Mohan has scored higher than Krishna. The difference between the ranks does not have any meaningful interpretation in the sense that it cannot tell the difference in absolute marks between the three candidates. Another example of the ordinal scale could be the CAT score given in percentile form. Suppose a candidate's score is 95 percentile in the CAT exam. What it means is that 95 per cent of the candidates that appeared in the CAT examination have a score below this candidate, whereas only 5 per cent have scored more than him. The actual score is how much less or more cannot be known from this statement. Examples of the ordinal scale include quality ranking, rankings of the teams in a tournament, ranking of preference for colours, soft drinks, socio-economic class and occupational status, to mention a few. Some of the examples of ordinal scales are listed below:

- Rank the following attributes while choosing a restaurant for dinner. The most important attribute may be ranked one, the next important may be assigned a rank of 2 and so on.

Attribute	Rank
Food quality	
Prices	
Menu variety	
Ambience	
Service	

- Rank the following by placing a 1 beside the attribute you think is the most important, a 2 beside the attribute you think is the second most important and so on while purchasing a two-wheeler.

Attribute	Rank
After sale service	
Prices	
Re-sale value	
Fuel efficiency	
Aesthetic appeal	

In the ordinal scale, the assigned ranks cannot be added, multiplied, subtracted or divided. One can compute median, percentiles and quartiles of the distribution. The other major statistical analysis which can be carried

out is the rank order correlation coefficient, sign test. As the ordinal scale measurement is higher than the nominal scale measurement, all the statistical techniques which are applicable in the case of nominal scale measurement can also be used for the ordinal scale measurement. However, the reverse is not true. This is because ordinal scale data can be converted into nominal scale data but not the other way round.

Interval scale: The interval scale measurement is the next higher level of measurement. It takes care of the limitation of the ordinal scale measurement where the difference between the score on the ordinal scale does not have any meaningful interpretation. In the interval scale the difference of the score on the scale has meaningful interpretation. It is assumed that the respondent is able to answer the questions on a continuum scale. The mathematical form of the data on the interval scale may be written as

$$Y = a + bX \quad \text{where } a \neq 0$$

The interval scale data has an arbitrary origin (non-zero origin). The most common example of the interval scale data is the relationship between Celsius and Fahrenheit temperature. It is known that:

$$C^0 = \frac{5}{9}(F^0 - 32)$$

Therefore,
$$C^0 = \frac{-160}{9} + \frac{5}{9}F^0$$

This is of the form $Y = a + bX$, where $a = \frac{-160}{9}$ and $b = \frac{5}{9}$ and hence it represents the interval scale measurement. In the interval scale, the difference in score has a meaningful interpretation while the ratio of the score on this scale does not have a meaningful interpretation. This can be seen from the following interval scale question:

- How likely are you to buy a new designer carpet in the next six months?

	Very unlikely	Unlikely	Neutral	Likely	Very likely
Scale A	1	2	3	4	5
Scale B	0	1	2	3	4
Scale C	-2	-1	0	1	2

Suppose a respondent ticks the response category 'likely' and another respondent ticks the category 'unlikely'. If we use any of the scales A, B or C, we note that the difference between the scores in each case is 2. Whereas, when the ratio of the scores is taken, it is 2, 3 and -1 for the scales A, B and C respectively. Therefore, the ratio of the scores on the scale does not have a meaningful interpretation. The following are some examples of interval scale data.

NOTES

NOTES

- How important is price to you while buying a car?

Least important	Unimportant	Neutral	Important	Most important
1	2	3	4	5

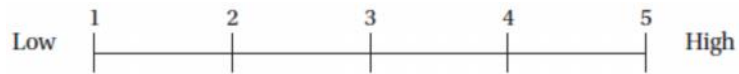
- How do you rate the work environment of your organization?

Very good	Good	Neither good nor bad	Bad	Very bad
5	4	3	2	1

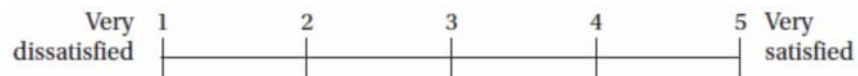
- The counter-clerks at ICICI Bank, (Vasant Kunj Branch) are very friendly.

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1	2	3	4	5

- Rate the life of the battery of your inverter.



- Indicate the degree of satisfaction with the overall performance of Wagon R.



- How expensive is the restaurant ‘Punjabi By Nature’?

Extremely expensive	Definitely expensive	Somewhat expensive	Somewhat inexpensive	Definitely inexpensive	Extremely inexpensive
1	2	3	4	5	6

- How likely are you to buy a new car within the next six months?

Definitely will buy	Probably will buy	Neutral	Probably will not buy	Definitely will not buy
1	2	3	4	5

The numbers on this scale can be added, subtracted, multiplied or divided. One can compute arithmetic mean, standard deviation, correlation coefficient and conduct a t-test, Z-test, regression analysis and factor analysis. As the interval scale data can be converted into the ordinal and the nominal scale data, therefore all the techniques applicable for the ordinal and the nominal scale data can also be used for interval scale data.

Ratio scale: This is the highest level of measurement and takes care of the limitations of the interval scale measurement, where the ratio of the measurements on the scale does not have a meaningful interpretation. The ratio scale measurement can be converted into interval, ordinal and nominal scale. But the other way round is not possible. The mathematical form of the ratio scale data is given by $Y = bX$. In this case, there is a natural zero (origin), whereas in the interval scale we had an arbitrary zero. Examples of the ratio scale data are weight, distance travelled,

income and sales of a company, to mention a few. Consider the following examples for ratio scale measurements:

- How many chemist shops are there in your locality?
- How many students are there in the MBA programme at IIFT?
- How much distance do you need to travel from your residence to reach the railway station?

All the mathematical operations can be carried out using the ratio scale data. In addition to the statistical analysis mentioned in the interval, the ordinal and the nominal scale data, one can compute coefficient of variation, geometric mean and harmonic mean using the ratio scale measurement. The basic characteristics, examples and the statistical techniques applicable under each of the four scales are summarized in Table 12.1.

Table 12.1 Types of Scale, Characteristics, Examples, Permissible Statistical Techniques

Scale	Basic Characteristics	Examples	Permissible Statistics
Nominal	Numbers are used to label and classify objects	Players of Team India, Caste, Religion, Gender, Marital Status, Store Types, Brands, etc.	Percentages, Mode, Chi-square, Contingency coefficient, Binomial test
Ordinal	Numbers indicate the relative position of the objects, however the difference in the magnitude of the score cannot be known	Preference Ranking, Image Ranking, Social Class, etc.	Percentile, Quartiles, Median, Rank order correlation, Friedman ANOVA
Interval	Difference between the objects can be known, however the ratio of the scores has no meaning	Attitude, Opinion, Index Numbers	Product moment correlation coefficient, t-test, z-test, ANOVA, Regression Analysis, Factor Analysis
Ratio	Ratios of the score value have a meaningful interpretation	Age, Income, Market Share, Sales, Cost, etc.	Geometric means, Harmonic Means and Coefficient of variation

Attitude

An attitude is viewed as an enduring disposition to respond consistently in a given manner to various aspects of the world, including persons, events and objects. A company is able to sell its products or services when its customers have a favourable attitude towards its products/services. In the reverse scenario, the company will not be able to sustain itself for long. It, therefore, becomes very important to measure the attitude of the customers towards the company's products/services. Unfortunately, attitude cannot be measured directly. There are many variables which the researcher

NOTES

NOTES

wishes to investigate as psychological variables and these cannot be directly observed. For example, we may have a favourable attitude towards a particular brand of toothpaste, but this attitude cannot be observed directly. In order to measure an attitude, we make an inference based on the perceptions the customers have about the product/services. The attitude is derived from the perceptions. If the consumers have a favourable perception towards the products/services, the attitude will be favourable. Therefore, the attitudes are indirectly observed.

Basically, attitude has three components: cognitive, affective and intention (or action) components.

Cognitive component: This component represents an individual's information and knowledge about an object. It includes awareness of the existence of the object, beliefs about the characteristics or attributes of the object and judgement about the relative importance of each of the attributes. In a survey, if the respondents are asked to name the companies manufacturing plastic products, some respondents may remember names like Tupperware, Modicare and Pearl Pet. This is called unaided recall awareness. More names are likely to be remembered when the investigator makes a mention of them. This is aided recall. It may be noted that the knowledge may not be limited only to the awareness. An individual can form beliefs or judgements about the characteristics or attributes of the plastic products manufacturing companies through advertisements, word of mouth, peer groups, etc. The examples of such beliefs could be that the products of Tupperware are of high quality, non-toxic and can be used in parties; a mutton dish can be cooked in a pressure cooker in less than 30 minutes; the Nano car gives a very high mileage as compared to the other small cars.

Affective component: The affective component summarizes a person's overall feeling or emotions towards the objects. The examples for this component could be: the food cooked in a pressure cooker is tasty, taste of orange juice is good or the taste of bitter gourd is very bad. If there are a number of alternatives to choose from, liking is expressed in terms of preference for one alternative over the other. Among the various soft drinks like Pepsi, Coke, Limca and Sprite, the respondents might have to indicate the most preferred soft drinks, the second preferred one and so on. This is an example of the affective component. The other example could be that the plastic products produced by Pearl Pet are cheaper than Tupperware products; however, the quality of Tupperware products is better than that of Pearl Pet.

Intention or action component: This component of an attitude, also called the behavioural component, reflects a predisposition to an action by reflecting the consumer's buying or purchase intention. It also reflects a person's expectations of future behaviour towards an object. How likely a person is to buy a designer carpet may range from most likely to not at all likely, reflecting the purchase intentions. However, when one is talking about the purchase intentions, a time horizon has to be kept in mind as the intentions may undergo a change over time. The intentions incorporate information regarding the respondent's willingness to pay for the product.

There is a relationship between attitude and behaviour. If a consumer does not have a favourable attitude towards the product, he/she will certainly not buy the product. However, having a favourable attitude does not mean that it would be reflected in the purchase behaviour. This is because intention to buy a product has to be backed by the purchasing power of the consumer. Having a favourable attitude towards Mercedes Benz does not mean that a person is going to purchase it even if he does not have the ability to buy a product. Therefore, the relationship between the attitude and the purchase behaviour is a necessary condition for the purchase of the product but it is not a sufficient condition. This relationship could hold true at the aggregate level but not at the individual level.

NOTES

Check Your Progress

1. What are the four types of measurement scales?
2. When is nominal measurement scale used?
3. What is a ratio scale?

12.3 METHODS OF SCALE CONSTRUCTION

One of the ways of classifications of scales is in terms of the number of items in the scale. Based upon this, the following classification may be proposed:

Single Item vs Multiple Item Scale

Single item scale: In the single item scale, there is only one item to measure a given construct. For example:

Consider the following question:

- How satisfied are you with your current job?
Very Dissatisfied
Dissatisfied
Neutral
Satisfied
Very satisfied

The problem with the above question is that there are many aspects to a job, like pay, work environment, rules and regulations, security of job and communication with the seniors. The respondent may be satisfied on some of the factors but may not on others. By asking a question as stated above, it will be difficult to analyse the problem areas. To overcome this problem, a multiple item scale is proposed.

Multiple item scale: In multiple item scale, there are many items that play a role in forming the underlying construct that the researcher is trying to measure. This is because each of the item forms some part of the construct (satisfaction) which the

NOTES

researcher is trying to measure. As an example, some of the following questions may be asked in a multiple item scale.

- How satisfied are you with the pay you are getting on your current job?
Very dissatisfied
Dissatisfied
Neutral
Satisfied
Very satisfied
- How satisfied are you with the rules and regulations of your organization?
Very dissatisfied
Dissatisfied
Neutral
Satisfied
Very satisfied
- How satisfied are you with the job security in your current job?
Very dissatisfied
Dissatisfied
Neutral
Satisfied
Very satisfied

Comparative vs Non-comparative Scales

The scaling techniques used in research can also be classified into comparative and non-comparative scales (Figure 12.1).

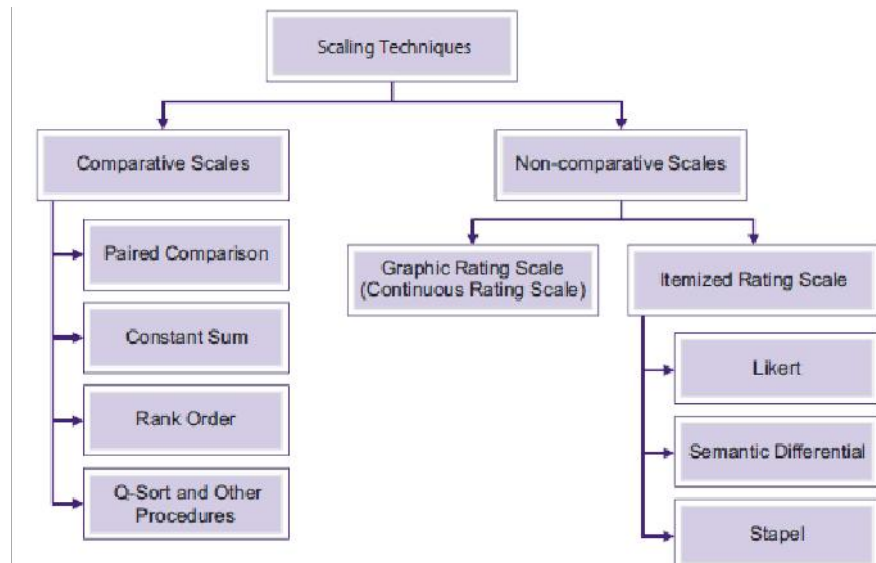


Fig. 12.1 Types of Scaling Techniques

Comparative Scales

In comparative scales it is assumed that respondents make use of a standard frame of reference before answering the question. For example:

A question like ‘How do you rate Barista in comparison to Cafe Coffee Day on quality of beverages?’ is an example of the comparative rating scale. It involves the direct comparison of stimulus objects. For example, respondents may be asked whether they prefer Chinese in comparison to Indian food. Consider the following set of questions generally used to compare various attributes of Domino’s Pizza and Pizza Hut.

- Please rate Domino’s in comparison to Pizza Hut on the basis of your satisfaction level on an 11-point scale, based on the following parameters: (1 = Extremely poor, 6 = Average, 11 = Extremely good). Circle your response:

a.	Variety of menu options	1	2	3	4	5	6	7	8	9	10	11
b.	Value for money	1	2	3	4	5	6	7	8	9	10	11
c.	Speed of service (delivery time)	1	2	3	4	5	6	7	8	9	10	11
d.	Promotional offers	1	2	3	4	5	6	7	8	9	10	11
e.	Food quality	1	2	3	4	5	6	7	8	9	10	11
f.	Brand name	1	2	3	4	5	6	7	8	9	10	11
g.	Quality of service	1	2	3	4	5	6	7	8	9	10	11
h.	Convenience in terms of takeaway location	1	2	3	4	5	6	7	8	9	10	11
i.	Friendliness of the salesperson on the phone	1	2	3	4	5	6	7	8	9	10	11
j.	Quality of packaging	1	2	3	4	5	6	7	8	9	10	11
k.	Adaptation of Indian taste	1	2	3	4	5	6	7	8	9	10	11
l.	Side orders/appetizers	1	2	3	4	5	6	7	8	9	10	11

Comparative scale data is interpreted generally in a relative kind. The comparative scale includes paired comparison, rank order, constant sum scale and Q-sort technique to mention a few.

We will discuss below each of the scales under comparative rating scales in detail:

Paired comparison scales: Here a respondent is presented with two objects and is asked to select one according to whatever criterion he or she wants to use. The resulting data from this scale is ordinal in nature. As an example, suppose a parent wants to offer one of the four items to a child—chocolate, burger, ice cream and pizza. The child is offered to choose one out of the two from the six possible pairs, i.e., chocolate or burger, chocolate or ice cream, chocolate or pizza, burger or ice cream, burger or pizza and ice cream or pizza. In general, if there are n items, the number of paired comparison would be $(n(n - 1)/2)$. Paired comparison technique is useful when the number of items is limited because it

NOTES

NOTES

requires a direct comparison and overt choice. In case the number of items to be compared is large (say 10), it would result in 45 paired comparisons which would further result in fatigue for the respondents. Further, in reality a respondent does not make the choice from two items at a time—there are multiple alternatives available to him.

There are many ways of analysing the paired comparison data. The analysis of paired comparison data would result in an ordinal scale and also in an interval scale measurement. This will be shown with the help of an example. Let us assume that there are five brands—A, B, C, D and E—and a paired comparison with two brands at a time is presented to the respondent with the option to choose one of them. As there are five brands, it will result in 10 paired comparisons. Suppose this is administered to a sample of 250 respondents with the results as presented in Table 12.2.

Table 12.2 Paired Comparison Data

	A	B	C	D	E
A	–	0.60	0.30	0.60	0.35
B	0.40	–	0.28	0.70	0.40
C	0.70	0.72	–	0.65	0.10
D	0.40	0.30	0.35	–	0.42
E	0.65	0.60	0.90	0.58	–

The above table may be interpreted by assuming that the cell entry in the matrix represents the proportion of respondents who believe that ‘the column brand is preferred over the row brand’. For example:

In brand A versus brand B comparison it can be said that 60 per cent of the respondents prefer brand B to brand A. Similarly, 30 per cent of the respondents prefer brand C to brand A and so on.

To develop the ordinal scale from the given paired comparison data in the above table, we can convert the entries in the table to 0 – 1 scores. This is to show whether the column brand dominates the row brand and vice versa. If the proportion is greater than 0.5 in the above table, a number of ‘1’ is assigned to that cell, which means that the column brand is preferred over the row brand. Whenever the proportion is less than 0.5 in above table, a number of ‘0’ is assigned to that cell, which means column brand does not dominate the row brand. The results are in Table 12.3.

Table 12.3 Conversion of Paired Comparison Data into 0 to 1 form

	A	B	C	D	E
A	–	1	0	1	0
B	0	–	0	1	0
C	1	1	–	1	0
D	0	0	0	–	0
E	1	1	1	1	–
Total	2	3	1	4	0

NOTES

To get the ordinal relationship among the brands, we total the columns. Here the ordinal scale of brands is $D > B > A > C > E$. This means brand D is the most preferred brand, followed by B, A, C and E.

In order to obtain the interval scale data from the paired comparison data as presented above, the entries in the table can be analysed by using a technique called Thurston’s law of comparative judgement, which converts the ordinal judgements into the interval data. Here the proportions are assumed as probabilities and using the assumption of normality, Z-scores can be computed. Z-value has symmetric distribution with a mean of ‘0’ and variance of ‘1’. If the proportion is less than 0.5, the corresponding Z-value has a negative sign and for the proportion that is greater than 0.5, the Z-score takes a positive value. The Z-scores for the paired comparison data is given in Table 12.4.

Table 12.4 Z-score for Paired Comparison Data

	A	B	C	D	E
A	0	0.255	-0.525	0.255	-0.38
B	-0.255	0	-0.58	0.525	-0.255
C	0.525	0.58	0	0.385	-1.28
D	-0.255	-0.525	-0.385	0	-0.2
E	0.38	0.255	1.28	0.2	0
Total Distance	0.395	0.565	-0.21	1.365	-2.115
Average Distance	0.079	0.113	-0.042	0.273	-0.423
Brand	D	B	A	C	E
Interval scale value with change of origin	0.696	0.536	0.502	0.381	0

The entries in Table 12.4 show the distance between two brands. Assuming that the scores can be added, the total distance is computed. The average distance is computed by dividing the total score by the number of brands. This way one obtains the absolute position of each brand. Now the highest negative values among all the column is added to each entry corresponding to the average value so that by change of origin, interval scale values can be obtained. This is shown in the last row and the values are of interval scale, indicating the difference between brands. Brand D is the most preferred brand and E is the least preferred brand and the distance between the two is 0.696. The distance between brand C and E equals 0.381.

NOTES

Rank order scaling: In the rank order scaling, respondents are presented with several objects simultaneously and asked to order or rank them according to some criterion. Consider, for example the following question:

- Rank the following soft drinks in order of your preference, the most preferred soft drink should be ranked one, the second most preferred should be ranked two and so on.

Soft Drinks	Rank
Coke	
Pepsi	
Limca	
Sprite	
Mirinda	
Seven Up	
Fanta	

Like paired comparison, this approach is also comparative in nature. The problem with this scale is that if a respondent does not like any of the above-mentioned soft drink and is forced to rank them in the order of his choice, then, the soft drink which is ranked one should be treated as the least disliked soft drink and similarly, the other rankings can be interpreted. This scale is very commonly used to measure preferences for brands as well as attributes. The rank order scaling results in the ordinal data.

Constant sum rating scaling: In constant sum rating scale, the respondents are asked to allocate a total of 100 points between various objects and brands. The respondent distributes the points to the various objects in the order of his preference. Consider the following example:

- Allocate a total of 100 points among the various schools into which you would like to admit your child. The more the points you allocate to a school, more preferred it is to be considered. The points should be allocated in such a way that the sum total of the points allocated to various schools adds up to 100.

Schools	Points
DPS	
Modern School	
Mother's International	
APEEJAY	
DAV Public School	
Laxman Public School	
Tagore International	
TOTAL POINTS	100

Suppose Mother's International is awarded 30 points, whereas Laxman Public School is awarded 15 points, one can make a statement that the respondent rates Mother's International twice as high as Laxman Public School. This type of data is not only comparative in nature but could also result in ratio scale measurement. This type of scale is widely used in allocating weights which the consumer may assign to the various attributes of a product.

Q-sort technique: The Q-sort technique was developed to discriminate among a large number of objects quickly. This technique makes use of the rank order procedure in which objects are sorted into different piles based on their similarity with respect to certain criterion. Suppose there are 100 statements and an individual is asked to pile them into five groups, in such a way, that the strongly agreed statements could be put in one pile, agreed statements could be put in another pile, neutral statements form the third pile, disagreed statements come in the fourth pile and strongly disagreed statements form the fifth pile, and so on. The data generated in this way would be ordinal in nature. The distribution of the number of statement in each pile should be such that the resulting data may follow a normal distribution. The number of piles need not be restricted to 5. It could be as large as 10 or more as the large number increases the reliability or precision of the results.

Non-comparative Scales

In the non-comparative scales, the respondents do not make use of any frame of reference before answering the questions. The resulting data is generally assumed to be interval or ratio scale. For example:

The respondent may be asked to evaluate the quality of food in a restaurant on a five point scale (1 = very poor, 2 = poor and 5 = very good). The non-comparative scales are divided into two categories, namely, the graphic rating scales and the itemized rating scales. The itemized rating scales are further divided into Likert scale, semantic differential scale and Stapel scale. All these come under the category of the multiple item scales.

Graphic rating scale

This is a continuous scale, also called graphic rating Scale. In the graphic rating scale the respondent is asked to tick his preference on a graph. Consider for example the following question:

- Please put a tick mark (✓) on the following line to indicate your preference for fast food.



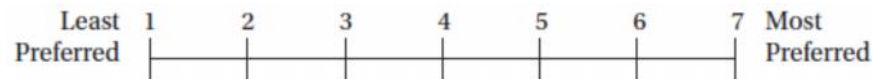
To measure the preference of an individual towards fast food one has to measure the distance from the extreme left to the position where a tick mark has been put. Higher the distance, higher would be the individual preference for fast food. This scale suffers from two limitations—one, if a respondent has put a tick

NOTES

NOTES

mark at a particular position and after ten minutes, he or she is given another form to put a tick mark, it will virtually be impossible to put a tick at the same position as was done earlier. Does it mean that the respondent's preference for fast food has undergone a change in 10 minutes? The basic assumption in this scale is that the respondents can distinguish the fine shade in differences between the preference/attitude which need not be the case. Further, the coding, editing and tabulation of data generated through such a procedure is a tedious task and researchers try to avoid using it. Another version of graphic scale could be the following:

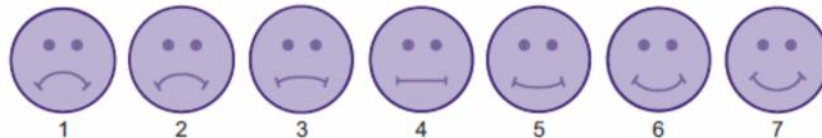
Ñ Please put a tick mark (•) on the following line to indicate your preference for fast food.



This is a slightly better version than the one discussed earlier. It will overcome the limitation of the scale to some extent. For example, if a respondent had earlier ticked between 5 and 6, it is likely that he would remember the same and the second time, he would tick very close to where he did earlier. This means that the difference in the two responses could be negligible.

Another way of presenting the graphic rating scale is through smiling face scale. The following example would illustrate the same.

Ñ Please indicate how much do you like fast food by pointing to the face that best shows your attitude and taste. If you do not prefer it at all, you would point to face one. In case you prefer it the most, you would point to face seven.



Itemized rating scale

In the itemized rating scale, the respondents are provided with a scale that has a number of brief descriptions associated with each of the response categories. The response categories are ordered in terms of the scale position and the respondents are supposed to select the specified category that describes in the best possible way an object is rated. Itemized rating scales are widely used in survey research. There are certain issues that should be kept in mind while designing the itemized rating scale. These issues are:

Number of categories to be used: There is no hard and fast rule as to how many categories should be used in an itemized rating scale. However, it is a practice to use five or six categories. Some researchers are of the opinion that more than five categories should be used in situations where small changes in attitudes are to be measured. There are others that argue that the respondents would find it difficult to distinguish between more than five categories. It is, however, a fact that the additional categories need not increase the precision with the attitude being measured. It is generally seen that researchers use five-category scales and in special cases, may increase or decrease the number of categories.

Odd or even number of categories: It has been a matter of debate among the researchers as to whether odd or even number of categories are to be used in survey research. By using even number of categories the scale would not have a neutral category and the respondent will be forced to choose either the positive or the negative side of the attitude. If odd numbers of categories are used, the respondent has the freedom to be neutral if he wants to be so. The Likert scale (to be discussed later) is a balanced rating scale with an odd number of categories and a neutral point. It is generally seen that if a respondent is not aware of the subject matter being measured by the scale, he would prefer to be neutral. However, if we have selected our unit of analysis to be one who is knowledgeable about the study being conducted and if he prefers to be neutral, we should not debar him from this opportunity.

Balanced versus unbalanced scales: A balanced scale is the one which has equal number of favourable and unfavourable categories. Examples of balanced and unbalanced scale are given below.

The following is the example of a balanced scale:

Ñ How important is price to you in buying a new car?

Very important

Relatively important

Neither important nor unimportant

Relatively unimportant

Very unimportant

In this question, there are five response categories, two of which emphasize the importance of price and two others that do not show its importance. The middle category is neutral.

The following is the example of the unbalanced scale.

Ñ How important is price to you in buying a new car?

More important than any other factor

NOTES

NOTES

Extremely important

Important

Somewhat important

Unimportant

In this question there are four response categories that are skewed towards the importance given to the price, whereas one category is for the unimportant side. Therefore, this question is an unbalanced question. In the unbalanced scale, the numbers of favourable and unfavourable categories are not the same. One could use an unbalanced scale depending upon the nature of attitude distribution to be measured. If the distribution is dominantly favourable, an unbalanced scale with more favourable categories than unfavourable categories should be appropriate. If an unbalanced scale is used, the nature and degree of the unbalance in the scale should be taken into account during the data analysis.

Nature and degree of verbal description: Many researchers believe that each category must have a verbal, numerical or pictorial description. Verbal description should be clearly and precisely worded so that the respondents are able to differentiate between them. Further, the researcher must decide whether to label every scale category, some scale categories, or only extreme scale categories. It is argued that a clearly defined response category increases the reliability of the measurement.

Forced versus non-forced scales: An important issue concerning the construction of an itemized rating scale is the use of a forced scale versus non-forced scale. In the forced scale, the respondent is forced to take a stand, whereas in the non-forced scale, the respondent can be neutral if he/she so desires. The argument for a forced scale is that those who are reluctant to reveal their attitude are encouraged to do so with the forced scale. Paired comparison scale, rank order scale and constant sum rating scales are examples of forced scales.

Physical form: There are many options that are available for the presentation of the scales. It could be presented vertically or horizontally. The categories could be expressed in boxes, discrete lines or as units on a continuum. They may or may not have numbers assigned to them. The numerical values, if used, may be positive, negative or both.

Suppose we want to measure the perception about Jet Airways using a multi-item scale. One of the questions is about the behaviour of the crew members. Given below is a set of scale configurations that may be used to measure their behaviour. The following are some of the examples where various forms of presenting the scales are shown:

The behaviour of the crew members of Jet Airways is:

1. Very bad _____ _____ _____ _____ _____ Very good
2. Very bad 1 2 3 4 5 Very good
3. Very bad

 Neither bad nor good

 Very good
4. Very bad Bad Neither bad nor good Good Very good
5. -2 -1 0 1 2
 Very bad Neither bad nor good Very good

NOTES

Below we will describe some of the itemized rating scales which are very commonly used in survey research.

Likert scale: This is a multiple item agree–disagree five-point scale. The respondents are given a certain number of items (statements) on which they are asked to express their degree of agreement/disagreement. This is also called a summated scale because the scores on individual items can be added together to produce a total score for the respondent. An assumption of the Likert scale is that each of the items (statements) measures some aspect of a single common factor, otherwise the scores on the items cannot legitimately be summed up. In a typical research study, there are generally 25 to 30 items on a Likert scale.

To construct a Likert scale to measure a particular construct, a large number of statements pertaining to the construct are listed. These statements could range from 80 to 120. The identification of the statements is done through exploratory research which is carried out by conducting a focus group, unstructured interviews with knowledgeable people, literature survey, analysis of case studies and so on. Suppose we want to assess the image of a company. As a first step, an exploratory research may be conducted by having an informal interview with the customers, and employees of the company. The general public may also be contacted. A survey of the literature on the subject may also give a set of information that could be useful for constructing the statements. Suppose the number of statements to measure the constructs is 100 in number. Now samples of representative respondents are asked to state their degree of agreement/disagreement on those statements. Table 12.5 gives a few statements to assess the image of the company.

It may be noted that only anchor labels and no numerical values are assigned to the response categories. Once the scale is administered, numerical values are assigned to the response categories. The scale contains statements’ some of which are favourable to the construct we are trying to measure and some are unfavourable to it.

NOTES

For example, out of the ten statements given, statements numbering 1, 2, 4, 6 and 9 in Table 12.5 are favourable statements, whereas the remaining are unfavourable statements. The reason for having a mixture of favourable and unfavourable statements in a Likert scale is that the responses by the respondent should not become monotonous while answering the questions. Generally, in a Likert scale, there is an approximately equal number of favourable and unfavourable statements. Once the scale is administered, numerical values are assigned to the responses. The rule is that a ‘strongly agree’ response for a favourable statement should get the same numerical value as the ‘strongly disagree’ response of the unfavourable statement. Suppose for a favourable statement the numbering is done as Strongly disagree = 1, Disagree = 2, Neither agree nor disagree = 3, Agree = 4 and Strongly agree = 5. Accordingly, an unfavourable statement would get the numerical values as Strongly disagree = 5, Disagree = 4, Neither agree nor disagree = 3, Agree = 2 and Strong agree = 1. In order to measure the image that the respondent has about the company, the scores are added.

Table 12.5 Likert Scale Statements to Measure the Image of the Company

No.	Statement	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
1.	The company makes quality products			✓		
2.	It is a leader in technology					✓
3.	It doesn't care about the general public		✓			
4.	The company leads in R&D to improve products				✓	
5.	The company is not a good paymaster	✓				
6.	The products of the company go through stringent quality tests				✓	
7.	The company has not done anything to curb pollution		✓			
8.	It does not care about the community near its plant	✓				
9.	The company's stocks are good to buy or own				✓	
10.	The company does not have good labour relations		✓			

For example, if a respondent has ticked (✓) statements numbering from one to ten as shown in Table 12.5, his total score would be 3 + 5 + 4 + 4 + 5 + 4 + 4 + 5 + 4 + 4 = 42 out of 50. Now if there are 100 respondents and 100 statements, the score on the image of the company can be worked out for each respondent by adding his/her scores on the 100 statements. The minimum score for each respondent will be 100, whereas the maximum score would be 500.

As mentioned earlier, a typical Likert scale comprises about 25–30 statements. In order to select 25 statements from the 100 statements, we need to discard some of them. The rule behind discarding the statements is that those

items that are non-discriminating should be removed. The procedure for choosing 25 (say number of statements) is shown.

As mentioned earlier, the score for each of the respondents on each of the statements can be used to measure his/her total score about the image of the company. The data may look as given in Table 12.6.

Table 12.6 shows that the total score for respondent no. 1 is 410, whereas for respondent no. 2 it is 209. This means that respondent no. 1 has a more favourable image for the company as compared to respondent no. 2. Now, in order to select 25 statements, let us consider statements numbering *i* and *j*. We note that the statement no. *j* is more discriminating as compared to statement no. *i*. This is because the score on statement *j* is very highly correlated with the total score as compared to the scores on statement *i*. Therefore, if we have to choose between *i* and *j*, we will choose statement no. *j*. From this we can conclude that only those statements will be selected which have a very high correlation with the total score. Therefore, the 100 correlations are to be arranged in the ascending order of magnitudes corresponding to each statement and only top 25 statements having a high correlation with the total score need to be selected.

Table 12.6 Total Score and Individual Score of Each Respondent on Various Statements

Scores of Statements										
Resp. No.	1	2	3	<i>i</i>	<i>j</i>	100	Total Score
1	-	-	-	5	4	-	410
2	-	-	-	4	2	-	209
3	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-
100	-	-	-	-	-	-	-

Another method of selecting the number of statements from a relatively large number of them is through the use of factor analysis. This aspect will be covered at the appropriate stage in the chapter on factor analysis.

Semantic differential scale: This scale is widely used to compare the images of competing brands, companies or services. Here the respondent is required to rate each attitude or object on a number of five- or seven-point rating scales. This scale is bounded at each end by bipolar adjectives or phrases. The difference between Likert and Semantic differential scale is that in Likert scale, a number of statements (items) are presented to the respondents to express their degree of agreement/disagreement. However, in the semantic differential scale, bipolar adjectives or phrases are used. As in the case of Likert scale, the information on the phrases and adjectives is obtained through exploratory research. At times there may be a

NOTES

NOTES











favourable or unfavourable descriptor (adjectives) on the right-hand side and on certain occasions these may be presented on the left-hand side. This rotation becomes necessary to avoid the halo effect. This is because the location of previous judgments on the scale may influence the subsequent judgements because of the carelessness of the respondents. The mid point of a bipolar scale is a neutral point. In the Likert scale, ten statements were used where respondents were asked to express their degree of agreement/disagreement regarding the image of the company. Taking the same example further, the semantic differential scale corresponding to those ten statements in Likert scale is shown below where the bipolar adjectives/phrases are separated by seven points. These points can be numbered as 1, 2, 3, ..., 7 or +3, +2, +1, 0, -1, -2, -3 for a favourable descriptor positioned on the left hand side. For an unfavourable descriptor the numberings would be reversed. A typical semantic differential scale where bipolar adjectives/phrases are positioned at the two extreme ends is given in Table 12.7.

Table 12.7 Select Bipolar Adjectives/Phrases of Semantic Differential Scale

1	Makes quality products	□ □ □ □ □ □ □	Does not make quality products
2	Leader in technology	□ □ □ □ □ □ □	Backward in technology
3	Does not care about general public	□ □ □ □ □ □ □	Cares about general public
4	Leads in R & D	□ □ □ □ □ □ □	Lagging behind in R&D
5	Not a good paymaster	□ □ □ □ □ □ □	A good paymaster
6	Products go through stringent quality test	□ □ □ □ □ □ □	Products don't go through quality test
7	Does nothing to curb pollution	□ □ □ □ □ □ □	Does a remarkable job in curbing pollution
8	Does not care about community near plants	□ □ □ □ □ □ □	Cares about community near plants
9	Company stocks good to buy	□ □ □ □ □ □ □	Not advisable to invest in company stock
10	Does not have good labour relations	□ □ □ □ □ □ □	Has good labour relations

Once the scale is constructed and administered to the representative respondents, the mean score for each of the descriptor is calculated. The scale is administered under the assumption that the numerical values assigned to the response categories are of interval scale in nature. This is generally the practice adopted by many researchers. However, if the response categories are treated as ordinal scale, instead of computing the arithmetic mean, median may be computed. In this example, we are treating the responses as the interval scale and hence the mean is computed. Once the mean for all the bipolar adjectives/phrases is computed we put the result in the form of a pictorial profile so as to make the comparison easy. At this time, all the favourable descriptors are kept on one side and all the unfavourable descriptors are positioned at the other. In our example, we have positioned all the favourable descriptors for the two companies whose image we want to compare on the left hand side. This is shown in Table 12.8.

Table 12.8 Pictorial Profile based on Semantic Differential Ratings

1	Makes quality products		Does not make quality products
2	Leader in technology		Backward in technology
3	Cares about general public		Does not care about general public
4	Leads in R & D		Lagging behind in R&D
5	A good paymaster		Not a good paymaster
6	Products go through stringent quality test		Products do not go through quality test
7	Done remarkable job in curbing pollution		Done nothing to curb pollution
8	Cares about community near plants		Does not care about community near plants
9	Company stocks good to buy		Not advisable to invest in company stock
10	Has good labour relations		Does not have good labour relations

_____ Company A _____ Company B

As per the results presented in the pictorial profile, Company A is better than Company B in the sense that it makes quality products, leads in R&D, its products go through stringent quality tests, its stocks are good to buy and it has good labour relations. Company B is ahead of Company A as it cares about general public and is a good paymaster. Company A is better than Company B as it is leads in technology whereas Company B is better than Company A as it has done a remarkable job in curbing pollution. However, these differences are not statistically significant.

Stapel scale: The Stapel scale is used to measure the direction and intensity of an attitude. At times, it may be difficult to use semantic differential scales because of the problem in creating bipolar adjectives.

RESTAURANT	
+5	+5
+4	+4
+3	+3
+2*	+2
+1	+1
Quality of Food	Quality of Service
-1	-1
-2	-2
-3	-3
-4	-4
-5	-5*

NOTES

NOTES

The Stapel scale overcomes this problem by using only single adjectives. This scale generally has 10 categories involving numbering –5 to +5 without a neutral point and is usually presented in a vertical form. The job of the respondent is to indicate how accurately or inaccurately each term describes the object by selecting an appropriate numerical response category. If a positive higher number is selected by the respondent, it means the respondent is able to describe it more favourably. Suppose a restaurant is to be evaluated on quality of food and quality of service, then the Stapel scale would be presented as shown on the previous page:

In the above scale, the respondents are asked to evaluate how accurately each word or phrase describes the restaurant in question. They will choose a value of +5 if the restaurant very accurately describes the attribute and –5 if it does not describe at all correctly the word in question. Suppose a respondent has chosen his options as indicated by *. This shows that the respondent slightly prefers the quality of food and is of the opinion that the quality of service is totally useless.

Measurement Error

Measurement error occurs when the observed measurement on a construct or concept deviates from its true values. The following is a list of the sources of measurement errors.

- Ñ There are factors like mood, fatigue and health of the respondent which may influence the observed response while the instrument is being administered.
- Ñ The variations in the environment in which measurements are taken may also result in a departure from the true value.
- Ñ There are situations when a respondent may not understand the question being asked and the interviewer may have to rephrase the same. While rephrasing the question the interviewer's bias may get into the responses. Also how the questionnaire is administered (telephone survey, personal interview with questionnaire or mail survey) will have its own impact on the responses.
- Ñ At times, some of the questions in the questionnaire may be ambiguous and some may be very difficult for the respondents to understand. Both of them can cause deviation from the correct response, thereby giving rise to measurement error.
- Ñ At times, the errors may be committed at the time of coding, entering of data from questionnaire to the spreadsheet on the computer and at the tabulation stage.

The observed measurement in any research need not be equal to the true measurement. The observed measurement can be written as

$$O = T + S + R$$

where, O = Observed measurement
T = True score
S = Systematic error
R = Random error

It may be noted that the total error consists of two components—systematic error and random error. Systematic error causes a constant bias in the measurement. Suppose there is a weighing scale that weighs 50 gm less for every one kg of product being weighed. The error would consistently remain the same irrespective of the kind of product and the time at which product is weighed. Random error on the other hand involves influences that bias the measurements but are not systematic. Suppose we use different weighing scales to weigh one kg of a product and if systematic error is assumed to be absent, we may find that recorded weights may fall within a range around the true value of the weight, thereby causing random error.

12.3.1 Criteria for Good Measurement: Reliability and Validity

There are three criteria for evaluating measurements: reliability, validity and sensitivity.

Reliability

Reliability is concerned with consistency, accuracy and predictability of the scale. It refers to the extent to which a measurement process is free from random errors. The reliability of a scale can be measured using the following methods:

Test–retest reliability: In this method, repeated measurements of the same person or group using the same scale under similar conditions are taken. A very high correlation between the two scores indicates that the scale is reliable. However, the following issues should be kept in mind before arriving at such a conclusion.

- Ñ What should be the appropriate time difference between the two observations is a question which requires attention. If the time difference between two consecutive observations is very small (say two or three weeks) it is very likely that the respondents would remember the previous answer and may give the same answer when the instrument is administered the second time. This will make the instrument reliable, which may not actually be the case. However, if the difference between the two observations is very large (say more than a year) it is quite likely that the respondent's answers to the various questions of the instrument might have actually undergone a change, resulting in poor reliability of the scale. Therefore, the researcher has to be very careful in deciding upon the time difference between the two observations. Generally, it is thought that a time difference of about five to six months is an ideal period.
- Ñ Another problem in this test is that the first measurement may change the response of the subject to the second measurement.

NOTES

NOTES

- Ñ The situational factors working on two different time periods may not be the same, which may result in different measurement in the two periods.
- Ñ The second reading on the same instrument from the same subject may produce boredom, anger or attempt to remember the answers given in an initial measurement.
- Ñ A favourable response with a brand during the period between the two tests might cause a shift in the individual rating by the subject.

Split-half reliability method: This method is used in the case of multiple item scales. Here the number of items is randomly divided into two parts and a correlation coefficient between the two is obtained. A high correlation indicates that the internal consistency of the construct leads to greater reliability. Another measure which is used to test the internal consistency of a multiple item scale is the coefficient alpha (\pm) commonly known as cronbach alpha. The cronbach alpha computes the average of all possible split-half reliabilities for a multiple item scale. This coefficient demonstrates whether the average score of all split-half of reliabilities converge to a certain point or not.

The coefficient alpha does not address validity. However, many researchers use this as a sole indicator of validity. The alpha coefficient can take values between 0 and 1. The following values of alpha with their interpretations are suggested below:

$\alpha = 0$ means	There is no consistency between the various items of a multiple item scale
$\alpha = 1$ means	There is complete consistency between various items of a multiple item scale
$0.80 \leq \alpha \leq 0.95$ implies	There is very good reliability between the various items of a multiple item scale
$0.70 \leq \alpha \leq 0.80$ implies	There is good reliability between the various items of a multiple item scale
$0.60 \leq \alpha \leq 0.70$ implies	There is fair reliability between the various items of a multiple item scale
$\alpha < 0.60$ means	There is poor reliability between the various items of a multiple item scale

Validity

The validity of a scale refers to the question whether we are measuring what we want to measure. Validity of the scale refers to the extent to which the measurement process is free from both systematic and random errors. The validity of a scale is a more serious issue than reliability. There are different ways to measure validity.

Content validity: This is also called face validity. It involves subjective judgement by an expert for assessing the appropriateness of the construct. For example, to measure the perception of a customer towards Jet Airways, a multiple item scale is developed. A set of 15 items is proposed. These items when combined in an index measure the perception of Jet Airways. In order to judge the content validity

of these 15 items, a set of experts may be requested to examine the representativeness of the 15 items. The items covered may be lacking in the content validity if we have omitted behaviour of the crew, food quality, and food quantity, etc., from the list. In fact, conducting the exploratory research to exhaust the list of items measuring perception of the airline would be of immense help in such a case.

Concurrent validity: It is used to measure the validity of the new measuring techniques by correlating them with the established techniques. It involves computing the correlation coefficient of two measures of the same phenomena (for example, perception of an airline and image of a company) which are administered at the same time. We prepare a 15 item scale to measure the perception of Jet Airways, which is assumed to be a valid one. Suppose a researcher proposes an alternative and shorter technique. The concurrent validity of the new technique would be established if there is a high correlation between the two techniques when administered at the same time under similar or identical conditions.

Predictive validity: This involves the ability of a measured phenomena at one point of time to predict another phenomenon at a future point of time. If the correlation coefficient between the two is high, the initial measure is said to have a high predictive ability. As an example, consider the use of the common admission test (CAT) to shortlist candidates for admission to the MBA programme in a business school. The CAT scores are supposed to predict the candidate's aptitude for studies towards business education.

Sensitivity

The sensitivity of a scale is an important measurement concept, particularly when changes in attitudes are under investigation. Sensitivity refers to an instrument's ability to accurately measure the variability in a concept. A dichotomous response category such as agree or disagree does not allow the recording of any attitude changes. A more sensitive measure with numerous categories on the scale may be required. For example, adding strongly agree, agree, neither agree nor disagree, disagree and strongly disagree categories will increase the sensitivity of the scale.

The sensitivity of scale based on a single question or a single item can be increased by adding questions or items. In other words, because composite measures allow for a greater range of possible scores, they are more sensitive than a single-item scale. Therefore, the sensitivity of the scale is generally increased by adding more response points or by adding scale items.

12.3.2 Thurstone, Guttman and Bogardus Scales

Let us analyse some more scales of relevance.

Thurstone Scale

Thurstone's method of pair comparisons can be considered a prototype of a normal distribution-based method for scaling-dominance matrices. Even though the theory behind this method is quite complex, the algorithm is definitely straightforward.

NOTES

NOTES

A Thurstone scale has a number of “agree” or “disagree” statements. It is a uni-dimensional scale to measure attitudes towards people. Developing the scale is time consuming and relatively complex compared to other scales (like the Likert scale). The steps are as below:

Ñ Although there are technically three scales, when people refer to the “Thurstone Scale” they are usually talking about the method of equal-appearing intervals. It is called “Equal appearing intervals” because when you choose the items for your test, you are picking items equally spaced apart.

The other two variations are:

Ñ The method of **successive intervals**: this method is more challenging to implement than equal-appearing intervals.

Ñ The method of **paired comparisons**: requires twice the judgments than the equal-appearing intervals method and can quickly become very consuming.

Guttman Scale

A Guttman Scale (named after Louis Guttman) is formed by a set of items if they can be ordered in a reproducible hierarchy. For example, in a test of achievement in mathematics, if examinees can successfully answer items at one level of difficulty (e.g., summing two 3-digit numbers), they would be able to answer the earlier questions (e.g., summing two 2-digit numbers).

The Guttman scale only applies to tests, such as achievement tests, that have binary outcomes and it is assumed that respondents can only and will always respond in the same way. A perfect Guttman scale consists of a unidimensional set of items that are ranked in order of difficulty from least difficult to most difficult. Therefore, a person scoring a “7” on a ten item Guttman scale, will agree with items 1-7 and disagree with items 8,9,10. This means that a person’s entire set of responses to all items can be predicted from their cumulative score because the model is deterministic. The extent to which a scale is reproducible can be estimated from the coefficient of reproducibility and, to ensure that there is a range of responses (not the case if all respondents only endorsed one item) the coefficient of scalability is used. In the creation of Guttman scales items that reduce reproducibility of scalability are re-written or discarded. In Guttman scaling is found the beginnings of item response theory which, in contrast to classical test theory, acknowledges that items in questionnaires do not all have the same level of difficulty. Non-deterministic (ie stochastic) models have been developed such as the Mokken scale and the Rasch model.

Bogardus Scale

The Bogardus scale is a social distance scale that measures prejudice—or, more precisely, the degrees of warmth, intimacy, indifference or hostility—between an individual and any social, racial or ethnic groups.

Developed by Emory Bogardus in 1924 and named after him, the Bogardus social distance scale is one of the oldest psychological attitude scales still in use. It is unidimensional, which means it can be used to measure exactly one concept (prejudice). But though it was created to measure prejudice against other racial groups, it is broad enough it can be used in reference to almost any societal group (homeless people, artists, atheists, circus dancers, ISIS members, etc.).

Check Your Progress

4. What is rank order scaling?
5. What are non-comparative scales?

NOTES

12.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. There are four types of measurement scales—nominal, ordinal, interval and ratio scales.
2. Nominal scale measurements are used for identifying food habits (vegetarian or non-vegetarian), gender (male/female), caste, respondents, brands, attributes, stores, the players of a hockey team and so on.
3. Ratio scale is the highest level of measurement and takes care of the limitations of the interval scale measurement, where the ratio of the measurements on the scale does not have a meaningful interpretation.
4. In the rank order scaling, respondents are presented with several objects simultaneously and asked to order or rank them according to some criterion.
5. In the non-comparative scales, the respondents do not make use of any frame of reference before answering the questions. The resulting data is generally assumed to be interval or ratio scale.

12.5 SUMMARY

- Ñ The term ‘measurement’ means assigning numbers or some other symbols to the characteristics of certain objects. When numbers are used, the researcher must have a rule for assigning a number to an observation in a way that provides an accurate description.

NOTES

- Ñ Scaling is an extension of measurement. Scaling involves creating a continuum on which measurements on objects are located.
- Ñ There are four types of measurement scales—nominal, ordinal, interval and ratio scales.
- Ñ Nominal scale is the lowest level of measurement. Here, numbers are assigned for the purpose of identification of the objects. Any object which is assigned a higher number is in no way superior to the one which is assigned a lower number.
- Ñ Ordinal scale is the next higher level of measurement than the nominal scale measurement. One of the limitations of the nominal scale measurements is that we cannot say whether the assigned number to an object is higher or lower than the one assigned to another option.
- Ñ The interval scale measurement is the next higher level of measurement. It takes care of the limitation of the ordinal scale measurement where the difference between the score on the ordinal scale does not have any meaningful interpretation.
- Ñ Ratio scale is the highest level of measurement and takes care of the limitations of the interval scale measurement, where the ratio of the measurements on the scale does not have a meaningful interpretation.
- Ñ In multiple item scale, there are many items that play a role in forming the underlying construct that the researcher is trying to measure.
- Ñ In comparative scales it is assumed that respondents make use of a standard frame of reference before answering the question.
- Ñ In the rank order scaling, respondents are presented with several objects simultaneously and asked to order or rank them according to some criterion.
- Ñ Likert scale is a multiple item agree–disagree five-point scale. The respondents are given a certain number of items (statements) on which they are asked to express their degree of agreement/disagreement.
- Ñ A Guttman Scale (named after Louis Guttman) is formed by a set of items if they can be ordered in a reproducible hierarchy.
- Ñ Thurstone’s method of pair comparisons can be considered a prototype of a normal distribution-based method for scaling–dominance matrices. Even though the theory behind this method is quite complex, the algorithm is definitely straightforward.

12.6 KEY WORDS

- Ñ **Likert scale:** This is a multiple item agree–disagree five-point scale. The respondents are given a certain number of items (statements) on which they are asked to express their degree of agreement/disagreement.

Ñ **Attitude:** An attitude is viewed as an enduring disposition to respond consistently in a given manner to various aspects of the world, including persons, events and objects.

Ñ **Q-sort technique:** It was developed to discriminate among a large number of objects quickly. This technique makes use of the rank order procedure in which objects are sorted into different piles based on their similarity with respect to certain criterion.

NOTES

12.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Differentiate between nominal and ordinal scales.
2. What are comparative scales?
3. State the issues of itemized rating scale.

Long-Answer Questions

1. Discuss the various techniques of scaling.
2. Describe the different methods of scale construction.
3. Discuss the criteria for evaluating measurements.

12.8 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

NOTES

UNIT 13 MEASURES OF CENTRAL TENDENCY, DISPERSION AND CORRELATION ANALYSIS

Structure

- 13.0 Introduction
- 13.1 Objectives
- 13.2 Mean, Median Mode
- 13.3 Range, Quartile, Mean and Standard Deviation
- 13.4 Karl Pearson's Coefficient of Correlation
- 13.5 Rank Correlation and Attributes
- 13.6 Solved Problems
- 13.7 Answers to Check Your Progress Questions
- 13.8 Summary
- 13.9 Key Words
- 13.10 Self Assessment Questions and Exercises
- 13.11 Further Readings

13.0 INTRODUCTION

Measures of central tendency are used to represent “central value” or “central location” that the data seem to be grouped around. There are different measures of central tendency. A measure of central tendency may be used to divide a group of observations into two equal groups.

Dispersion is a statistical term that describes the size of the range of values expected for a particular variable. Correlation is a statistical technique that can show whether and how strongly pairs of variables are related.

13.1 OBJECTIVES

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

- Analyse the properties of arithmetic mean
- Describe the computation of median with the help of examples
- Discuss the process of the computation of range and quartile deviation
- Explain the measure of Karl Pearson's coefficient of correlation

13.2 MEAN, MEDIAN MODE

Popularly known as *average*, arithmetic mean (or simply, mean) is the most important and frequently used measure of central tendency. It possesses certain useful properties, which account for its wider use. Mean is obtained by taking the sum of all observations comprising a given set of data and dividing the sum by the total number of observations. This method remains essentially the same whether the data refer to a sample or a finite population.

Consider, *for example*, the ungrouped raw data given in Table 13.1. Comprising 50 sample observations in cm, their sum comes to 8356.3. When divided by 50, the resultant mean height comes to 167.26 cm.

Table 13.1 Heights of a Sample of 50 Students of a Management Programme

166.7	167.2	167.0	166.5	166.5	167.4	168.2	167.0	167.6
162.2	164.8	163.7	166.9	165.4	169.3	166.4	167.4	168.0
170.6	170.2	170.0	161.3	167.1	170.5	162.8	164.7	
165.8	165.2	169.6	169.0	167.6	165.8	170.3	170.8	
166.8	166.4	165.2	165.9	171.7	163.8	164.5	170.2	
170.9	167.9	166.8	170.8	168.1	169.2	163.2	163.9	

Representing these sample observations by x_1, x_2, \dots, x_{50} , the sample mean, denoted as \bar{x} , is computed as

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_{50}}{50}.$$

Using the sigma operation,

$$\bar{x} = \frac{\sum_{i=1}^{50} x_i}{50} = \frac{8356.3}{50} = 167.126 \text{ cm.}$$

In general, the mean of n sample observations is

$$\bar{x} = \frac{\sum_{i=1}^{50} x_i}{n} \quad (13.1)$$

Analogously, the mean of a finite population, denoted as μ , is

$$\mu = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N}, \quad (13.2)$$

in which x_1, x_2, \dots, x_N represent the N observations comprising the population.

NOTES

Illustration: Consider the weekly earnings of a sample of 80 female weavers in Table 13.1. The sum of $n = 80$ observations is $\sum x_i = 86665$. Using Eq. (13.1), the mean earnings are

$$\bar{x} = \frac{\sum x_i}{n} = \frac{86665}{80} = 1083.312$$

For the frequency distribution of height data as given in Table 13.2, the necessary computations for obtaining mean are made therein. The computational method may be notationally expressed as:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \dots + f_kx_k}{f_1 + f_2 + \dots + f_k} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} \quad (13.3)$$

Herein, x_i 's are the class mid-points, f_i 's the corresponding class frequencies, and k the number of classes such that $i = 1, 2, 3, \dots, k$.

Table 13.2 Frequency Distribution of Heights of 50 Students of a Management Programme

Class Intervals	Class Frequencies	Class Mid-Points	$f_i x_i$
$L_1 - L_2$	f_i	$x_i = (L_1 + L_2)/2$	(2 × 3)
(1)	(2)	(3)	(4)
161–162.9	3	161.95	485.85
163–164.9	7	163.95	1147.65
165–166.9	14	165.95	2323.30
167–168.9	12	167.95	2015.40
169–170.9	10	169.95	1699.50
171–172.9	4	171.95	687.80
$n = \sum f_i = 50$			$\sum f_i x_i = 8359.50$

The various steps for using Eq. (13.3) consist of

- obtaining x_i for each class as $[(L_1 + L_2)/2]$,
- multiplying each x_i by corresponding f_i to get $f_i x_i$,
- taking the sum of $f_i x_i$ as $\sum f_i x_i$ and
- dividing the sum $\sum f_i x_i$ by $\sum f_i = n$.

Since $\sum f_i x_i = 8359.50$ and $\sum f_i = 50$,

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{8359.50}{50} = 167.19 \text{ cm.}$$

For a frequency distribution based on N population observations, the corresponding equation for the computation of population mean μ is

$$\mu = \frac{f_1 x_1 + f_2 x_2 + \dots + f_k x_k}{f_1 + f_2 + \dots + f_k} = \frac{\sum_{i=1}^k f_i x_i}{\sum_{i=1}^k f_i} \quad (13.4)$$

Whereas x_i 's in Eq. (13.4) represent the class mid-points of the population distribution, here $\sum f_i = N$ as against $\sum f_i = n$ in Eq. (13.3). Only to remind, N stands for the total number of observations in the population and n represents those in the sample.

Correction for Wrong Entries

The method of computation of mean readily enables us to correct the value of mean for any wrong entries that may have occurred during the course of summation. *For example*, on obtaining the mean of a set of n sample observations, it may be discovered that one or two observations were misread and wrong entries made in arriving at the sum $\sum x_i$.

In a situation of this type, computation of mean by using Eq. (13.1) or Eq. (13.2) easily permits making correction for the wrong entries. The wrong entries and the corresponding correct observations must, however, be known to us.

If a given wrong entry is denoted as x_i^0 and the corresponding correct observation as x_i , the correct mean \bar{x}_c may be obtained as

$$\bar{x}_c = \frac{n\bar{x} - x_i^0 + x_i}{n}$$

for a single wrong entry. For two wrong entries, the correct mean

$$\bar{x}_c = \frac{n\bar{x} - (x_1^0 + x_2^0) + (x_1 + x_2)}{n}$$

in which x_1^0 and x_2^0 are the two wrong entries made against the corresponding correct observations x_1 and x_2 , respectively.

It may be noted that for a larger number of wrong entries made, it is worthwhile to compute \bar{x} afresh. For, going into a correction process as explained above will make the task more difficult and cumbersome.

Illustration: Let us illustrate by taking a hypothetical case. Suppose the mean \bar{x} computed for $n = 50$ sample observations is 40. Further assume that due to an oversight, an actual observation of 53 is wrongly read as 83 and $\bar{x} = 40$ computed with 83 entered in place of 53.

Thus, for $x_i^0 = 83$ and $x_i = 53$, the corrected mean \bar{x}_c can be obtained as

$$\bar{x}_c = \frac{n\bar{x} - x_i^0 + x_i}{n} = \frac{(50 \times 40) - 83 + 53}{50} = 39.4.$$

NOTES

NOTES

Properties of Arithmetic Mean

An important characteristic of mean is that it satisfies some useful properties. No other measure of central tendency except mean does this. This holds both for the ungrouped and the grouped data, and irrespective of whether the data refer to a sample or a population. Stating them here only with reference to sample ungrouped data, these properties are as follows:

- (i) That the sum of the deviations of a set of n observations (x_1, x_2, \dots, x_n) from their mean \bar{x} is equal to zero. Thus, $(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x})$ being the deviations, it requires that

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = \sum_{i=1}^n (x_i - \bar{x}) = 0.$$

For example, at an elementary level, when mean for $n = 5$ observations (4, 6, 3, 5, 2) is = 4. The deviations $(x_i - \bar{x})$ being (4 - 4), (6 - 4), (3 - 4), (5 - 4), (2 - 4), their sum $\sum(x_i - 4) = 0$.

A simple algebraic operation also shows it as

$$\begin{aligned} \sum(x_i - \bar{x}) &= \sum x_i - \sum \bar{x} \\ &= \sum x_i - n\bar{x} \\ &= n\bar{x} - n\bar{x} \\ &= 0 \end{aligned}$$

- (ii) That the sum of the squares of deviations of a set of n observations (x_1, x_2, \dots, x_n) from any number, say A , is the least only when $A = \bar{x}$. That is, $\sum(x_i - A)^2$, which represents the sum of the squares of the deviations of x_i from any number A , is the least when $A = \bar{x}$.

This property is extremely important as it is used for developing measures of dispersion discussed in the next chapter. Its genesis lies in the fact that $\sum(x_i - \bar{x}) = 0$. Given that A is not \bar{x} , $\sum(x_i - A)$ will be anything but zero. Obviously, $\sum(x_i - A)^2$ will be the least only when $A = \bar{x}$, because in that case $\sum(x_i - A) = \sum(x_i - \bar{x}) = 0$.

Alternatively, $\sum(x_i - A)^2$ will be minimum only when its first derivative with respect to A is zero. That is, it will be minimum when

$$-2\sum(x_i - A) = 0$$

or $\sum(x_i - A) = 0$

or $\sum x_i - nA = 0$

or $nA = \sum x_i$

or $A = \sum x_i / n = \bar{x}$

- (iii) That if there are k sets of sample data consisting of n_1, n_2, \dots, n_k observations with $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$ as their respective means, the combined mean \bar{x}_c of all the $(n_1 + n_2 + \dots + n_k)$ observations in k sets of data is

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \dots + n_k\bar{x}_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k n_i x_i}{\sum_{i=1}^k n_i} \quad (13.5)$$

Illustration: Suppose a sample of 25 girl students of a primary school shows an average weight of 42 kg. Assume further that another sample of 15 boys of the same school gives an average weight of 46 kg. Given these data, we may find the average weight of all the 40 students, 25 girls and 15 boys, by pooling the data for the two samples. Thus, if $\bar{x}_1 = 42$ kg for $n_1 = 25$ students and \bar{x}_2 for $n_2 = 15$ students, the combined mean \bar{x}_c for $k = 2$ may be obtained by using Eq. (13.5). That is,

$$\bar{x}_c = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{(25 \times 42) + (15 \times 46)}{(25 + 15)} = 43.5.$$

- (iv) If y represents a linear transformation on x (that is, $C = a + bx$), then the mean of y is given by the same transformation as on mean \bar{x} . That is, when

$$y = a + bx$$

$$\bar{y} = a + b\bar{x}$$

wherein a is the additive constant and b is the multiplicative constant.

To show this, we have

$$\begin{aligned} \bar{y} &= \frac{\sum y}{n} = \frac{\sum (a + bx)}{n} \\ &= \frac{\sum a}{n} + \frac{b \sum x}{n} = \frac{na}{n} + b\bar{x} \\ &= a + b\bar{x} \end{aligned}$$

- (v) If A is any value from within or beyond a given set of data consisting of n observations (E_1, E_2, \dots, x_n) , and if d_i 's represent the deviations of each x_i from A [that is, $d_i = (x_i - A)$], then

$$\bar{x} = A + \bar{d} = A + \frac{\sum d_i}{n}. \quad (13.6)$$

Since Eq. (13.6) also applies to grouped data, in that case

$$\bar{x} = A + \bar{d} = A + \frac{\sum f_i d_i}{\sum f_i}.$$

NOTES

NOTES

To show it, we have

$$d_i = x_i - A$$

or $x_i = A + d_i$

or $\sum x_i = An + \sum d_i$

or $\frac{\sum x_i}{n} = \frac{An}{n} + \frac{\sum d_i}{n}$

or $\bar{x} = A + \bar{d}$

A distinct advantage of using Eq. (13.6) for grouped data in place of Eq. (13.3) is that it simplifies the computational work, especially when A is one of the midpoints of the various class intervals. This is possible because Eq. (13.6) involves change of origin, which reduces each x_i by the constant A . It amounts to shifting the base from 0 to A .

A still easier way to compute mean is to define u_i , as

$$u_i = \frac{d_i}{C} \quad \text{or} \quad u_i = \frac{(x_i - A)}{C}$$

Here, not only has the origin been changed from x_i to d_i , as $d_i = (x_i - A)$, even the scale has been reduced by dividing d_i 's by C .

As C represents the width of the class interval, reducing the deviations by it requires that C remains the same for all the classes. Thus, when u_i 's are defined as

$$u_i = \frac{(x_i - A)}{C}$$

or

$$x_i = A + Cu_i$$

Then, by virtue of (iv) above, mean for grouped data with constant C , is

$$\begin{aligned} \bar{x} &= A + C\bar{u} \\ &= A + C \left(\frac{\sum f_i u_i}{\sum f_i} \right) \end{aligned} \quad (13.7)$$

It is obtained as

$$u_i = (x_i - A)/C$$

or $x_i = A + Cu_i$

or $\sum f_i x_i = A \sum f_i + C \sum f_i u_i$

$$\text{or } \frac{\sum f_i x_i}{\sum f_i} = A + C \frac{\sum f_i u_i}{\sum f_i}$$

$$\bar{x} = A + C\bar{u}$$

Illustration: Take once again the frequency distribution given in Table 13.3 where the width of class interval $C = 10$ remains the same for all the classes. Carefully observe the computations made in Table 13.4 to identify the various steps taken in sequence.

Table 13.4 Computation of using Eqs. (13.6) and (13.7)

Class Interval	Frequencies	Mid-Points	Deviations	$f_i d_i$	$u_i = d_i/C$	$f_i u_i$
$L_1 - L_2$	f_i	x_i	$d_i - x_i - A$	(2×4)	$-(x_i - A)/C$	(2×6)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1050-59	6	1054.5	-30	-180	-3	-18
1060-69	9	1064.5	20	180	2	18
1070-79	15	1074.5	-10	-150	-1	-15
1080-89	25	1084.5 - A	0	0	0	0
1090-99	13	1094.5	10	130	1	13
1100-09	7	1104.5	20	140	2	14
1110-19	5	1114.5	30	150	3	15
$n = \sum f_i = 80$				$\sum f_i d_i = -90$		$\sum f_i u_i = -9$

Substituting the values in Eqs. (13.6) and (13.7), we have:

Using Eq. (13.6)

$$\bar{x} = A + \bar{d}$$

$$\text{or } \bar{x} = A + \frac{\sum f_i d_i}{\sum f_i}$$

$$= 1084.5 + \frac{-90}{80}$$

$$= 1083.375.$$

Using Eq. (13.7)

$$\bar{x} = A + C\bar{u}$$

$$\text{or } \bar{x} = A + C \left(\frac{\sum f_i u_i}{\sum f_i} \right)$$

$$= 1084.5 + 10 \frac{-9}{80}$$

$$= 1083.375.$$

It may be noted that the value of \bar{x} computed either way is the same as obtained by using Eq. (13.3). Also note that Eqs. (13.6) and (13.7) provide *more efficient method(s)* of computing mean \bar{x} . This is evident from the step-wise computations made in Table 13.4, compared to those made in Table 13.3. When done manually, Cols. (4) and (5) in the case of Eq. (13.6) and Cols. (6) and (7) in the case of Eq. (13.7) involve simpler computations than those in Col. (4) in the case of Eq. (13.3).

Do yourself: For the distribution in Table 13.2, obtain \bar{x} by using Eqs. (13.6) and (13.7) strictly in the manner as in Table 13.4. Compare the mean obtained by using both the equations to see that it is the same as computed by using Eq. (13.3). Also check up that if A is assumed to be 165, $\sum f_i d_i$ and $\sum f_i u_i$ are 109.50 and 54.750, respectively.

NOTES

NOTES

In the case of a distribution containing open-ended class(es), the usual methods of obtaining mean as contained in Eqs. (13.3), (13.6), and (13.7) do not apply. For the obvious reason that mid-points of the open-ended class(es) cannot be found. Thus, mean is indeterminate in the case of such distributions. However, if absolutely necessary, some idea of mean can be had by computing it without giving any consideration to the open-ended class(es).

Importantly, it may be noted that mean serves as the best representative of central tendency only when a distribution is bell-shaped or roughly bell-shaped. For, these distributions tend to have the majority of observations concentrated somewhere around the middle. Further, mean serves as a reliable measure of comparing two or more frequency distributions when i) they possess approximately identical shape, and ii) the data are expressed in the same units of measurement.

More Efficient Method(s) Restated

It should now be clear that Eq. (13.3) works as a convenient method of obtaining the mean for the grouped data when the mid-points and/or the corresponding class frequencies are relatively small in magnitude. The computation of mean by using Eq. (13.3) becomes cumbersome when it is not so. In other words, the use of Eqs. (13.6) and (13.7) is justified for easy computations where the class-mid points and/or corresponding frequencies are large in magnitude.

Equation (13.6) provides a more friendly method of obtaining mean for grouped data when the width of class interval C varies and the magnitudes involved are large. Equation (13.7) simplifies computation of mean where the width of the class interval C remains the same over all the classes and the data magnitudes are large.

As stated earlier, the more efficient methods contained in Eqs. (13.6) and (13.7) are based on redefining the class mid-points x_i 's as

$$d_i = (x_i - A)$$

in the case of unequal class intervals, and as

$$u_i = \frac{d_i}{C} \quad \text{or} \quad u_i = \frac{(x_i - A)}{C}$$

in the case of equal class intervals. The d_i 's represent the deviations of class midpoints x_i from any arbitrary value A , usually referred to as *assumed mean*.

For a clear understanding of using the more efficient method, the application of Eq. (13.6) is demonstrated again with reference to the distribution given in Table 13.5. With all the necessary computations made therein, substituting the values in Eq. (13.6) gives us the mean value as under:

$$\bar{x} = A + \frac{\sum f_i d_i}{\sum f_i} = 25 + \left(\frac{475}{100} \right) = 29.75.$$

Table 13.5 Computation of using Eq. (13.6)

Sales in ₹Lac $L_1 - L_2$	No of Units f_i	Class Mid-Points x_i	$d_i = (x_i - A)$	$f_i d_i (2 \times 4)$
(1)	(2)	(3)	(4)	(5)
00-10	05	05.0	-20.0	-100.0
10-20	10	15.0	-10.0	-100.0
20-30	25	25.0 - A	0	0
30-35	30	32.5	7.5	225.0
35-40	20	37.5	12.5	250.0
40-50	10	45.0	20.0	200.0
$n = \Sigma f_i = 100$			$\Sigma f_i d_i = 475.0$	

NOTES

Caution: It may be noted that the width of the class interval does not remain the same over all the classes. This calls for the use of the method in Eq. (13.6), being more efficient, instead of that in Eq. (13.7).

Since the width of the class interval C is not the same for all the classes, the steps involved in using the more efficient method in Eq. (13.6) are as follows:

- Obtain the various class mid-points x_i [as $(L_1 + L_2)/2$.
- Designate any x_i , preferably the mid-point of the middle class, or the class with the highest frequency, as assumed mean A .
- Obtain deviations d_i as $d_i = (x_i - A)$ [and u_i as $u_i = (x_i - A)/C$ when using Eq. (13.7)].
- Multiply each deviation d_i by the corresponding class frequency f_i to obtain $f_i d_i$ [and $f_i u_i$ when using Eq.(13.7)].
- Substitute the needed values in Eq. (13.6) [or in Eq. (13.7), provided C is the same over all the classes] to obtain the value of \bar{x} .

Weighted Arithmetic Mean/Average

There are situations when all the n individual observations in a set of data are not equally important. This does not allow arithmetic mean to serve as the best representative of the given data. Where individual observations vary in importance, they are assigned weights according to the level of importance of each in the computation of their mean.

For example, the compensation package of those engaged in different managerial cadre positions in a company may vary according to one's technical and professional qualifications. Depending on the work requirements one has to handle while working in any position, the company may have only a few working in senior managerial positions and getting very high compensation package. It may have relatively more of those working in supervisory positions and drawing relatively much lower compensation package.

If we were to obtain the average compensation package for all managerial positions, a simple average of all possible operative compensation packages will

NOTES

give a very deceptive picture. The best option in the case is to accord due importance to each compensation package by multiplying its monetary value with the number of persons working on any given package.

Thus, the number of persons working on different compensation packages serve here as weights. Though the weights, denoted as w_i , are like class frequencies, they are indeed not the same. This is despite the fact that weights enter into the computation of mean as do the class frequencies f_i . Accordingly, the arithmetic mean of a set of observations computed by taking into account the corresponding weights is known as the weighted arithmetic mean/average.

Irrespective of whether the given data refer to a sample or a population, the weighted mean is denoted as \bar{x}_w and is obtained as

$$\bar{x}_w = A + \frac{\sum w_i x_i}{\sum w_i} \quad (13.8)$$

where w_i 's are the weights, x_i 's are the sample observations in the case of ungrouped data and the mid-points in the case of grouped data.

It may be noted that the weighted mean for the grouped data may also be computed as

$$\bar{x}_w = A + \frac{\sum w_i d_i}{\sum w_i} \quad (13.9)$$

in which $d_i = (x_i - A)$. It can also be obtained as

$$\bar{x}_w = A + C \left(\frac{\sum w_i u_i}{\sum w_i} \right), \quad (13.10)$$

in which $u_i = (x_i - A)/C$ and C is the class interval that remains constant across all the classes of a distribution.

Illustration: Consider a company with managerial staff of 50 persons, who are distributed according to the value of annual compensation packages given under Col. (2) of Table 13.6. The number of persons working on each compensation package represents the weights, which are given under Col. (3).

Table 13.6 Computation of Arithmetic Mean

Managerial Positions	Value of Annual Compensation x_i (₹ lac)	No. of Persons Working w_i	$w_i x_i$
(1)	(2)	(3)	(4)
Senior Level Managers	4.50	5	22.50
Middle Rung Managers	2.75	10	27.50
Junior Executives	1.50	13	19.50
Supervisory Staff	0.75	22	16.50
	$\Sigma x_i = 9.50$	$\Sigma w_i = 50$	$\Sigma w_i x_i = 86.00$

In order to find the average annual compensation package for the entire managerial staff, we substitute the values to have

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i} = \frac{86.00}{50} = ₹ 1.72 \text{ lac.}$$

A simple average of the four compensation packages is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{9.50}{4} = ₹ 2.375 \text{ lac.}$$

which is relatively much higher than the weighted average. The difference in the two mean values is owing to the difference in weights. Had there been no difference in weights, \bar{x} and \bar{x}_w would have been the same.

Median, and its Computation

The median, denoted as M_p is a location average. It is the middle value in an ordered array of a set of observations. For locating median, observations comprising the given data are arranged first in an array, in a descending or ascending order of magnitude.

If the total number of observations is odd, median is the middle observation in the array. For example, in an array consisting of seven observations (2, 9, 10, 11, 15, 17, 21), median is the fourth observation (11) from either side. In general, in the case of sample data consisting of n observations median is the $(n + 1)/2$ th observation in the array. It is the $(N + 1)/2$ th observations in an arrayed population consisting of N observations.

If the number of observations is even, median is the mean of the middle two observations in the array. This holds irrespective of whether the given data pertain to a population or a sample. Had 23 been the 8th observation so as to have even number of observations in the above illustration, median would have been the average of the two middle observations, that is, $(11 + 15)/2 = 13$.

Computation of median for the grouped data is based on the concept of *median class*, which is located by obtaining *less than* cumulative frequencies. Median class is the one whose corresponding cumulative frequency covers $\sum f/2$.

Given the median class, median is obtained by substituting the values in

$$M_d = L_1 + \left[\frac{(\sum f_i/2) - C_f}{f_m} \right] C_m. \quad (13.11)$$

- Herein, i) L_1 is the lower limit of the median class,
ii) C_f is the cumulative frequency up to the class immediately preceding the median class, and
iii) C_m represents the width of the median class and f_m its class frequency.

NOTES

NOTES

Importantly, the method of computation of median as in Eq. (13.11) is the same both for sample and population distributions. The only difference is that $\sum f_i = n$ for a sample distribution, and $\sum f_i = N$ for a population distribution.

The use of Eq. (13.11) requires taking the following steps, and in that order:

- Find *less than* cumulative frequencies for each class, and divide the total number of frequencies ($\sum f_i$) by 2.
- Locate the class interval whose corresponding *less than* cumulative frequencies contain $(\sum f_i/2)$ or $(\sum f_i + 1)/2$, as the case may be. This class is known as the median class.
- Note the cumulative frequencies C_f up to the class that immediately precedes the median class.
- Also note the lower limit L_1 the class frequency f_m , and interval width C_m of the median class.
- Substitute the required in Eq. (13.11) to solve for median M_d .

Illustration: For the sample distribution given in Table 13.3, the computation of median is as demonstrated in Table 13.7.

Table 13.7 Computation of Median M_d

Class Intervals	Class Frequencies	Cumulative Frequencies
$L_1 - L_2$	f_i	f_c
1050–1059	6	6
1060–1069	9	15
1070–1079	15	30
1080–1089	25	55 ! (M_d class)
1090–1099	13	68
1100–1109	7	75
1110–1119	5	80
$n = \sum f_i = 80$		

Since $\sum f_i = 80$ is an even number, we obtain $(\sum f_i/2)$, and not $(\sum f_i + 1)/2$. As $(\sum f_i/2) = 80/2 = 40$ is contained in the cumulative frequency $f_c = 55$ against the (80 – 89) class, this class is the median class. With $C_f = 30$, $f_m = 25$, $C_m = 10$, and $L_1 = 1080$, we have

$$\begin{aligned}
 M_d &= L_1 + \left[\frac{(\sum f_i/2) - C_f}{f_m} \right] C_m \\
 &= 1080 + \left[\frac{(40 - 30)}{25} \right] \times 10 = 1084.
 \end{aligned}$$

Properties of Median

Like mean, median too has some interesting properties. These are as follows:

- i) That median for any set of data divides it into two equal halves. One half consists of observations smaller than the median. Observations in the other half are larger than the median.
- ii) That the sum of absolute deviations about the median is minimum. It may be recalled that arithmetic mean minimizes the sum of squared deviations, not the absolute deviations. Median, on the other hand, minimizes the sum of absolute deviations, that is, $\sum |x_i - M_a|$. The vertical bars on either side means summing without care to *plus* or *minus* signs.
- iii) That, as is apparent from the manner of its computation, median is not affected by the presence of unequal or open-ended class intervals.
- iv) That the value of median is determined by the position or location of observations in the array, and not by their size or magnitude. This is contrary to the method of computation of mean, in which mean is affected by the size of observations.
- v) That, unlike in the case of mean, the presence of extreme values in the data do not affect median so much. This qualifies median to be a more appropriate average for unevenly distributed data. Since such data have open-ended distributions, their mean is indeterminate.

NOTES

Locating Median from an Ogive

An ogive is a cumulative frequency curve usually of *less than* type. In order to locate median from an ogive, a cumulative frequency curve of *less than* type is drawn. A point is established on the vertical axis at $(\sum f_i/2)$, from where a line is drawn parallel to the horizontal axis allowing it to meet the ogive. A perpendicular is then drawn from the said point to read the value of median where it meets the horizontal axis.

Do Yourself: Re-lay the frequency distribution given in Table 13.2 with a column of cumulative frequencies of *less than* type as in Table 13.7, and determine the value of median in the manner explained above. Cross check your result to see that (167–168.9) is the median class with 167.167 as the median.

Mode, and its Computation

Mode is also a location average. In an ungrouped data, mode is that observation which occurs the maximum number of times. *For example*, in the data consisting of the observations (120, 125, 121, 115, 114, 115), mode is 115 as it occurs twice compared to others which occur only once.

Mode can be conveniently obtained by arranging a given set of observations in an array and counting the number of times each observation occurs. Having done that, mode is located as that particular observation which has occurred the maximum number of times.

NOTES

Locating mode in a frequency distribution is based on the concept of *modal class*. It is a class having the highest corresponding class frequency. In the event of there being two such classes, the one with relatively higher frequency in the preceding or succeeding (adjacent) class shall be the modal class.

For example, refer to the frequency distribution in Table 13.3. Assume further that the frequency against (1080–1089) class is 15, instead of 25. The rest being the same, there are now two classes, (1070–1079) and (1080–1089), having the highest (15) class frequencies. In this situation, (1080–1089) is the modal class, for the frequency (13) against the succeeding class (1090–1099) is more than the frequency (9) against the preceding class (1960–1969).

Once the modal class is located, the value of mode, denoted as M_o , is obtained by substituting the values in

$$M_o = L_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right] C_m \quad (13.12)$$

- Herein, i) L_1 is the lower limit and C_m the width of the modal class
 ii) Δ_1 represents the difference, ignoring *plus/minus* sign, between the frequency of the modal class and the class preceding it, and
 iii) Δ_2 is the difference between the frequency of the modal class and the class succeeding it.

Thus, the various steps in the computation of mode by using Eq. (13.12) are as follows:

- Locate the class having the maximum frequency, which is the modal class.
- Note the lower limit L_1 and width C_m of the modal class.
- Obtain Δ_1 and Δ_2 as defined in ii) and iii) above, respectively.
- Substitute the required values in Eq. (13.12) to solve for mode.

It may be noted that the above method of computation of mode is the same irrespective of whether the data belong to a sample or a population.

Illustration: For the distribution given in Table 13.8, the computation of mode is as shown therein. Since (1080–1089) class has the maximum (25) frequencies, it is the modal class. With $L_1 = 1080$, $\Delta_1 = (25 - 15) = 10$, $\Delta_2 = (25 - 13) = 12$, and $C_m = 10$, the value of mode is obtained by substituting the values in Eq. (13.12).

Thus, we have

$$\begin{aligned} M_o &= L_1 + \left[\frac{\Delta_1}{\Delta_1 + \Delta_2} \right] C_m \\ &= 1080 + \left[\frac{10}{10 + 12} \right] \times 10 = 1084.545. \end{aligned}$$

Table 13.8 Computation of Mode M_o

Class Intervals $L_1 - L_2$ (1)	Frequencies f_i (2)
1050–1059	6
1060–1069	9
1070–1079	15 $\Delta_1 = (25 - 15) = 10$
1080–1089	25 (<i>Modal Class</i>)
1090–1099	13 $\Delta_2 = (25 - 13) = 12$
1100–1109	7
1110–1119	5
$\Sigma f_i = 80$	

NOTES

Do Yourself: For the distribution in Table 13.2, check up that $\Delta_1 = 7$, $\Delta_2 = 2$ and mode $M_o = 166.556$.

Properties of Mode

Importantly, mode also possesses some unique properties. These are as follows:

- i) Mode is indeterminate in an ungrouped data possessing two or more observations occurring for the maximum number of times. By definition, there is no mode in a set of data wherein all observations occur with the same frequency.
- ii) In the case of grouped data, mode is the same as the mid-point of the modal class when the immediately preceding and succeeding classes have equal frequencies.
- iii) When the class preceding the modal class has higher frequency than that succeeding it, the value of mode is more than the mid-point of the modal class, and vice-versa.
- iv) Like median, the existence of unequal and open-ended classes does not interfere with the computation of mode. However, mode becomes indeterminate where the modal class is an open-ended class.

Further, mode is the least used measure of central tendency as compared to mean and median. The reason being that mode does not lend itself to any useful interpretations as median and, more so, mean do. While mode is relevant only when one or two observations occur more frequently than others, it adds nothing useful to describe the data when most or all observations occur almost the same number of times.

NOTES

Check Your Progress

1. How is mean obtained?
2. State an important characteristic of mean.

13.3 RANGE, QUARTILE, MEAN AND STANDARD DEVIATION

The Range, and How to Obtain It

Range is defined as the difference between the smallest and the largest observations in a given set of data. Obtaining range from an ungrouped data thus requires identifying only these two extreme values, and taking the difference between them. *For example*, if the smallest value in a set of data is 23 and the largest 73, range is $(73 - 23) = 50$.

In the case of grouped data, range can be obtained in the following two ways:

- i) *In the first*, range is found by taking the difference between the upper limit of the last class and the lower limit of the first class. This is because the lowest and the highest observations are not identifiable in the case of grouped data. For the distribution in Table 13.3, the lower limit of the first class is 1050 and the upper limit of the last class is 1119, resulting in $(1119 - 1050) = 69$ as the range.

It is obvious that range so obtained will be higher than the one based on the corresponding raw data. The reason being that in grouped data the lower limit of the first class interval is almost always smaller than the actual smallest value. Likewise, the upper limit of the last class interval is almost always higher than the actual highest value.

- ii) *In the second*, range is found by taking the difference between the midpoints of the first and the last class. This does yield a result closer to the actual range since it reduces the margin by which it is in error when computed by using the first method. Accordingly, for the distribution in Table 13.3, range is $[(1110 + 1119)/2] - [(1050 + 1059)/2] = 60$. It is an improved method of obtaining range, but only as long as the observations falling in these two extreme classes are uniformly distributed.

Besides being simple to compute and understand, range is as good a measure of dispersion as any other where the data consist of a few observations. This accounts for range being extensively used in establishing quality control charts. The major drawback of range, however, is that its computation is being based

only on two extreme values. Despite this limitation, range is often used to express the degree of dispersion in specific data such as stock prices. It is also commonly used in engineering and medical reports.

Quartile Deviation, and Computation

Quartile deviation, denoted as Q_D is defined as

$$Q_D = \frac{(Q_3 - Q_1)}{2} \quad (13.13)$$

Q_1 and Q_3 being the first and the third quartiles, quartile deviation is also known as *semi-quartile range*. The method of computation of Q_D in Eq. (13.13) is the same both for sample and population data. For the distribution in Table 13.3, Q_1 and Q_3 have already been computed in earlier as $Q_1 = 1073.33$, and $Q_3 = 1093.85$. Substituting these values in Eq. (13.13),

$$Q_D = \frac{(Q_3 - Q_1)}{2} = \frac{1093.85 - 1073.33}{2} = 10.26$$

Do yourself: Compute Q_1 and Q_3 for the distribution given in Table 13.9 and also obtain quartile deviation.

Table 13.9 Computation of Q_1 , Q_3 , and Q_D

Class Intervals $L_1 - L_2$	Frequencies f_i	Cumulative Frequencies f_c	
(1)	(2)	(3)	
161-162.9	3	3	
163-164.9	7	10	
165-166.9	14	24	(Q_1 class, as $(\Sigma f) \times (1/4) = 12.5$)
167-168.9	12	36	
169-170.9	10	46	(Q_3 class, as $(\Sigma f) \times (3/4) = 37.5$)
171-172.9	4	50	
$\Sigma f_i = 50$			

Since the i th quartile Q_i is computed as

$$Q_i = L_1 + \left(\frac{(\Sigma f_i) \times (i/4) - C_f}{f_i} \right) C_i,$$

$$Q_1 = L_1 + \left(\frac{(\Sigma f_i) \times (1/4) - C_f}{f_i} \right) C_1$$

$$= 165 + \left(\frac{12.5 - 10}{10} \right) \times 2 = 165.375$$

NOTES

NOTES

Similarly,

$$Q_3 = L_1 + \left(\frac{(\sum f_1) \times (3/4) - C_f}{f_3} \right) C_3$$

$$= 169 + \left(\frac{37.50 - 36}{10} \right) 2 = 169.300.$$

Substituting the values in Eq. (13.13),

$$Q_D = \frac{(Q_3 - Q_1)}{2} = \frac{169.300 - 165.375}{2} = 1.971.$$

In a symmetrical distribution, Q_1 and Q_3 are equidistantly placed from Q_2 the median. That is, the distance between Q_3 and Q_2 (or $Q_3 - Q_2$) is the same as the distance between Q_2 and Q_1 (or $Q_2 - Q_1$). It follows that $(M_d + Q_D)$ is the same as Q_3 , and $(M_d - Q_D)$ is the same as Q_1 .

These relationships do not, however, hold good in the case of skewed distributions which are typical of most business and economic time series data. But so long as a distribution is moderately skewed, $(M_d + Q_D)$ will yield a value quite close to Q_3 , and $(M_d - Q_D)$ quite close to Q_1 .

On the same logic as quartile deviation Q_D , it is possible to define what may be termed as *percentile deviation*. Denoted as P_D , it may be obtained as

$$P_D = \frac{(P_{90} - P_{10})}{2}$$

in which P_{90} and P_{10} are the 90th and 10th percentiles.

Mean Absolute Deviation, and Computation

The mean absolute deviation (*MAD*) measures the average deviation of a set of observations about their mean, ignoring the *plus/minus* sign of the deviations. It is computed by subtracting the mean from each individual observation, summing all the deviations ignoring the *sign*, and dividing the sum by the total number of observations. The sign of the deviations is ignored because otherwise the sum of the deviations from the mean [that is, $\sum(x_i - \bar{x})$] will be zero.

Thus, mean absolute deviation for a set of sample data consisting of n observations is computed as

$$MAD = \frac{\sum |x_i - \bar{x}|}{n} \quad (13.14)$$

in the case of ungrouped data. It is obtained as

$$MAD = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i} \quad (13.15)$$

in the case of grouped data, where x_i 's are the mid-points and $\sum f_i = n$.

Illustration: For the frequency distribution in Table 13.3, the various computations for mean absolute deviation are as shown in Table 13.10.

Table 13.10 Computation of MAD ($\bar{x} = 1083.375$)

Class Intervals $L_1 - L_2$	Frequencies f_i	Mid-Points x_i	$f_i x_i$ (2 × 3)	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $ (2 × 5)
(1)	(2)	(3)	(4)	(5)	(6)
1050-59	6	1054.5	06327.0	28.875	173.250
1060-69	9	1064.5	09580.5	18.875	169.870
1070-79	15	1074.5	16117.5	08.875	133.125
1080-89	25	1084.5	27112.5	01.125	028.125
1090-99	13	1094.5	14228.5	11.125	144.625
1100-09	7	1104.5	07731.5	21.125	147.875
1110-19	5	1114.5	05572.5	31.125	155.625
$n = \sum f_i = 80$			$\sum f_i x_i = 86670.0$ $\bar{x} = 1083.375$	$\sum f_i x_i - \bar{x} = 952.50$	

Substituting the values in Eq. (13.15),

$$MAD = \frac{\sum f_i |x_i - \bar{x}|}{\sum f_i} = \frac{952.50}{80} = 11.91.$$

Thus, the various steps for obtaining mean absolute deviation by using Eq. (13.15) are as follows:

- Establish class mid-points x_i , as $(L_1 + L_2)/2$.
- Obtain the product $f_i x_i$ and sum up to get $\sum f_i x_i$ and the mean \bar{x} .
- Find the absolute deviations as $|x_i - \bar{x}|$, get the product $f_i |x_i - \bar{x}|$ for all the classes, and sum up to obtain $\sum f_i |x_i - \bar{x}|$.
- Substitute the values in Eq. (13.15) to solve for MAD .

For a population consisting of N observations with mean μ , the mean absolute deviation is obtained as

$$MAD = \frac{\sum |(X_i - \mu)|}{N} \quad (13.16)$$

in the case of ungrouped data. It is computed as

$$MAD = \frac{\sum |(X_i - \mu)|}{N} \quad (13.17)$$

in the case of grouped data, where x_i 's are the class mid-points and $\sum f_i = N$.

Do Yourself: For the frequency distribution given in Table 13.9, compute mean absolute deviation by systematically making the necessary computations as in Table 13.10. Cross check that $\sum f_i x_i = 8359.50$, $\sum f_i |x_i - \bar{x}| = 111.52$ and $MAD = 2.230$.

NOTES

NOTES

Mean absolute deviation is relatively easy to compute and simple to understand. It is, however, not frequently used because of variance and standard deviation being the more precise and exact measures of dispersion. *MAD* has some definite utility in the area of inventory control.

Standard Deviation, and Computation

Standard deviation, *denoted as s*, is the positive square root of variance. If the variance is 49, standard deviation is $\sqrt{49} = 7$. Similarly, if variance is 121, standard deviation is $\sqrt{121} = 11$. This means that in order to find the standard deviation of a set of data, variance has to be computed first and then its square root taken.

Applications of Standard Deviation

Variance and standard deviation as the two principal measures of dispersion indicate the magnitude of deviations of a set of observations in terms of their distance from the mean. Standard deviation is expressed in the same unit of measurement as mean. Variance is, however, stated in square units (that is, Rs^2 , or $Qtl s^2$).

A relatively small variance means a high degree of uniformity in the data, with smaller overall divergence of individual observations from their mean. A high variance, on the other hand, indicates a greater degree of divergence of individual observations from the mean. This helps decide which of the two sets of data with the same mean value, is represented more adequately by their respective means. That is, a set of data with smaller variance is represented more adequately by its mean than the one with relatively larger variance.

Another important area where standard deviation has been extensively used relates to drawing statistical inferences. Generally, population characteristics (parameters) such as mean μ and standard deviation s are unknown. The populations being infinite, or time and cost factors prohibiting census operations, these parameters are estimated on the basis of information obtained through a sample.

Since it is necessary to know the reliability of sample based estimates of different population parameters, a perfectly symmetrical (normal) distribution is of immense help in determining the reliability of the estimates. This is made possible by the fact that the normal distribution covers varying percentage of observations that lie within $1s$, $2s$, and $3s$ on either side of its mean μ .

Approximately 68.27 per cent of the total observations are covered between $\mu + s$ and $\mu - s$ (that is, one standard deviation on either side of the mean μ). Similarly, 95.45 per cent and 99.73 per cent of the observations are covered between $\mu \pm 2s$ (two standard deviations on either side of the mean) and $\mu \pm 3s$ (three standard deviations on either side of the mean), respectively. These per cent area relationships are shown in Figure 13.1.

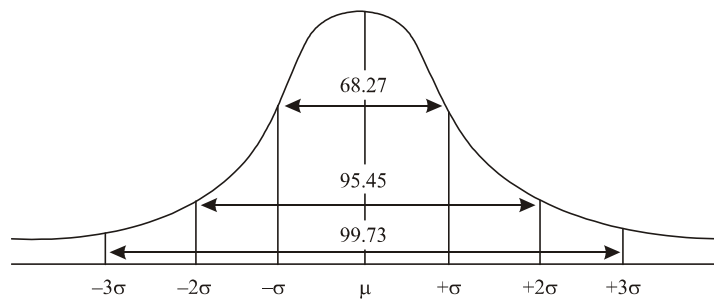


Fig. 13.1 Per cent Area Relationships

NOTES

Chebyshev's Theorem

The above area relationships are an extension of Chebyshev theorem, which gives an idea of the extent of dispersion of a set of data in terms of their mean \bar{x} and standard deviation s , regardless of the units in which the data are expressed. The theorem states that a given set of n observations on any variable X , the proportion of observations lying between k standard deviation from mean \bar{x} (that is, between $\bar{x} - ks$ and $\bar{x} + ks$) is at least $1 - (1/k^2)$, where k represents any positive value greater than 1.

Sheppard's Correction

Variance and standard deviation computed from grouped data always contain some error because of grouping of individual observations into different classes, called grouping. *Sheppard's correction* is a factor used for correcting variance for grouping errors. It is $1/12$ of the square of the width of class interval C , which is deducted from the computed variance.

That is,

$$\text{Corrected variance} = \text{Computed variance} - C^2/12, \quad (13.18)$$

wherein $C^2/12$ is known as Sheppard's correction for variance. It is applicable only in the case of frequency distributions of continuous variables. Statisticians are, however, not in agreement over the utility of using the correction factor.

This owes to the fear that it may lead to over-correction and may thus introduce fresh error. What is generally agreed upon is that the correction factor is not to be used without thorough examination of the problem.

Check Your Progress

3. How is range defined?
4. When is the mean absolute deviation used?

13.4 KARL PEARSON'S COEFFICIENT OF CORRELATION

NOTES

This measure is based on the fact that when a distribution drifts away from symmetry, its mean, median, and mode tend to deviate from each other. This comes about as a result of the presence of exceptionally high or low observations affecting the value of mean the most, and that of the mode the least.

The value of mean tends to be the highest, and that of mode the lowest, when some observations in a given set of data are exceptionally high. Consequently, a distribution having exceptionally high observations has a longer tail towards the right. On the contrary, mean tends to be the lowest, and mode the highest, when a set of data contains some exceptionally low observations. Resultantly, a distribution with exceptionally low observations has a longer tail towards the left.

Thus, it is the direction in which mode drifts from mean that determines whether a distribution will have positive or negative skewness. Using this outcome, the Pearsonian coefficient of skewness, denoted as Sk_p , is defined as

$$Sk_p = \frac{\text{Mean} - \text{Mode}}{s}, \quad (13.19)$$

in which s is the standard deviation. The empirical relationship among mean, median, and mode may be modified as

$$Sk_p = \frac{3(\text{Mean} - \text{Median})}{s}. \quad (13.20)$$

It may be noted that Sk_p in Eq. (13.19) is a relative measure of skewness. It indicates that the greater the divergence between mean and mode, the greater the skewness. It is negative when mode is more than mean, and positive when mode is less than mean.

When skewness is defined as in Eq. (13.19), it is known as Pearsonian *first coefficient* of skewness. It is called Pearsonian *second coefficient* of skewness when defined as in Eq. (13.20). Both being ratios, Sk_p is a pure number which varies between the limits -3 and $+3$. It takes positive sign in positively skewed distributions and negative sign in negatively skewed distributions. Sk_p is zero in symmetrical distributions, since mean, median, and mode are all equal in the case of such distributions.

13.5 RANK CORRELATION AND ATTRIBUTES

Correlation of Ranks r_s

Correlation of ranks is applied either when quantification of some information is not possible or where exact magnitudes are not ascertainable. A possible answer

in any such situation is to do ranking with reference to a particular characteristic. Ranks may be assigned either by two persons to a single characteristic or by a single person to two different characteristics.

As ranks are assigned in any first N natural numbers, ranking done either way offers two rank series. This allows obtaining a measure of correlation between ranks, known as Spearman's rank correlation. Denoted as r_s , it measures the degree of relationship between two rank series.

Consider, for example, the ranks X_1, X_2, \dots, X_N given to a characteristic A , and Y_1, Y_2, \dots, Y_N given to another characteristic B . The ranks being integers from 1 to N , a rank correlation between A and rank series can be obtained as

$$r_s = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} \quad (13.42)$$

Eq. (13.21) is obtained as follows:

Let \bar{X} and \bar{Y} be the arithmetic means of ranks given to characteristics A and B , respectively, so that

$$\bar{X} = \frac{\sum X}{N} \quad \text{or} \quad \sum X = N\bar{X},$$

and
$$\bar{Y} = \frac{\sum Y}{N} \quad \text{or} \quad \sum Y = N\bar{Y}.$$

Since the mean \bar{X} or \bar{Y} of first N integers is $(N+1)/2$,

$$\sum X = \frac{N(N+1)}{2} \quad \text{and} \quad \sum Y = \frac{N(N+1)}{2}.$$

That is,

$$\sum X = \sum Y = \frac{N(N+1)}{2}.$$

Similarly, as the variance S_X^2 or S_Y^2 of first N integers is $\frac{N^2 - 1}{12}$, we have

$$S_X^2 = \frac{\sum (X - \bar{X})^2}{N} = \frac{N^2 - 1}{12},$$

or

$$NS_X^2 = \sum (X - \bar{X})^2 = \frac{N(N^2 - 1)}{12} = \frac{N^3 - N}{12}.$$

NOTES

NOTES

Similarly,

$$NS_Y^2 = \sum (Y - \bar{Y})^2 = \frac{N(N^2 - 1)}{12} = \frac{N^3 - N}{12}.$$

Thus,

$$NS_X^2 = \sum (X - \bar{X})^2 \quad NS_Y^2 = \sum (Y - \bar{Y})^2 = \left(\frac{N^3 - N}{12} \right). \quad (i)$$

Let d_i be the difference in the i th rank given to A and B so that

$$d_i = (X - Y).$$

Since $\bar{X} = \bar{Y}$, $(X - Y)$ can be rewritten as

$$(X - Y) = (X - \bar{X}) - (Y - \bar{Y}). \quad (ii)$$

Writing x for $(Y - \bar{Y})$ and y for (\bar{Y}) , (i) may be restated as

$$\sum x^2 = \sum y^2 = \frac{1}{12}(N^3 - N). \quad (iii)$$

Taking the square on both sides of (ii), we have

$$\sum d_i^2 = \sum x^2 + \sum y^2 - 2\sum xy$$

or

$$2\sum xy = \sum x^2 + \sum y^2 - \sum d_i^2$$

or

$$2\sum xy = \left(\frac{N^3 - N}{12} \right) + \left(\frac{N^3 - N}{12} \right) - \sum d_i^2. \text{ [using (i)]}$$

or

$$\sum xy = \frac{1}{2} \left[\left(\frac{N^3 - N}{6} \right)^2 - \sum d_i^2 \right]. \quad (iv)$$

Now using (iii) and (iv) in Eq. (13.21) for the coefficient of correlation r , we have

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{\frac{1}{2} \left[\left(\frac{N^3 - N}{6} \right)^2 - \sum d_i^2 \right]}{\sqrt{\left[\frac{1}{12}(N^3 - N) \right] \left[\frac{1}{12}(N^3 - N) \right]}}$$

$$\begin{aligned}
 &= \frac{\frac{1}{2} \left[\left(\frac{N^3 - N}{6} \right) - \sum d_i^2 \right]}{\frac{1}{12} (N^3 - N)} = \frac{\frac{1}{2} \left[\frac{N^3 - N - 6 \sum d_i^2}{6} \right]}{\frac{1}{12} (N^3 - N)} \\
 &= \frac{N^3 - N - 6 \sum d_i^2}{(N^3 - N)} = 1 - \frac{6 \sum d_i^2}{(N^3 - N)} \\
 &= 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}.
 \end{aligned}$$

NOTES

Illustration: A leading company engaged in the production of detergents had 10 vacancies of salesmen for which ($N =$) 15 persons were called for personal interview. The interview board consisted of the Sales Manager and a Psychologist. The ranks given by the two to all the 15 candidates who attended the interview, compared as given in Cols. (2) and (3) of Table 13.10.

Table. 13.10 Computation of Rank Correlation, r_s

Sr. No. in the Interview List	Ranking by the Sales Manager (X_i)	Ranking by the Psychologist (Y_i)	$d_i = X_i - Y_i$	d_i^2
(1)	(2)	(3)	(4)	(5)
1	1	2	-1	1
2	3	3	0	0
4	2	1	1	1
5	4	5	-1	1
8	6	4	2	4
9	5	6	-1	1
10	7	8	-1	1
11	9	7	2	4
13	8	9	-1	1
14	11	10	1	1
15	10	12	-2	4
17	12	11	1	1
18	14	13	1	1
19	13	14	-1	1
20	15	15	0	0
				$\sum d_i^2 = 22$

Using the two sets of rank data, the computations required for obtaining the coefficient of rank correlation r_s are as given in Cols. (4) and (5) of Table 13.10.

Substituting $\sum d_i^2 = 22$ and $N = 15$ in Eq. (13.21), we have

$$\begin{aligned}
 r_s &= 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)} = 1 - \frac{(6)(22)}{(15)(225 - 1)} \\
 &= 1 - \frac{132}{3360} = \frac{3228}{3360} = 0.96.
 \end{aligned}$$

NOTES

Rank Correlation in the Case of Tied Ranks

There may be situations when the ranks assigned in two or more cases are the same. *For example*, suppose the ranking of 5 candidates in an interview conducted by an expert are 1, 2, 4, 2, 5. Here two ranks, second and fourth, are the same (2, 2). This being a situation of tie of ranks, such ranks may be called *tied ranks*.

When this obtains, we modify the tied ranks to the average of ranks that would have been assigned to them otherwise. Following this, the modified ranks will be 1, 2.5, 4, 2.5, 5.

The above modification for tied ranks is needed to ensure that the sum of ranks remains unaffected whether or not there are any tied ranks. Since the suggested modification affects the standard deviation of the concerned set of ranks, the factor $\sum d_i^2$ in Eq. (13.21) is to be corrected as

$$\sum d_c^2 = \sum d_i^2 + \left(\frac{t^3 - t}{12} \right), \quad (13.23)$$

in which $\sum d_c^2$ denotes the corrected $\sum d_i^2$, and $(t^3 - t)/12$ is the correction factor wherein t represents the number of tied ranks in set of ranks.

It may be noted that in obtaining $\sum d_c^2$, the correction factor $(t^3 - t)/12$ is to be added to $\sum d_i^2$ for every set of tied ranks in any paired rank data. *For example*, in the case of two sets of tied ranks,

$$\sum d_c^2 = \sum d_i^2 + \left(\frac{t^3 - t}{12} \right) + \left(\frac{t^3 - t}{12} \right). \quad (13.24)$$

Accordingly, the coefficient of rank correlation r_s in the case of tied ranks is computed as

$$r_s = 1 - \frac{6(\sum d_c^2)}{N(N^2 - 1)}. \quad (13.25)$$

Illustration: A selection board consisting of two experts for the post of general manager in a company interviewed 10 candidates whom the two experts assigned ranks as in Cols. (1) and (2) of Table 13.11. Since there are tied ranks occurring in the two rankings, the coefficient of rank correlation r_s is obtained by using Eq. (13.45), as under:

Table 13.11 Computation of Coefficient of Rank Correlation r_s

Sr. No.	Original Ranks		Adjusted Ranks		$d_i = (X_i - Y_i)$	d_i^2
	Expert I	Expert II	Expert I (X_i)	Expert II (Y_i)		
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	7	8	7	8	-1	1
2	9	10	9	10	-1	1
3	2	4	2	4	-2	4
4	4	6	4	6	-2	4
5	5	4	5.5	4	1.5	2.25
6	5	4	5.5	4	1.5	2.25
7	8	7	8	7	1	1
8	10	9	10	9	1	1
9	3	1	3	1	2	4
10	1	2	1	2	1	1
	54	55	55	55		$\Sigma d_i^2 = 21.5$

NOTES

Since rank 5 occurs twice and rank 4 occur thrice, the coefficient of correlation r_s is

$$r_s = 1 - \frac{6 \left[\Sigma d_i^2 + \left(\frac{t^3 - t}{12} \right) + \left(\frac{t^3 - t}{12} \right) \right]}{N(N^2 - 1)}$$

$$= 1 - \frac{6 \left[21.5 + \left(\frac{2^3 - 2}{12} \right) + \left(\frac{3^3 - 3}{12} \right) \right]}{10(10^2 - 1)} = 0.864.$$

13.6 SOLVED PROBLEMS

13.1 The marks obtained by 10 students in a semester examination in statistics are

70, 65, 68, 70, 75, 73, 80, 70, 83, and 86.

Find a) range, b) mean absolute deviation, and c) variance.

Solution

a) The lowest and the highest observations being 65 and 86, range is $(86 - 65) = 21$.

b) For $\bar{x} = 74$, the absolute deviations $|x_i - \bar{x}|$ are 4, 9, 6, 4, 1, 1, 6, 4, 9, and 12. Their sum being $\Sigma |x_i - \bar{x}| = 56$, the mean absolute deviation

$$MAD = \frac{\Sigma |x_i - \bar{x}|}{n} = \frac{56}{10} = 5.6$$

NOTES

c) Given the absolute deviations in b) above, their squares are 16, 81, 36, 16, 1, 1, 36, 16, 81, and 144,

whose sum $\sum (x_i - \bar{x})^2 = 428$, so that variance

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n} = \frac{428}{10} = 42.8$$

13.2 In a sample, 100 students doing a master programme in management were tested in a general knowledge paper carrying 100 marks. At the end of the exercise, they were found distributed according to marks obtained as under. Find a) the range of the distribution, b) quartile deviation, c) percentile deviation, and d) mean absolute deviation.

Marks Obtained	30–34	35–39	40–44	45–49	50–54	55–59	60–64
No. of Students	5	8	12	20	27	20	8

Solution

a) Range of the distribution can be found in two ways:

First, range as the difference between the lower limit of the first class and the upper limit of the last class is $(64 - 30) = 34$.

Second, range as the difference between the mid-points of the first and the last class intervals is $(62 - 32) = 30$.

b) Quartile deviation is obtained in terms of the first and the third quartiles, for which the necessary computations are as follows:

Class Intervals	f_i	f_c	
30–34	05	05	
35–39	08	13	→ P_{10} Class
40–44	12	25	→ Q_1 Class
45–49	20	45	
50–54	27	72	
55–59	20	92	→ Q_3 and P_{90} Class
60–64	08	100	

Substituting for the first quartile,

$$Q_1 = L_1 + \left(\frac{(\sum f_i) \times (1/4) - C_f}{f_i} \right) C_1 = 40 + \left(\frac{25 - 13}{12} \right) 5 = 45.00,$$

and for the third quartile,

$$Q_3 = L_1 + \left(\frac{(\sum f_i) \times (3/4) - C_f}{f_3} \right) C_3 = 55 + \left(\frac{75 - 72}{20} \right) 5 = 55.75.$$

Now substituting for quartile deviation,

$$Q_D = \frac{(Q_3 - Q_1)}{2} = \frac{55.75 - 45.00}{2} = 5.375.$$

c) Obtaining P_{90} and P_{10} as

$$P_{90} = L_1 + \left(\frac{(\sum f_i) \times (90/100) - C_f}{f_{90}} \right) C_{90} = 55 + \left(\frac{90 - 72}{20} \right) 5 = 59.5$$

and

$$P_{10} = L_1 + \left(\frac{(\sum f_i) \times (10/100) - C_f}{f_{90}} \right) C_{10} = 35 + \left(\frac{10 - 5}{8} \right) 5 = 38.125,$$

and substituting for percentile deviation, we have

$$P_D = \frac{(P_{90} - P_{10})}{2} = \frac{59.50 - 38.12}{2} = 10.69.$$

d) The mean absolute deviation requires the following computations:

Class Intervals	f_i	x_i	$f_i x_i$	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $
30-34	05	32	0160	17.4	087.00
35-39	08	37	0296	12.4	099.20
40-44	12	42	0504	07.4	088.80
45-49	20	47	0940	02.4	048.00
50-54	27	52	1404	02.6	070.20
55-59	20	57	1140	07.6	152.00
60-64	08	62	0496	12.6	100.80
$n = \sum f_i = 100$		$\sum f_i x_i = 4940$		$\sum f_i x_i - \bar{x} = 646.00$	
$\bar{x} = 49.40$					

Thus, absolute mean deviation

$$MAD = \frac{\sum f_i |(x_i - \bar{x})|}{\sum f_i} = \frac{646.00}{100} = 6.46.$$

NOTES

NOTES

13.3 For the sample distribution of marks obtained in general knowledge by 100 management students as in Prob. 13.2, find the variance using a) the long method, b) the alternative method, and c) the efficient method.

Solution

a) The necessary computations for using the long method are as given below:

Class Intervals	f_i	x_i	$f_i x_i$	$(x_i - \bar{x})$	$f_i(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$
30–34	05	32	0160	-17.4	302.76	1513.80
35–39	08	37	0296	-12.4	153.76	1230.08
40–44	12	42	0504	-07.4	054.76	0657.12
44–49	20	47	0940	-02.4	005.76	0115.20
50–54	27	52	1404	02.6	006.76	0182.52
54–59	20	57	1140	07.6	057.76	1155.20
60–64	08	62	0496	12.6	158.76	1270.08
$n = \sum f_i = 100$			$\sum f_i x_i = 4940$	$\sum f_i(x_i - \bar{x})^2 = 6124.00$		
				$\bar{x} = 49.40$		

Substituting for variance,

$$s^2 = \frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i} = \frac{6124}{100} = 61.24.$$

b) The necessary computations for obtaining variance by using the alternate simple method are as follows:

Class Intervals	f_i	x_i	$f_i x_i$	x_i^2	$f_i x_i^2$
30–34	05	32	0160	1024	05120
35–39	08	37	0296	1369	10952
40–44	12	42	0504	1764	21168
44–49	20	47	0940	2209	44180
50–54	27	52	1404	2704	73008
54–59	20	57	1140	3249	64980
60–64	08	62	0496	3844	30752
$n = \sum f_i = 100$			$\sum f_i x_i = 4940$	$\sum f_i x_i^2 = 250160$	

Substituting for variance,

$$s^2 = \frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i} \right)^2 = \frac{250160}{100} - \left(\frac{4940}{100} \right)^2 = 61.24.$$

c) Since the class interval C remains the same over all the classes, computations required for obtaining variance by using the efficient method are as under:

Class Intervals	f_i	x_i	$u_i = (x_i - A)/C$	$f_i u_i$	u_i^2	$f_i u_i^2$
30-34	05	32	-3	-15	9	45
35-39	08	37	-2	-16	4	32
40-44	12	42	-1	-12	1	12
44-49	20	47	0	0	0	0
50-54	27	52	1	27	1	27
54-59	20	57	2	40	4	80
60-64	08	62	3	24	9	72
$n = \sum f_i = 100$				$\sum f_i u_i = 48$		$\sum f_i u_i^2 = 268$

Substituting for variance,

$$s^2 = C^2 \left[\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum f_i u_i}{\sum f_i} \right)^2 \right] = 25 \left[\frac{268}{100} - \left(\frac{48}{100} \right)^2 \right] = 61.24.$$

13.4 A set of data consists of 10 observations. The mean of these observations was reported as 215 and that of their squares as 46850. Find the standard deviation.

Solution

The mean of $n = 10$ observations is $\bar{x} = \sum x_i / n = 215$, and the mean of their squares is $\bar{x}^2 = \sum x_i^2 / n = 46850$, the standard deviation s is

$$s = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2} = \sqrt{46850 - (215)^2} = 25$$

13.5 A distribution of 64 observations has a class interval which remains the same over all the classes. The mean of the squares of the step-deviations of class mid-points from an assumed mean is 185, with their mean as 11. Find the size of the class interval if the standard deviation of the distribution is 32.

Solution

Letting C be the size of the class interval, x_i 's the mid-points, and u_i 's the step deviations defined as $(x_i - A)/C$ with A as the assumed mean. Given are $\sum f_i = 64$, $\sum f_i u_i / \sum f_i = 11$, $\sum f_i u_i^2 / \sum f_i = 185$, and standard deviation $s = 32$.

NOTES

NOTES

Substituting the values for class interval C , we have

$$s = C \sqrt{\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum f_i u_i}{\sum f_i}\right)^2}$$

or

$$32 = C \sqrt{185 - (11)^2}$$

or

$$C = 32/8 = 4.$$

13.6 The mean and standard deviation of a set of 60 observations were found to be 120 and 32, respectively. Adjust both the quantities for a wrong entry of 80 in place of 110.

Solution

For adjusting mean $\bar{x} = 120$, we have $\sum x = n\bar{x} = (60 \times 120) = 7200$. With a wrong entry of 80 in place of 110, revised $\sum x_i = 7200 + (110 - 80) = 7230$.

and adjusted mean $\bar{x} = \frac{\sum x_i}{n} = \frac{7230}{60} = 120.50$.

Similarly, for adjusting variance $s^2 = 32$, $\sum x_i^2 = n(s^2 + \bar{x}^2) = 60(32 + 14400) = 865920$. With a wrong entry of 80 in place of 110, revised $\sum x_i^2 = 865920 - (80)^2 + (110)^2 = 871620$ and adjusted variance $s = \frac{\sum x_i^2}{n} - \bar{x}^2 = \frac{871620}{60} - (120.5)^2 = 6.75$.

13.7 A software company having 140 software engineers on its payroll, reported that on an average it pays a salary of ₹ 2.5 lac per annum with a coefficient of variance of 35 per cent. A sample of 60 engineers was selected and their average salary found to be ₹ 3.2 lac per annum with a standard deviation of ₹ 0.45 lac. Find the average salary and standard deviation of the remaining engineers.

Solution

Letting n_1 be the size of sample observed, so that $n = 140$, $n_1 = 60$, and $n_2 = 80$. Other quantities given are:

Combined mean $\bar{x}_c = 2.5$

Combined coefficient of variance $V_c = 35\%$

Mean of sample size $n_1, \bar{x}_1 = 3.2$

Standard deviation of $n_1, s_1 = 0.45$ (and $s^2 = 0.2025$).

With $V_c = 35$ per cent and $\bar{x}_c = 2.5$, the combined standard deviation is

$$V_c = (s_c / \bar{x}_c) \times 100$$

or

$$\begin{aligned} s_c &= (V_c \times \bar{x}_c) / 100 \\ &= (35 \times 2.5) / 100 = 0.875. \end{aligned}$$

Since the combined mean \bar{x}_c is

$$\bar{x}_c = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2},$$

we substitute the values and solve for \bar{x}_2 . Thus,

$$2.5 = \frac{60 \times 3.2 + 80 \times \bar{x}_2}{140}$$

or

$$\bar{x}_2 = ₹1.975 \text{ lac}$$

Now given the combined variance s_c^2 as

$$s_c^2 = \frac{n_1 (s_1^2 + d_1^2) + n_2 (s_2^2 + d_2^2)}{n_1 + n_2}$$

we substitute the values and solve for s_2^2 . With $d_1^2 = (\bar{x}_1 - \bar{x}_c)^2 = (3.2 - 2.5)^2 = 0.49$

and $d_2^2 = (\bar{x}_2 - \bar{x}_c)^2 = (1.975 - 2.5)^2 = 0.28$, we have

$$(0.875)^2 = \frac{60(0.2025 + 0.49) + 80(s_2^2 + 0.28)}{140}$$

$$107.18 = \frac{41.55 + 80s_2^2 + 22.40}{140}$$

$$s_2^2 = 0.154 \text{ or } s_2 = 0.735.$$

13.8 A factory accountant who distributed increased DA to 20 low-paid employees reported to the management that on an average each employee got additional DA of ₹ 85.65 with a standard deviation of ₹ 35.20. Having had a relook on the details, he discovered that in the case of two employees wrong DA figures of ₹ 120 and ₹ 95 were entered in place of correct figure of ₹ 98 and ₹ 105, respectively. Find the correct average DA paid and standard deviation thereof.

NOTES

NOTES

Solution

Given are $\bar{x} = 85.65$, $s = 35.20$ (or $s^2 = 1239.04$), and $n = 20$.

Since $\bar{x} = \sum x_i/n$, $\sum x_i = n\bar{x} = (85.65 \times 20) = 1713.00$.

Then, corrected sum $\sum x_i = 1713.00 - (120 + 95) + (98 + 105) = 1701.00$,

and corrected mean $\bar{x} = \frac{\sum x_i}{n} = \frac{1701}{20} = 85.05$.

Since $s^2 = (\sum x_i^2/n) - \bar{x}^2$, $\sum x_i^2 = n(s^2 + \bar{x}^2) = 20(1239.04 + 7335.92) = 171499.22$.

Then, corrected sum $\sum x_i^2 = 171499.22 - [(120)^2 + (95)^2] +$
 $[(98)^2 + (105)^2] = 168703.22$,

and corrected standard deviation $s = \sqrt{\frac{\sum x_i^2}{n} - \bar{x}^2} = \sqrt{\frac{168703.22}{20} - (85.05)^2} = 34.66$.

13.9 On examining the sales data for 20 sales regions for a particular year, a sales analyst found that sales observed a symmetric distribution. During the course of sales analysis, he found per region average sales to be worth ₹ 2.5 lac with a variance of ₹ 0.65 lac. The difference between the highest and the lowest sales was noted as ₹ 1.22 lac. Finding them too unusual, he wanted to know the average sales and variance, ignoring the highest and the lowest sales. Help the sales analyst.

Solution

It may be noted that we are given the range of the distribution $R = 1.22$. In the case of a symmetrical distribution, mean lies in the middle of the range so that half of the range ($R/2 = 1.22/2$) = 0.61 lies on either side of the mean $\bar{x} = 2.5$.

Thus, the highest sales are $(\bar{x} + R/2) = 2.5 + 0.61 = ₹ 3.16$ lac and the lowest sales are $(\bar{x} - R/2) = 2.5 - 0.61 = ₹ 1.84$ lac. It is required to exclude these two extreme sales from the computation of mean and variance of sales.

Since $\sum x_i = n\bar{x} = 20 \times 2.5 = 50$, the revised sum $\sum x_i = 50 - (3.16 + 1.84) = 45.00$. Thus, the revised average sales \bar{x} for 18 regions is

$$\bar{x} = \frac{\sum x_i}{n} = \frac{45}{18} = ₹ 2.50 \text{ lac.}$$

Similarly, as $\sum x_i^2 = n(s^2 + \bar{x}^2) = 20(0.65 + 6.25) = 138.00$, the revised sum $\sum x_i^2 = 138 - [(3.16)^2 + (1.84)^2] = 124.63$.

Thus, the revised sales variance for 18 regions is

$$s^2 = \frac{\sum x_i^2}{N} - \bar{x}^2 = \frac{124.63}{18} - (2.50)^2 = ₹ 0.67 \text{ lac.}$$

13.10 Using the frequency distribution in Prob. 4.5, as reproduced below, find a) the range, b) quartile deviation, c) mean absolute deviation, d) variance, and e) standard deviation.

Class Intervals	50–52	53–55	56–58	59–61	62–64
Frequencies	5	10	21	8	6

Solution

The lower limit of the first class is 50 and the upper limit of the last class is 64. Being difference of the two, range is $(64 - 50) = 14$. Alternatively, as difference between the mid-points of the first class (51) and the last class (63), range is $(63 - 51) = 12$.

As already computed under Prob. 4.5, $Q_1 = 55.25$, $Q_3 = 59.56$. Thus, the quartile deviation

$$Q_d = \frac{(Q_3 - Q_1)}{2} = \frac{59.56 - 55.25}{2} = 2.16.$$

c) The mean \bar{x} as computed under Prob. 4.5 is 57.00. Then, the necessary computations for obtaining MAD are as follows:

Class Intervals	f_i	x_i	$ x_i - \bar{x} $	$f_i x_i - \bar{x} $
50–52	5	51	6	30
53–55	10	54	3	30
56–58	21	57	0	0
59–61	8	60	3	24
62–64	6	63	6	36
$\Sigma f_i = 50$		$\Sigma f_i x_i - \bar{x} = 120$		

NOTES

NOTES

Substituting for mean absolute deviation,

$$MAD = \frac{\sum f_i |(x_i - \bar{x})|}{\sum f_i} = \frac{120}{50} = 2.4.$$

d) The required computations for obtaining variance are as under:

$L_1 - L_2$	f_i	x_i	$f_i x_i$	x_i^2	$f_i x_i^2$
50-52	5	51	255	2601	13005
53-55	10	54	540	2916	29160
56-58	21	57	1197	3249	68229
59-61	8	60	480	3600	28800
62-64	6	63	378	3969	23814
$\Sigma f_i = 50$		$\Sigma f_i x_i = 2850$		$\Sigma f_i x_i^2 = 163008$	

Substituting for variance,

$$\begin{aligned} s^2 &= \frac{\sum f_i x_i^2}{\sum f_i} - \left(\frac{\sum f_i x_i}{\sum f_i} \right)^2 \\ &= \frac{163008}{50} - \left(\frac{2850}{50} \right)^2 \\ &= 3260.16 - 3249 = 11.16. \end{aligned}$$

e) Since standard deviation s is the under-root of variance s^2 ,

$$s^2 = \sqrt{s^2} = \sqrt{11.16} = 3.34.$$

13.11 Which one of the two distribution series given below is more consistent?

Class Intervals	Series A	Series
10-20	10	18
20-30	16	22
30-40	34	38
40-50	38	34
50-60	24	20
60-70	18	8

Solution

In order to find which of the two series is more consistent, we compute and compare the coefficients of variation as below:

Class Intervals	Series A					Series B		
	x_i	$u_i = (x_i - A)/C$	f_i	$f_i u_i$	$f_i u_i^2$	f_i	$f_i u_i$	$f_i u_i^2$
10–20	15	-3	10	-30	90	18	-54	162
20–30	25	-2	16	-32	64	22	-44	88
30–40	35	-1	34	-34	34	38	-38	38
40–50	45 = A	0	38	0	0	34	0	0
50–60	55	1	24	24	24	20	20	20
60–70	65	2	18	36	72	8	16	32
			140	-36	284	140	-100	340

NOTES

For series A:

$$\bar{x} = A + C \left(\frac{\sum f_i u_i}{\sum f_i} \right) = 45 + 10 \left(\frac{-36}{140} \right) = 42.43,$$

$$s = C \sqrt{\frac{\sum f_i u_i^2}{\sum f_i} - \left(\frac{\sum f_i u_i}{\sum f_i} \right)^2} = 10 \sqrt{\frac{284}{140} - \left(\frac{-36}{140} \right)^2} = 14,$$

and

$$V_1 = \frac{s}{\bar{x}} \times 100 = \frac{14}{42.43} \times 100 = 33.00\%.$$

For series B:

$$\bar{x} = A + C \left(\frac{\sum f_i u_i}{\sum f_i} \right) = 45 + 10 \left(\frac{-100}{140} \right) = 37.86,$$

$$\bar{x} = A + C \left(\frac{\sum f_i u_i}{\sum f_i} \right) = 45 + 10 \left(\frac{-100}{140} \right) = 37.86,$$

and

$$V_2 = \frac{s}{\bar{x}} \times 100 = \frac{13.85}{37.86} \times 100 = 36.58\%.$$

Since $V_1 < V_2$, series A is more consistent as compared to series B.

NOTES

13.12 A distribution of 100 households according to their annual savings revealed an average saving of ₹ 4,500 per annum and a variance of ₹ 650. Another distribution of 200 households according to their annual income showed an average income of ₹ 48,000 per annum and variance of ₹ 1,800. Compare the two distributions to identify which one of them has a greater degree of dispersion.

Solution

For the distribution of savings with a mean of ₹ 4500 and variance of 650 (or $s = 25.49$), coefficient of variation

$$V_1 = \left(\frac{s}{\bar{x}} \right) 100 = \left(\frac{25.49}{4500} \right) 100 = 0.566.$$

Similarly, for the distribution of income with a mean of ₹ 48,000 and variance of 1800 (or $s = 42.43$), coefficient of variation

$$V_2 = \left(\frac{s}{\bar{x}} \right) 100 = \left(\frac{42.43}{48000} \right) 100 = 0.088.$$

Since $V_1 > V_2$, it means household savings show greater dispersion than household income. That is, the distribution of income is more stable.

13.13 A company manufacturing a popular brand of home air-conditioners wanted to have an idea of the income background of their brand users. A sample of 100 users revealed that the distribution of their family income (from all sources) had moderately positive skewness with income variance of ₹ 1,200. Find a) the approximate value of quartile deviation of income, and b) an estimate of mean absolute deviation in income.

Solution

For a moderately skewed distribution, the empirical relationship between quartile deviation and standard deviation is $Q_d = 2/3$ (standard deviation) = $2/3$ (34.64) = 23.09.

Similarly, by its empirical relationship with standard deviation, mean absolute deviation is $MAD = 4/5$ (standard deviation) = $4/5$ (34.64) = 27.71.

13.14 Coefficients of variation of two series are 75 per cent and 90 per cent, with standard deviations 15 and 18, respectively, a) Obtain the mean of the two series, b) Also comment which of the two is better represented by its mean.

Solution

Let the two series be designated as I and II.

For series I, the coefficient of variation V_1 is

$$V_1 = \frac{s_1}{\bar{x}_1} \times 100 \text{ or } 75 = \frac{15}{\bar{x}} \times 100$$

or
$$\bar{x}_1 = \frac{1500}{75} = 20.$$

For series II, the coefficient of variation V_2 is

$$V_2 = \frac{s_2}{\bar{x}_2} \times 100 \text{ or } 90 = \frac{18}{\bar{x}_2} \times 100$$

or
$$\bar{x}_1 = \frac{1800}{90} = 20.$$

This means both the series have the same mean.

Even though both the series have the same mean, the one with lower standard deviation is better represented by its mean.

NOTES

Check Your Progress

5. When is correlation of ranks applied?
6. On what fact is Karl Pearson's coefficient of correlation based on?

13.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Mean is obtained by taking the sum of all observations comprising a given set of data and dividing the sum by the total number of observations.
2. An important characteristic of mean is that it satisfies some useful properties. No other measure of central tendency except mean does this.
3. Range is defined as the difference between the smallest and the largest observations in a given set of data.
4. The mean absolute deviation (*MAD*) measures the average deviation of a set of observations about their mean, ignoring the *plus/minus* sign of the deviations.

NOTES

5. Correlation of ranks is applied either when quantification of some information is not possible or where exact magnitudes are not ascertainable.
6. This measure is based on the fact that when a distribution drifts away from symmetry, its mean, median, and mode tend to deviate from each other.

13.8 SUMMARY

- Popularly known as *average*, arithmetic mean (or simply, mean) is the most important and frequently used measure of central tendency. It possesses certain useful properties, which account for its wider use.
- The method of computation of mean readily enables us to correct the value of mean for any wrong entries that may have occurred during the course of summation.
- An important characteristic of mean is that it satisfies some useful properties. No other measure of central tendency except mean does this. This holds both for the ungrouped and the grouped data, and irrespective of whether the data refer to a sample or a population.
- There are situations when all the n individual observations in a set of data are not equally important. This does not allow arithmetic mean to serve as the best representative of the given data.
- Where individual observations vary in importance, they are assigned weights according to the level of importance of each in the computation of their mean.
- The median, *denoted as Md* , is a location average. It is the middle value in an ordered array of a set of observations. For locating median, observations comprising the given data are arranged first in an array, in a descending or ascending order of magnitude.
- Mode is also a location average. In an ungrouped data, mode is that observation which occurs the maximum number of times.
- Mode can be conveniently obtained by arranging a given set of observations in an array and counting the number of times each observation occurs. Having done that, mode is located as that particular observation which has occurred the maximum number of times.
- Range is defined as the difference between the smallest and the largest observations in a given set of data. Obtaining range from an ungrouped data thus requires identifying only these two extreme values, and taking the difference between them.

- The mean absolute deviation (*MAD*) measures the average deviation of a set of observations about their mean, ignoring the *plus/minus* sign of the deviations.
- Variance and standard deviation as the two principal measures of dispersion indicate the magnitude of deviations of a set of observations in terms of their distance from the mean.
- Variance and standard deviation computed from grouped data always contain some error because of grouping of individual observations into different classes, called grouping. *Sheppard's correction* is a factor used for correcting variance for grouping errors.
- Correlation of ranks is applied either when quantification of some information is not possible or where exact magnitudes are not ascertainable.

NOTES

13.9 KEY WORDS

- **Sheppard's correction:** It is a factor used for correcting variance for grouping errors. It is $1/12$ of the square of the width of class interval C , which is deducted from the computed variance.
- **Standard deviation:** Standard deviation, denoted as s , is the positive square root of variance.

13.10 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Write a short note on weighted arithmetic mean.
2. State the properties of median.
3. State the applications of standard deviation.
4. Write a short note on correlation of ranks.

Long-Answer Questions

1. Analyse the properties of arithmetic mean.
2. Describe the computation of median with the help of examples.
3. Discuss the process of the computation of range and quartile deviation.
4. Explain the measure of Karl Pearson's coefficient of correlation.

NOTES

13.11 FURTHER READINGS

- Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.
- Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.
- Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.
- Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.

BLOCK - V
PREPARATION OF A RESEARCH REPORT

*Preparation of
Research Report*

**UNIT 14 PREPARATION OF
RESEARCH REPORT**

NOTES

Structure

- 14.0 Introduction
- 14.1 Objectives
- 14.2 Research Report
 - 14.2.1 Types of Reports
 - 14.2.2 Characteristics of a Good Report
 - 14.2.3 Mechanics of Writing a Report
- 14.3 Analysis of Data
- 14.4 Answers to Check Your Progress Questions
- 14.5 Summary
- 14.6 Key Words
- 14.7 Self Assessment Questions and Exercises
- 14.8 Further Readings

14.0 INTRODUCTION

A research report is a formal statement of the research process and its results. A research activity is concluded by presenting results that include both major and minor recommendations. The reporting of a research study depends on the purpose with which it was undertaken. Research studies when reported, follow certain standard patterns, styles and formats for maintaining parity in reporting and for easy grasp by others who are concerned with those studies. They narrate the problem studied, method used for studying it, and the findings and conclusions of the study.

The purpose of writing a research report is to communicate to interested people, the methodology and results of the study in such a manner as to enable them to understand the research process and to determine the validity of the conclusions.

The aim of the report is not to convince the reader about the value of the result, but to convey to him what was done, why it was done and what was the outcome.

14.1 OBJECTIVES

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

- Describe the different types of reports
- Analyse the characteristics of a good report

- Describe the mechanics of writing a report
- Explain the process of data analysis

NOTES

14.2 RESEARCH REPORT

A **report** can be defined as a written document which presents information in a specialized and concise manner. 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 short and conventional format. A report should cover all mandatory matters but nothing extra should be written. For writing a report, at 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.

14.2.1 Types of Reports

Reports can be divided into different categories. The two main types of reports are:

- Informational report
- Interpretive report

Informational report

A 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 as follows:

- **Inspection reports:** Reports which show the outcome of products or equipment to assure their proper functioning or to describe their quality are called inspection reports. This type of report is mainly used in manufacturing organizations.
- **Inventory reports:** Reports which are made to keep stock of various things like furniture, equipment, stationery, utensils and other accessories are called inventory reports.
- **Assessment reports:** 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 reports which are made to measure the performance of the employees in an organization for different purposes like appraisal or promotion are called performance reports.

Interpretive report

An interpretive report contains 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 informational reports because they contain different elements. The possible elements that can be used in interpretive reports are:

- Cover
- Frontpiece
- Title page
- Copyright notice
- Forwarding letter
- Preface
- Acknowledgements
- Table of contents
- List of illustrations
- Abstract and summary
- Introduction
- Discussion
- Conclusions
- Recommendations
- Appendices
- List of references
- Bibliography
- Glossary
- Index

14.2.2 Characteristics of a Good Report

The characteristics of a good report can be classified under the following four heads:

- Language and style of the report
- Structure of the report
- Presentation of the report
- References in the report

Each of the above 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

NOTES

NOTES

writer needs to avoid overcrowding of information that may make the reader feel confused and lost in reading the data, thereby losing its charm. A systematic scrutiny of each of these aspects of a report is, therefore, necessary.

1. Language and style of a report

A report must have a logical structure with a 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 and informative title for the content of the 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

2. Structure of a 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 subheadings:
 - 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:**
 - o Start with the technical chapters/sections.
 - o Follow with the discussion.
 - 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, etc.
 - o Bibliography, i.e., list of all the sources of material, you referred to in your report.

3. Presentation of a report

As stated earlier, mere data overloading or just a lucid style of writing is not only necessary for 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 the 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, etc.

NOTES

NOTES

- 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:**
 - o Sheets should be of white A4 size and 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.

4. References in a 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 should avoid any violation of copyright laws and plagiarism. The necessary rule of thumb in this regard can be stated as follows:

- o **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 have understood what you have read and know how to apply it, to your own context.
 - Direct quotes should be used sparingly.
- o **Direct quotes:**
 - **Short direct quotes:** These need to be placed between quotation marks. For example, 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 forty 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.

NOTES

14.2.3 Mechanics of Writing a Report

There are several parameters that are strictly followed while 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 $8\frac{1}{2}'' \times 11''$ 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 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:
 - o 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.

NOTES

o In case of a multivolume 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*, June 2004, Whitefish, Volume 2, p.124.

o In case of works arranged alphabetically:

- For works arranged alphabetically such as encyclopedias and dictionaries, no page reference is usually 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.

o 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:

Shahad, P.V. 'Rajesh Jain's Ecosystem', in *Business Today*, Vol. 14, December 18, p. 28, 2005.

o 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 multiple authorship is indicated by 'et al' or 'and others'.

- Author's name in normal order
- Title of work, underlined to indicate italics
- Place and date of publication
- Pagination references

For example:

Alexandra K. Wigdor, *etal.*, *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 to. Some English and Latin abbreviations are often used in bibliographies and footnotes to eliminate any repetition.

Table 14.1 shows the various English and Latin abbreviations used in bibliographies and footnotes.

NOTES

Table 14.1 *English and Latin Abbreviations used in Bibliographies and Footnotes*

Abbreviations	Meaning
Anon.,	Anonymous
Ante.,	Before
Art.,	Article
Aug.,	Augmented
bk.,	Book
bull.,	Bulletin
cf.,	Compare
ch.,	Chapter
col.,	Column
diss.,	Dissertation
ed.,	editor, edition, edited
ed. cit.,	edition cited
e.g.	exempli gratia: for example
eng.,	Enlarged
et al.,	and others
et seq.,	et sequens: and the following
ex.,	Example
f.,ff.,	figure(s)
fn.,	Footnote
ibid.,ibidem	in the same place
id.,idem.,	the same
ill.,illus., or illust(s)	illustrated, illustration(s)
Intro., intro.,	introduction
l.,ll.,	line(s)
loc. cit.,	in the place cited; used as op.cit.,
MS.,MSS.,	Manuscript(s)
N.B. nota bene	note well
n.d.,	no date
n.p.,	no place
no pub.,	no publisher
no(s) .,	number(s)
o.p.,	out of print
op.cit:	in the work cited
p.pp	page(s)
passim:	here and there
Post:	After

NOTES

- **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.

Check Your Progress

1. What are the two main types of reports?
2. State the characteristics of a good report.

14.3 ANALYSIS OF DATA

Being a subject of much practical utility and having wide-ranging applications, statistics displays a unique strength. It suffers from an important weakness as well. All in all, it is spreading its tentacles far and wide.

1. *The strength:* The greatest strength of statistics as a subject lies in developing a statistical mode of thinking, in imparting an orientation to the mind to think statistically. This is of specific relevance in a modern society where governance is no longer circumscribed by the day-to-day administration of the matters of the state. Governments and other state agencies now remain constantly engaged in various activities encompassing the whole gamut of the functions of a corporate manager and a development planner.

This calls for collection and compilation of massive data on all such characteristics of the subjects of the state (such as the level of education, income, occupation, sex, age, marital status, or the like) as are necessary for effective planning of the developmental activities of the state. All these data, often collected over time, are carefully studied and systematically analysed with a view to seeking useful insights and reaching statistically valid conclusions for sound decision-making.

Thus, the wide diversity of data we face and the statistical tools that are applied for data analysis do, together, impel us to think statistically. We are gently coaxed into the statistical thinking mode, while:

- (i) bringing out the pattern of variations in the available data in a given problem situation on one or more relevant characteristic(s), and

- (ii) training the mind in comparative dimensions of data analysis, examining the consequent variations, drawing inferences, and establishing plausible relationships.

So long as the process of collection and analysis of data is devoid of comparative inputs, it fails to offer useful and reliable results for any meaningful decision activity. For, a set of data compiled without serious thought and deeper insights, proves a mere waste of efforts. Apart from providing defences against any such possibilities, statistics cultivates a resilient mind with an astute statistical sense alive to the dangers involved.

2. *The weakness:* The general feeling of distrust in it is an important weakness of statistics. It emanates from the often-held view that the data, to which statistical methods are applied, lack the desired element of accuracy. As a result, the conclusions and inferences drawn from data analysis cannot be claimed as being adequately reliable. In support of these fears, a commoner may cite frequent cases of media reports and other officially sponsored public relation material which, he feels, are generally based on inadequate, manipulated, and unreliable data.

To the extent that this apprehension may be taken as based on tactual situations, the real culprit are those who compile, collect, and project data in a given light. Even if data inaccuracy is otherwise taken as being too serious a flaw of statistics, there is really no escape from it. The whole process of data collection, compilation, and tabulation is, indeed, too porous, and does allow room for numerous errors. These can at best be minimized, but can not be eliminated altogether.

Since complete accuracy can not be ensured in the absolute sense, considerations of reliability and trust are relevant only in the relative terms. The facts being what they are, the saying that ‘working on some information is better than doing without any information,’ is a useful common sense maxim. This should, and often does, greatly soften our attitude towards the lack of reliability of statistical data and the consequent distrust in statistics.

3. *The increasing tentacles:* Despite the feeling that statistical data are not often very trustworthy, statistics has evolved fairly objective methods of evaluating the element of errors that erodes data reliability. Assuming sincerity and fair play on the part of those involved in data collection and processing, the available means of estimating the extent of errors in data-based results have greatly reduced the reason for distrust in statistical data. All said and done on this score, statistics has, as of now, established itself as a generic and versatile subject of study. The more one gets to know of it, the more one imbibes of its subtle impact in terms of the mental ability to draw fairly valid conclusions even from limited data. And, it is precisely owing to this reason that the applications of statistical methods have fast spread its tentacles to the various important areas of human interest.

NOTES

NOTES

The Language of Statistics

In developing necessary methods of data analysis and interpretation, statistics makes use of some common *notations* that have become fairly standardized. As notations are the shorthand expressions of concepts and statements, they constitute the *language of statistics*. We must, therefore, adequately apprise ourselves of these notations, since they go into the making of the essential body of statistical methods and constitutes its *lingua franca*.

Notations for Variables

Statistical data, as the raw material of statistics, are the observations on variables, continuous or discrete. While a variable is represented by the Roman capital letter X , the possible values that it may take are denoted by X_1, X_2, X_3, \dots

In developing methods involving two variables, the second variable is, generally, represented by Y and the values that it may take by Y_1, Y_2, Y_3, \dots . Where interest lies in simultaneously dealing with three variables, Z may be used to represent the third variable with values notated as Z_1, Z_2, Z_3, \dots

Alternatively, the three variables may be represented by X_1, X_2 , and X_3 , with possible values that these may take denoted as $X_{11}, X_{12}, X_{13}, \dots, X_{21}, X_{22}, X_{23}, \dots$, and $X_{31}, X_{32}, X_{33}, \dots$, respectively.

For example, national income and investment as two different variables can be represented by X and Y . In the event of export entering into the picture as the third variable, it may be appropriate to represent national income by X_1 , investment by X_2 , and exports by X_3 .

Notations for Observed Data

So long as we refer to variables, these are expressed in terms of notations in the manner stated above. But as soon as a variable X is observed to have taken definite values, the observed values are represented by the Roman small letter x as x_1, x_2, x_3, \dots . Similarly, the observed values of variable Y are denoted as y_1, y_2, y_3, \dots

For example, when X represents national income, what it is actually going to be next year, next-to-next year, or in any future year, is not known. As such, X is a variable that can take any value X_1, X_2, X_3, \dots tomorrow or the day after. When national income for the next year becomes known as a *definite figure*, it is denoted as x_1 and is read as a particular observed value of the variable X . For the next-to-next year, national income actually observed is denoted as E_2 , and so on. It follows that when X represents a particular variable, X_1, X_2, X_3, \dots are the unknown possible values it may take, and x_1, E_2, E_3, \dots are the values actually observed.

In statistics, we often deal with variables, and observed data on variables, by developing methods aimed at describing their characteristics. A distinction so made between X_1, X_2, X_3 and x_1, x_2, x_3, \dots is important and must be understood

as a statistical process. For, it is essentially hidden behind the body of methods that comprise the scope and tool-kit of statistics.

Notations in Sampling

The size of sample and that of population are important in developing methods applied to a set of sample or population data. The necessary requirement for effecting sampling is that the observations comprising the population are known and finite. Accordingly, population size is denoted as N and the sample size as n .

Importantly, variables represent infinite populations. Accordingly, X as a variable can take an infinite number of unknown values X_1, X_2, X_3, \dots . But in the context of sampling applied to a finite population of size N , a significant departure is made in the notations used for the population values (units or observations). Even as N represents the number of values comprising a population that are known/observed data, these are denoted by the capital letter X as $X_1, X_2, X_3, \dots, X_N$ and not by the lower case letter x as $x_1, x_2, x_3, \dots, x_N$.

Representing N observed values of a population by $X_1, X_2, X_3, \dots, X_N$ is necessary to distinguish them from the n sample values to be denoted as $x_1, x_2, x_3, \dots, x_n$. Also, owing to the fact that for purposes of sampling, a given finite population deserves to be seen differently from x_1, x_2, x_3, \dots as the observed values of a variable as a statistical process. Compared to this, $X_1, X_2, X_3, \dots, X_N$ should be understood as clearly identifiable N observations of a finite population.

For example, consider the 2,000 students of a given college whose body weight is a characteristic of our interest. Body weight as such is a variable, but for purposes of sampling and observing the weight of those selected in the sample, the population comprises all the 2,000 college students. However, when each student is observed for body weight, the resultant $N = 2,000$ weight figures are in fact the population observations, which are denoted as $X_1, X_2, X_3, \dots, X_N$. But as soon as this is done, sampling loses all its meaning and relevance because in that case it becomes a full count.

Notations for Sample Statistics

Here the word statistics has a different meaning. It refers to all such values or quantities as we may compute from a sample of size n randomly drawn from a finite population consisting of N observations. *For example*, arithmetic mean of n sample observations $x_1, x_2, x_3, \dots, x_n$, denoted as \bar{x} , is obtained by taking the sum of all the n observations and dividing it by n . Likewise, a sample proportion is computed as the ratio of the number of items meeting (or not meeting) a certain characteristic to the total number of items in the sample. Such as the proportion of defective items in a sample of n items taken from a given lot offered for sale.

Sampling being random, sample arithmetic mean in repeated experiments is a random variable. It is then denoted as \bar{X} with possible values represented by $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots$. It is only after selecting a sample and computing the sample mean

NOTES

NOTES

\bar{x} that it represents a particular observed value of \bar{X} as a variable. Accordingly, the values actually observed by \bar{X} in repeated sampling experiments are denoted as $\bar{X}_1, \bar{X}_2, \bar{X}_3, \dots$. Similarly, for proportion \bar{p} as a random variable, with $\bar{P}_1, \bar{P}_2, \bar{P}_3, \dots$ as the unknown values it may take in repeated sampling. Once a sample has been taken and the sample proportion found, it is expressed as \bar{p} , with $\bar{P}_1, \bar{P}_2, \bar{P}_3, \dots$ as the values actually observed by \bar{p} .

Check Your Progress

3. When are we coaxed into statistical thinking mode?
4. Why are size of a sample and population important?

14.4 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Reports can be divided into different categories. The two main types of reports are:
 - Informational report
 - Interpretive report
2. The characteristics of a good report can be classified under the following four heads:
 - Language and style of the report
 - Structure of the report
 - Presentation of the report
 - References in the report
3. We are gently coaxed into the statistical thinking mode, while:
 - o Bringing out the pattern of variations in the available data in a given problem situation on one or more relevant characteristic(s), and
 - o Training the mind in comparative dimensions of data analysis, examining the consequent variations, drawing inferences, and establishing plausible relationships.
4. The size of sample and that of population are important in developing methods applied to a set of sample or population data.

14.5 SUMMARY

- A report can be defined as a written document which presents information in a specialized and concise manner.
- There is a difference between report writing and other compositions because a report is written in a short and conventional format. A report should cover all mandatory matters but nothing extra should be written.
- A report that consists of a collection of data or facts and is written in an orderly way is called an informational report.
- An interpretive report contains 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.
- Each of the above 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.
- 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 should avoid any violation of copyright laws and plagiarism.
- Being a subject of much practical utility and having wide-ranging applications, statistics displays a unique strength. It suffers from an important weakness as well.
- In developing necessary methods of data analysis and interpretation, statistics makes use of some common *notations* that have become fairly standardized. As notations are the shorthand expressions of concepts and statements, they constitute the language of statistics.
- Statistical data, as the raw material of statistics, are the observations on variables, continuous or discrete.
- In developing methods involving two variables, the second variable is, generally, represented by Y and the values that it may take by Y_1, Y_2, Y_3, \dots . Where interest lies in simultaneously dealing with three variables, Z may be used to represent the third variable with values notated as Z_1, Z_2, Z_3, \dots .
- The size of sample and that of population are important in developing methods applied to a set of sample or population data. The necessary requirement for effecting sampling is that the observations comprising the population are known and finite.

NOTES

NOTES

14.6 KEY WORDS

- **Report:** A report can be defined as a written document which presents information in a specialized and concise manner.
- **Assessment reports:** 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 reports which are made to measure the performance of the employees in an organization for different purposes like appraisal or promotion are called performance reports.

14.7 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the different types of informational reports?
2. Comment on the structure of a report.
3. How should the references in a report be handled?
4. Write a short note on the use of punctuations in a report.

Long-Answer Questions

1. Write a descriptive note on the different types of reports.
2. Analyse the characteristics of a good report.
3. Describe the mechanics of writing a report.
4. How is data analysed? Discuss.

14.8 FURTHER READINGS

Shajahan, S. 2004. *Research Methods for Management*. Mumbai: Jaico Publishing House.

Kothari, C.R. 2004. *Research Methodology: Methods and Techniques*, 2nd ed. New Delhi: New Age International Publishers.

Ahuja, R. 2001. *Research Methods*. New Delhi: Rawat Publications.

Sharma, K.R. 2002. *Research Methodology*. Jaipur: National Publishing House.