RESEARCH METHODOLOGY

MBA First Year
Semester-II, Paper-IV

Lesson Writers

Dr. G. Malathi
Assistant Professor,
Dept of Management,
Central University of AP

Dr. K. Naga Sundari
Head,
Dept of MBA,
Maris Stella College, Vijayawada

Dr. G. Ranganath
Associate Professor,
Central University of Karnataka

Dr. S. AnithaDevi
Professor,
TJPS College, Guntur

Editor
Prof. R. Siva Ram Prasad
Dean, Faculty of Commerce & Management Studies,
Acharya Nagarjuna University

Academic Advisor
Prof. R. Siva Ram Prasad
Dean, Faculty of Commerce & Management Studies,
Acharya Nagarjuna University

DIRECTOR, I/c.
Prof. V. VENKATESWARLU
M.A., M.P.S., M.S.W., M.Phil., Ph.D.
CENTRE FOR DISTANCE EDUCATION
ACHARYA NAGARJUNA UNIVERSITY
NAGARJUNA NAGAR 522 510
Ph: 0863-2346222, 2346208
0863- 2346259 (Study Material)
Website www.anucde.info
E-mail: anucdedirector@gmail.com

MBA : RESEARCH METHODOLOGY

First Edition : 2025

No. of Copies :

This book is exclusively prepared for the use of students of MBA, Centre for Distance Education, Acharya Nagarjuna University and this book is meant for limited circulation only.

Published by:

Prof. V. VENKATESWARLU
Director, I/c
Centre for Distance Education,
Acharya Nagarjuna University

Printed at:

FOREWORD

Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A+' grade from the NAAC in the year 2024, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 221 affiliated colleges spread over the two districts of Guntur and Prakasam.

The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.Sc., B.A., B.B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.

To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.

It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson-writers of the Centre who have helped in these endeavors.

Prof. K. Gangadhara Rao
M.Tech., Ph.D.,
Vice-Chancellor I/c
Acharya Nagarjuna University.

M.B.A - GENERAL
206EM24: RESEARCH METHODOLOGY
SYLLABUS

COURSE LEARNING OUTCOMES (CLOs):
On successful completion of the course the learner will be able to:

- To impart knowledge for enabling students to develop data analytics skills and
- meaningful interpretation to the data sets so as to solve the Business/Research problem.
- Develop an understanding on various kinds of Research, Objectives of doing research,
- Research process, Research designs and Sampling.
- To provide basic knowledge on qualitative research techniques
- To provide adequate knowledge on Measurement & Scaling techniques as well as
- quantitative data analysis.
- To familiarize students with Statistical packages such as SPSS.

Unit-I: Foundations of Research: Meaning, types of research, Research Process. Nature and Scope of Research Methodology - Problem Formulation, Research Objectives - Hypotheses, Characteristics of good hypotheses, Research Design - Types of Research Design.

Unit-II: Variable Types: lndependent & Dependent variables Exogenous & Endogenous Variables, Qualitative and Quantitative Research: Qualitative research- Quantitative research- Concept of measurement, Validity and Reliability. Levels of measurement, Nominal, Ordinal, Interval. Ratio.

Unit-III: Sampling: Characteristics of a good sample, Types of sampling- Probability Sampling Types, Non-Probability Sampling Types, Determining size of the sample, Primary and Secondary Sources-Methods of Data Collection- Primary data collection techniques & Secondary data Collection, Questionnaire Design

Unit -IV: Multivariate Data analysis: ANOVA one way /two way, Multiple Correlation & Multiple Regression, 'Discriminant Analysis, Factor Analysis, Conjoint Analysis, Multidimensional Scaling and Clustering Techniques.

Unit-V: Automated Data Analysis Using SPSS: Data Preparation- Univariate analysis (frequency tables, bar charts, pie charts, percentages, Bivariate analysis- Cross tabulations and Chi-square test including testing hypothesis of association. Performing T several t -Tests, ANOVA, Correlation, Regression in SPSS

Reference Books:

1. l. CT Kothari & Gaurav Garg, Research Methodology, New Age International Publishers,
2. New Delhi.
3. Cooper, "Business Research Methods"o Tata McGraw Hill, New Delhi, 2010.
4. C.R.Kothari, "Research Methodology: Methods and Techniques", New Age International
5. Publishers, New Delhi, 2006.
6. 4. Gupta S.P. "statistical Methods", Sultan Chand, New Delhi, 2010.
7. 5. f.S. Witt<inson & P.L. Bhandarkar, "Methodology and Techniques of Social Research".
8. 6. Richard A.Johnson & Dean W.Wichern, "Applied Multivariate Statistical Analysis",
9. Prentice Hall International Inc., 2007.

206EM24

MODEL QUESTION PAPER
M.B.A

RESEARCH METHODOLOGY

Time: 3hours. Max. Marks 70 Marks

SECTION A

(5 × 3 = 15 Marks)

1. Explain the various types of research designs
2. What is the formulation of a research problem?
3. Differentiate between independent and dependent variables.
4. Explain the applications of discriminant analysis
5. Describe the sampling frame.
6. Explain ANOVA
7. Explain the characteristics of a good sample.
8. Explain the applications of SPSS.
9. Types of hypotheses in research
10. Describe primary data.

SECTION B

(5 × 10 = 50 Marks)

Answer ALL questions

11. (a) Describe the different kinds of research using examples.

(OR)

(b) Discuss the steps in the research process.

12. (a) Describe reliability and validity in research.

(OR)

(b) Describe the nominal, ordinal, interval, and ratio levels of measurement.

13. (a) Describe both probability and non-probability sampling methods.

(OR)

(b) Explain the various primary and secondary data collection techniques.

14. (a) A researcher wants to examine how Sales Training Hours (X_1) and Advertising Expenditure (X_2) influence Monthly Sales (Y). The following data for 8 salespersons is given below. Find the multiple correlation coefficient $R_{3.12}$ between Y and X_1 and X_2

Y: Monthly Sales (₹000)	55	62	58	70	75	50	80	45
X_1: Training Hours	5	6	5	7	8	4	9	3
X_2: Advertising Spend (₹000)	20	25	22	30	35	18	40	15

(OR)

(b) Describe factor analysis and its applications

15. (a) Describe how to perform a t-test in SPSS.

(OR)

(b) Explain the Chi-Square tests and cross-tabulations with examples of how to interpret them.

SECTION C – CASE STUDY

(1 × 15 = 15 Marks)

A researcher aims to examine the preferred shopping method among university students (Online shopping versus Offline shopping). The researcher gathers data from 120 students via a straightforward questionnaire.

The questionnaire includes items on:

- Convenience
- Price
- Variety of products
- Satisfaction level
- Frequency of shopping

The researcher aims to compare online and offline shoppers and determine which factors affect their preferences.

Questions:

a) Identify the type of research conducted.
b) Suggest an appropriate sampling technique for selecting 120 students.
c) Identify the independent and dependent variables in this study.
d) Mention suitable measurement scales for convenience and satisfaction.
e) Suggest any two statistical tools (like t-test, charts, percentages) that can be used to analyse the collected data.
f) How can the researcher present the results using SPSS?

CONTENTS

LESSON-1

RESEARCH METHODOLOGY: AN OVERVIEW

OBJECTIVES OF THE LESSON

After reading this chapter, a learner will be able to:

1. Understand the Foundations of Research. Define Research Clearly
2. Recognize the Characteristics of Research
3. Explain the Objectives of Research. Differentiate Types of Research
4. Understand the Entire Research Process and the Role of Research in Library & Information Science.

STRUCTURE OF THE LESSON

1.1 FOUNDATION/ INTRODUCTION OF RESEARCH

All societies from the primitive to the most modern sophisticated societies have progressed only on the acquisition of knowledge and its application, depending upon their capability to understand their environments and control them through concerted efforts. Initially knowledge acquisition was more on the basis of observation, experience, learning by trial and error, simple logics of deduction and inference, etc. But with the increasing ability to conduct research and getting positive results and the ability to apply them in solving problems, although confined to a few individuals, human societies were slowly advancing materially.

With science and technology opening up new directions of growth and development from the 15th century in Western Europe and its influence in other parts of the world, methods of research have become a mode of acquiring knowledge through scientific methods. It was largely an individual flair that pushed up the frontiers of knowledge albeit with very limited facilities for research. With the advent of universities, research became one of their important functions, besides their teaching, training, and publications functions. Increasing pursuit of research has resulted in the growth of a body of literature over the years on research methodology, which has now developed into a subject in its own right.

In the course of time, institutions, associations and cognate bodies, have been established to deal with various development problems through research, with financial aids from

governments and industry. Today there are research institutions, which have been set up to deal exclusively with research in different subjects, including library and information science. In this Chapter, we are trying, in a general way, to study the subject of research methodology in all its dimensions. Formal definitions of research, need to pursue research to expand the horizons of knowledge, contours of research processes with an understanding of the conceptual framework model of research methodology, characteristics of research, scientific research, research design and other related aspects are discussed in this Unit. Another important point to be noted in a study of research methodology by students of library and information science is not only to get the necessary skills in doing research in their own field but also to be of assistance and help to the research community offering high quality information service. This aspect is also elaborated in this Chapter.

There are likely to be some overlapping of ideas in discussing these aspects in the different sections of this unit. They are reiterations and should be understood in the contexts in which each of these ideas is discussed.

1.2 DEFINITION OF RESEARCH

Webster's Third International Dictionary of the English Language defines research as *"Studious inquiry or examination, especially critical and exhaustive investigation or experimentation, having for its aim the discovery of new facts, and their correct interpretation, the revision of accepted conclusions, theories, or laws in the light of newly discovered facts, or practical applications of new or revised conclusions, theories, or laws."*

According to the Random House Dictionary of the English Language, *Research is a systematic inquiry into a subject in order to discover or revise facts, theories, etc.*

In the Encyclopedia of Social Sciences, Research is defined as *"the manipulation of things, concepts or symbols for the purpose of generalization to extend, correct or verify knowledge whether that knowledge aids in the construction of a theory or in practice of an art."*

Best and Kahn, in their book Research in Education define research *"as the systematic and objective analysis and recording of controlled observations that may lead to the development of generalization, principles or theories, resulting in prediction and possibly ultimate control of events."*

Busha in his publication Research Methods in Librarianship says that Research is *"a systematic quest for knowledge that is characterized by disciplined enquiry. Efficient and effective approach to expand knowledge is the conduct of special, planned and structured investigations."*

Cook outlines research as an honest, exhaustive, intelligent searching for facts and their meanings or implications, with reference to a problem.

He sees the word 'Research' as an acronym, each letter of the word, standing for a particular aspect as given below:

R = Rational way of thinking
E = Expert and Exhaustive treatment
S = Search and solution
E = Exactness

A = Analysis
R = Relationship of facts
C = Critical observation, Careful planning, Constructive attitude and Condensed generalization
H = Honesty and Hard working

Ranganathan *describes research to represent a critical and exhaustive investigation to discover new facts, to interpret them in the light of known ideas, theories and laws, to revive the current laws and theories in the light of the newly discovered facts to apply the conclusion to practical purpose.*

1.3 NATURE OR CHARACTERISTICS OF RESEARCH

Research is a process of collecting, analyzing and interpreting information to answer questions. But to qualify as research, the process must have certain characteristics: it must, as far as possible, be controlled, rigorous, systematic, valid and verifiable, empirical and critical.

- **Controlled** - in real life there are many factors that affect an outcome. The concept of control implies that, in exploring causality in relation to two variables (factors), you set up your study in a way that minimizes the effects of other factors affecting the relationship. This can be achieved to a large extent in the physical sciences (cookery, bakery), as most of the research is done in a laboratory. However, in the social sciences it is extremely difficult as research is carried out on issues related to human beings living in society, where such controls are not possible. Therefore, as you cannot control external factors, you attempt to quantify their impact.

- **Valid and verifiable** - this concept implies that whatever you conclude on the basis of your findings is correct and can be verified by you and others.

- **Empirical** - this means that any conclusions drawn are based upon hard evidence gathered from information collected from real life experiences or observations.

- **Critical** - critical scrutiny of the procedures used and the methods employed is crucial to a research enquiry. The process of investigation must be foolproof and free from drawbacks. The process adopted and the procedures used must be able to withstand critical scrutiny.

- **Rigorous** - you must be scrupulous in ensuring that the procedures followed to find answers to questions are *relevant, appropriate and justified.* Again, the degree of rigor varies markedly between the physical and social sciences and within the social sciences.

- **Systematic** - this implies that the procedure adopted to undertake an investigation follow a certain logical sequence. The different steps cannot be taken in a haphazard way. Some procedures must follow others.

1.4 OBJECTIVES OF RESEARCH

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

1. To gain familiarity with a phenomenon or to achieve new insights into it (Studies with this object in view are termed as exploratory or formulative research studies);

2. To portray accurately the characteristics of a particular individual, situation or a group (Studies with this object in view are known as descriptive research studies);

3. To determine the frequency with which something occurs or with which it is associated with something else (Studies with this object in view are known as diagnostic research studies);

4. To test a hypothesis of a causal relationship between variables (Such studies are known as hypothesis-testing research studies)

1.5 TYPES OF RESEARCH

Research can be classified into various categories depending on the perspective under which the research activity is initiated and conducted. The categorization depends on the following perspectives in general:

1. **Application of research study**
2. **Objectives in undertaking the research**
3. **Inquiry mode employed for research**

1. **Classification based on Application:**
    - **Pure / Basic / Fundamental Research:** As the term suggests a research activity taken up to look into some aspects of a problem or an issue for the first time is termed as basic or pure. It involves developing and testing theories and hypotheses that are intellectually challenging to the researcher but may or may not have practical application at the present time or in the future. The knowledge produced through pure research is sought in order to add to the existing body of research methods. Pure research is theoretical but has a universal nature. It is more focused on creating scientific knowledge and predictions for further studies.
    - **Applied / Decisional Research: Applied** research is done on the basis of pure or fundamental research to solve specific, practical questions; for policy formulation, administration and understanding of a phenomenon. It can be exploratory, but is usually descriptive. The purpose of doing such research is to find solutions to an immediate issue, solving a particular problem, developing new technology and look into future advancements etc. This involves forecasting and assumes that the variables shall not change.

Key Differences between Basic and Applied Research

a. Basic Research can be explained as research that tries to expand the already existing scientific knowledge base. On the contrary, applied research is used to mean the scientific study that is helpful in solving real-life problems.

b. While basic research is purely theoretical, applied research has a practical approach.

c. The applicability of basic research is greater than the applied research, in the sense that the former is universally applicable whereas the latter can be applied only to the specific problem, for which it was carried out.

d. The primary concern of the basic research is to develop scientific knowledge and predictions. On the other hand, applied research stresses on the development of technology and technique with the help of basic science.

e. The fundamental goal of the basic research is to add some knowledge to the already existing one. Conversely, applied research is directed towards finding a solution to the problem under consideration.

2. Classification based on Objectives:

- **Descriptive Research:** This attempts to explain a situation, problem, phenomenon, service or program, or provides information viz. living condition of a community, or describes attitudes towards an issue but this is done systematically. It is used to answer questions of who, what, when, where, and how associated with a particular research question or problem. This type of research makes an attempt to collect any information that can be expressed in quantifiable terms that can be used to statistically analyze a target audience or a particular subject. Descriptive research is used to observe and describe a research subject or problem without influencing or manipulating the variables in any way. Thus, such studies are usually correlation or observational. This type of research is conclusive in nature, rather than inquisitive. E.g. explaining details of budget allocation changes to departmental heads in a meeting to assure clarity and understanding for reasons to bring in a change.

- **Co relational Research:** This is a type of non-experimental research method, in which a researcher measures two variables, understands and assesses the statistical relationship between them with no influence from any extraneous variable. This is undertaken to discover or establish the existence of a relationship/ interdependence between two or more aspects of a situation. For example, the mind can memorize the bell of an ice cream seller or sugar candy vendor. Louder the bell sound, closer is the vendor to us. We draw this inference based on our memory and the taste of these delicious food items. This is specifically what co relational research is, establishing a relationship between two variables, ―bell sound‖ and ―distance of the vendor‖ in this particular example. Co relational research is looking for variables that seem to interact with each other so that when you see one variable changing, you have a fair idea how the other variable will change.

- **Explanatory:** is the research whose primary purpose is to explain why events occur, to build, elaborate, extend or test a theory. It is more concerned with showcasing, explaining and presenting what we already have. It is the process of turning over 100 rocks to find perhaps 1 or 2 precious gemstones. Explanatory survey research may look into the factors that contribute to customer satisfaction and determine the relative weight of each factor, or seek to model the variables that lead to people shifting to departmental stores from small shops from where they have been making purchases till now. An exploratory survey posted to a social networking site may uncover the fact that an organizations customers are unhappy thus helping the organization take up necessary corrective measures.

- **Exploratory Research**: Exploration has been the human kinds passion since the time immemorial. Looking out for new things, new destinations, new food, and new cultures has been the basis of most tourist and travel journeys. In the subjective terms exploratory research is conducted to find a solution for a problem that has not been studied more clearly, intended to establish priorities, develop operational definitions and improve the final research design. Exploratory research helps determine the best research design, data-collection method and selection of subjects. For such research, a researcher starts with a general idea and uses this research as a medium to identify issues that can be the hub for future research. An important aspect here is that the researcher should be willing to change his/her direction subject to the revelation of new data or insight. Such research is usually carried out when the problem is at a beginning stage. It is often referred to as grounded theory approach or interpretive research as it used to answer questions like what, why and how. For example: a fast-food outlet owner feels that increasing the variety of snacks will enable increase in sales, however he is not sure and needs more information. Thus, the owner starts studying local competition, talks to the existing customers, friends etc. to find out what are their views about the current menu and what else do they wish to be included in the menu and also assess whether he would be able to generate higher revenues.

3. Classification based on Inquiry Mode:

- **Structured approach:** The structured approach to inquiry is usually classified as quantitative research. Here everything that forms the research process- objectives, design, sample, and the questions that you plan to ask of respondents- is predetermined. It is more appropriate to determine the extent of a problem, issue or phenomenon by quantifying the variation e.g. how many people have a particular problem? How many people hold a particular attitude? E.g. asking a guest to give feedback about the dishes served in a restaurant.

- **Unstructured approach:** The unstructured approach to inquiry is usually classified as qualitative research. This approach allows flexibility in all aspects of the research process. It is more appropriate to explore the nature of a problem, issue or phenomenon without quantifying it. Main objective is to describe the variation in a phenomenon, situation or attitude e.g., description

of an observed situation, the historical enumeration of events, an account of different opinions different people has about an issue, description of working condition in a particular industry. E.g. when guest is complaining about the room not being comfortable and is demanding a discount the staff has to verify the claims empathically.

In many studies you have to combine both qualitative and quantitative approaches. For example, suppose you have to find the types of cuisine / accommodation available in a city and the extent of their popularity. Types of cuisine are the qualitative aspect of the study as finding out about them entails description of the culture and cuisine. The extent of their popularity is the quantitative aspect as it involves estimating the number of people who visit restaurant serving such cuisine and calculating the other indicators that reflect the extent of popularity.

4. Other Types of Research:

- **Descriptive v/s Analytical:** Descriptive research includes surveys and factfinding enquiries of different kinds. The major purpose of descriptive research is description of the state of affairs as it exists at any given time. The term Ex post facto research is used in social sciences and business research for descriptive research studies. The researcher only reports about the factors identified and cannot modify the details available thus it makes it clear that he does not have any control over such variables Most ex post facto research projects are used for descriptive studies in which the researcher strives to find out information about, for example, frequency of dining out, preferences of individuals, etc. Ex post facto studies also include attempts by researchers to discover causes even when they cannot control the variables. The methods of research utilized in descriptive research are survey methods of all kinds, including comparative and co relational methods. In analytical research, on the other hand, the researcher has to use facts or information already available, and analyze these to make a critical evaluation of the material.

- **Applied v/s Fundamental:** Research can either be applied (or action) research or fundamental (to basic or pure) research. Applied research aims at finding a solution for an immediate problem facing a society or an industrial/business organization, whereas fundamental research is mainly concerned with generalizations and with the formulation of a theory.

- "Gathering knowledge for knowledges sake is termed pure or basic research". Research concerning some natural phenomenon or relating to pure mathematics are examples of fundamental research. Similarly, research studies, concerning human behavior carried on with a view to make generalizations about human behavior, are also examples of fundamental research, but research aimed at certain conclusions (say, a solution) facing a concrete social or business problem is an example of applied research. Research to identify social, economic or political trends that may affect a

particular institution or the copy research (research to find out whether certain communications will be read and understood) or the marketing research or evaluation research are examples of applied research. Thus, the central aim of applied research is to discover a solution for some pressing practical problem, whereas basic research is directed towards finding information that has a broad base of applications and thus, adds to the already existing organized body of scientific knowledge.

- **Quantitative v/s Qualitative:** Quantitative research is based on the measurement of quantity or amount. It is applicable to phenomena that can be expressed in terms of quantity. E.g. Studying the number of enquiries received for room bookings through different modes like internet, emails, calls, letters, or different sources like travel and tours operators, companies and government organizations etc.

- **Qualitative research,** on the other hand, is concerned with qualitative phenomenon, i.e., phenomena relating to or involving quality or kind. E.g. studying the stress levels and reasons for variable performances of staff in different shifts in the same department of a hotel. The same individuals may perform differently with the change of shift timings. It can involve performing research about changing preferences of customers as per the change of season.

- Another example is attitude or opinion research i.e. research intended to find out how people feel or what they think about a particular subject or institution is also qualitative research. Through behavioral research we can evaluate the diverse factors which motivate people to behave in a particular manner or which make people like or dislike a particular thing. It is therefore important that to be relevant in qualitative research in practice the researcher should seek guidance from qualified individuals from the field opted.

- **Conceptual vs. Empirical:** Conceptual research is associated to some theoretical idea(s) or presupposition and is generally used by philosophers and thinkers to develop new concepts or to get a better understanding of an existing concept in practice. On the other hand, **Empirical research** draws together the data based on experience or observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. It is also known as experimental research as it is essential to get facts firsthand, at their source, and actively to go about doing certain things to stimulate the production of desired information. Here the researcher develops a hypothesis and assimilates certain outcomes to start with followed by efforts to get adequate facts (data) to prove or disprove his hypothesis. An experimental design is then developed based on variables that can modify or concur the results to prove that he has given a valid statement. This also affirms that he has a reasonable control over the variables and can get different results by giving different values to them. Empirical research is appropriate

when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies is today considered to be the most powerful support possible for a given hypothesis.

1.6 RESEARCH PROCESS

Research, whether natural or social, involves a systematic process that focuses on objectively collecting information to fulfil well designed objectives. Regardless of type and theme or research (for example, experimental, evaluation or action research) specific steps are followed in the process of research, which are well- defined. Each step of the research leads to another, which makes the entire research process scientific and reliable. Any other individual who following the same steps repeats the study, would find similar results. This is referred to as replicating the study.

The research process is the paradigm of research project. It is a journey being undertaken by all the researchers in a zest to find answers to the research questions/ objectives. The journey begins with the formulation of the research problem and ends with finding answers. The path to finding answers to the research questions constitutes research methodology. In order to cover this journey in an objective manner definite steps have been devised for all the researchers. The sequence of these steps by and large remains same, though certain variations can be exercised by the researcher based on the requirements of the study and their experience. The important steps involved in the process of social research includes nine steps which are as follows:

Formulation of the Research Problem

↓

Literature Review

↓

Developing Research Questions and Formulating Objectives

↓

Conceptualizing the Research Design

↓

Constructing Hypothesis

↓

Universe and Sampling Designs

↓

Constructing the Instrument of Data Collection

↓

Data Processing and Analysis

↓

Writing Research Report

1.7 SUMMARY:

This chapter introduces the fundamental concepts of research and its methodology. It explains how human societies evolved through systematic knowledge creation, observation, and scientific inquiry. With the development of universities and research institutions, research has become a structured and essential activity. The chapter provides multiple definitions of research, emphasizing that it is a systematic, scientific, and objective pursuit of knowledge. Research must be controlled, rigorous, systematic, empirical, valid, and critical to qualify as credible.

Research objectives are classified into exploratory, descriptive, diagnostic, and hypothesis-testing categories. The chapter also details different types of research based on application (basic vs applied), objectives (descriptive, correlational, explanatory, exploratory), and inquiry mode (quantitative vs qualitative; structured vs unstructured). Additional distinctions include conceptual vs empirical research and descriptive vs analytical research. Finally, the chapter outlines a nine-step research process: formulating the problem, reviewing literature, setting research questions, designing the research, framing hypotheses, sampling, constructing data collection tools, processing/analyzing data, and writing the report. The chapter emphasizes that research methodology is essential not only for conducting research but also for supporting researchers through quality information services

1.8 TECHNICAL TERMS

- **Research** – A systematic and scientific process of investigating a problem to generate new knowledge.
- **Validity** – The extent to which a research tool measures what it is intended to measure.
- **Reliability** – The consistency or repeatability of research results when tested multiple times.
- **Exploratory Research** – Investigation done to gain new insights into an unclear or poorly understood problem.
- **Descriptive Research** – Research that systematically describes characteristics of people, events, or situations.
- **Diagnostic Research** – Study aimed at determining the frequency or reasons behind a problem or phenomenon.
- **Hypothesis-testing Research** – Research designed to test the relationship between variables through hypotheses.

1.9 SELF- ASSESSMENT QUESTIONS

Multiple Choice Questions (MCQs)

1. Research can be defined as:
a) Random observation
b) Studious inquiry for discovering new facts
c) Personal opinion
d) Unstructured guessing
2. The primary purpose of exploratory research is to:
a) Test hypotheses
b) Discover insights into a phenomenon
c) Describe a population
d) Measure frequency
3. Which of the following is NOT a characteristic of research?
a) Systematic
b) Critical
c) Controlled
d) Imaginary
4. Applied research aims at:
a) Expanding general knowledge
b) Solving immediate practical problems
c) Developing pure theories
d) Only descriptive analysis
5. Empirical research is based on:
a) Hypothetical ideas only
b) Experiences and observations
c) Opinions of experts only
d) Theories without validation

SHORT - ANSWER QUESTIONS

1. Define research in your own words.
2. Differentiate between basic and applied research.
3. What are the key characteristics of scientific research?
4. What is exploratory research?
5. Explain the difference between quantitative and qualitative research.
6. What are the main steps involved in the research process?

LONG ANSWER - QUESTIONS

1. How did early societies acquire knowledge before formal research methods developed?
2. Define research and explain the characteristics of Research?
3. What is the main aim of research?

4. What is the aim of descriptive research studies?
5. How do basic and applied research differ in purpose and approach?
6. What are the main characteristics of qualitative and quantitative research?
7. What are the key steps in the research process?

CASE STUDY

Case Study: Research in a Public Library

A public library is experiencing a decline in membership over the past two years. The librarian wants to understand why users are not renewing memberships and how services can be improved. The librarian has never conducted research before but follows the systematic process described in the chapter.

Tasks:

1. **Identify the research problem.**
2. **Suggest whether this study is basic or applied research and why.**
3. **Recommend whether a quantitative, qualitative, or mixed-method approach should be used.**
4. **Develop one possible hypothesis.**
5. **State which data collection tools would be appropriate.**

REFERENCES

1. Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Age International Publishers.
2. Best, J. W., & Kahn, J. V. (2006). *Research in Education* (10th ed.). Pearson Education
3. Creswell, J. W. (2014). Research Design: Qualitative, Quantitative and Mixed Methods Approaches (4th ed.). Thousand Oaks, CA: Sage

1.10 SUGGESTED READINGS

1. Kothari, C.R. 2004. Research Methodology Methods and Techniques, New Age International (P) Limited: New Delhi.
2. Rao K.V. 1993. Research Methodology in Commerce and Management, Sterling Publishers Private Limited: New Delhi.
3. Sadhu, A.N. and A. Singh, 1980. Research Methodology in Social Sciences, Sterling Publishers Private Limited: New Delhi.

RANGANATHAM GANGINENI

LESSON-2

RESEARCH METHOD AND METHODOLOGY

OBJECTIVES OF THIS LESSON

After studying lesson, students will be able to:

1. Differentiate between research methods and research methodology. Explain the logic, rationale, and significance of choosing specific research methods
2. Identify various non-experimental research methods. Understand experimental methods
3. Recognize how data collection, analysis, tools, and procedures link to methodology.
4. Explain the philosophical, theoretical, and strategic foundations of research methodology. valuate the credibility, validity, and dependability of a research study and justify the selection of appropriate methods. Describe the differences between methods and methodology

STRUCTURE OF THE LESSON

2.1 RESEARCH METHOD

The tactics, procedures, or techniques used in collecting data or evidence for analysis to reveal new knowledge or generate a better understanding of a topic are referred to as research methods. The term "research methods" refers to all of the researcher's processes undergoing when investigating his study issue. As a result, techniques will be regarded as the heart of research methodology. There are several research procedures, each of which employs a unique set of data gathering instruments. So, when preparing your approaches, you will need to make critical judgments over different methods used in the research. For instance, how do you collect data (Qualitative vs. quantitative) and how do you analyze the data (Descriptive vs. experimental). Thus, research methodology isn't simply about the research methods, but it is to consider the logic behind using the methods. Researchers need to identify the logic and rationality of using one approach over another.

2.2 RESEARCH METHODOLOGY

The term 'Methodology' is frequently used in research projects, and it gives a complete overview of the rationality of the research. Research Methodology refers how scientific research is carried out. It involves rationally adopting several methods to tackle research challenges in a systematic manner. Methodology aids in comprehending not just the results of scientific investigation, but also the method itself. In a broader way, methodology entails searching/linking methods and studying specific theories used in research field areas, designing the best strategy according to research goals and objectives.

Practically say 'research methodology' answer "how" of any piece of research is related to the research technique and how a researcher plans a study in a structured way to ensure reliable outcomes that address the study's goals and objectives. Apart from that, research methodology should explain the design decisions by illustrating that the approaches and techniques chosen match the study objectives and goals and provide valid and accurate results. A suitable research methodology yields scientifically solid results, whereas a lousy methodology yields none. Hence, it should highlight some basic (What, How, Why) questions of our research. Such as:

1. What was the purpose of your research?
2. What types of 'research method' can we use and why?
3. What types of data should we consider for your research analysis purpose?
4. What were the data collecting methods?
5. How did you analyze the collected data?
6. What kind of resources has been used in your research?

In addition, methodologies encompass all approaches, strategies, and instruments used by a researcher to finish the experiment and solve the study challenge. While methods display the entire study framework, explaining its various research design components. After all, 'methodology' enables researchers to verify a study's accuracy and helps them identify the research method's credibility, validity, and dependability. Another way, we can say that methodology will justify choosing the proper methods. On the other side,' research methods' give detailed information on the research design, participants, materials, equipment, variables, and processes.

2.3 DIFFERENCE BETWEEN RESEARCH METHOD AND METHODOLOGY

Key Differences between Research Method and Research Methodology

1. Research method focuses on the techniques and tools used to gather data, while research methodology deals with the overall strategy and framework of research.
2. Research method involves specific procedures and steps to obtain information,

whereas research methodology guides the researcher in selecting appropriate research methods.

3. Research method determines the type of data to be collected and analyzed, while research methodology provides a systematic approach to conducting research.
4. Examples of research methods include surveys, experiments, interviews, and observations, while research methodology encompasses qualitative, quantitative, and mixed methods.
5. Research method describes the process of data collection and analysis, whereas research methodology relates to the theoretical and philosophical underpinnings of research.
6. Research method is more concrete and tangible, focusing on the practical aspects of research, while research methodology is more abstract and conceptual, focusing on the theoretical aspects.
7. Research method is a subset of research methodology, which encompasses the entire research process.
8. Research method is specific to a particular research project or study, while research methodology is applicable across different research projects and studies.
9. Research method determines the reliability and validity of research findings, whereas research methodology determines the overall validity and soundness of research.
10. Research method is concerned with the tools and techniques employed, while research methodology focuses on the framework and approach utilized in the research process.

2.4 METHODS OF RESEARCH

Methods of research can be classified into two categories: Non-experimental methods and experimental methods.

2.4.1 Non-Experimental Methods

2.4.1.1 Naturalistic Observation

Sometimes all researchers need to know is what is happening to a group of animals or people. The best way to look at his behaviour of animals or people is to watch them behave in their normal environment. In naturalistic observation a scientist observes behaviour in real world settings and makes no effort to manipulate or control the situation. Researchers conduct naturalistic observation at homes, day-care centers and so on. For example, if someone wanted to know how adolescents behave with members of the opposite sex in a social setting the researcher might go the mall on a weekend night.

The most important advantage of naturalistic observation is that it allows researchers to get a realistic picture of how behaviour occurs because they are actually watching that behaviour. In many cases animals or people who know they are being watched will not behave normally anyway in a process called the observer effect so often the observer needs to remain hidden from view. In these cases, researcher might use one way mirror, or they might actually become participant in the group. This technique is called participant observation.

One of the major disadvantages of the naturalistic observation is the possibility of observer bias. That happens when the person doing the observing has a particular opinion about what he or she is going to see or expects to see. Sometimes that person sees only those actions that supports that expectation and ignores actions that don't fit.

Another disadvantage is that each naturalistic setting is unique and unlike any other. Observations that are made at one time in one setting may not hold true for another time even if the setting is similar because the conditions are not going to be exactly the same time after time, researchers don't have that kind of control over the natural world.

2.4.1.2 Archival Research

In this method the researchers do not actually collect data themselves but they obtain data from public records, archives and so on. The researches merely analyse the data attempts to draw certain conclusions from them. The method can be valuable in many respects. For instance, there is no other way to collect data on suicides and homicides. Archival Data are those that are present in existing records or archives. The researcher simply examines or selects the data for analysis.

Archival research may already exist or logistics or ethics may make it infeasible to conduct an experiment relating the variables of interest.

Archival research has limitations; First most archival data are collected for naturalistic reasons. Governments are private agencies collect the data for their own purpose and such data often do not suit the purposes of the scientist. Second because archival research is by nature carried out after the fact ruling out alternative hypotheses for particular observed correlations may be difficult. A researcher who relies on archival data is at the mercy of any biases that may have occurred in collecting the data. Police records are notoriously subject to bias. Many categories of crime are seldom reported to the police.

2.4.1.3 Content Analysis

Content analysis sometimes known as document analysis is a method of systematic, examination of communications or of current records or documents. Instead of questioning respondents according to some scale items or observing their behaviour directly the content – analyzer takes the communications or documents prepared by the respondents and systematically find out the frequency or proportion of their appearances.

In content or documents analysis the primary sources of data are: letters, autobiographies, diaries, compositions, records, reports, printed forms, themes or other academic work, books, periodicals, bulletins or catalogues, syllabus, court decisions, pictures, films, cartoons etc. It is the obligation of the researchers to establish the trustworthiness of these data that have been drawn. Content analysis can also be used with responses of projective test with all kinds of verbal materials and with materials specially produced for research problems.

Merits and Demerits

- First content analysis is applicable to a wide variety of materials such as creativity, attitude, and ethnocentrisms, stereotypes, curriculum changes values, interest, religiosity, college budgets etc.
- Second content analysis can also be used to examine the effect of experimental manipulation upon the dependent variables. If the investigator wants to study the effect of practice upon the improvement of handwriting of children, content analysis may be of no less importance than any experimental design.
- Third content analysis is also used to validate other methods of observation. Suppose one wants to validate a self-disclosure inventory. It is expected that people in general would not like to give personal information against which the test can be validated. But subjects can be asked some projective-type of questions and the responses can be content-analyzed. Subsequently the test can be validated against the content- analyzed response.
- Despite these merits content analysis should be used with caution because of the complexities involved.

2.4.2 Surveys

Survey methods are widely used gathering scientific information. It involves collection of data by asking questions and recording people's answers to them. They are used for various purposes on frequent goal of this kind of research is to estimate population characteristics. For example, the goal of survey might be to determine the percentage of people who hold supporting or opposing positions on particular social issues, such as provision of reservation for women in job. The census and public opinion done by various agencies are good examples of surveys.

Surveys can also be used to test hypotheses about the relationships among variable. One may try to find out the effect of some event on people's behaviour. For example, surveys have been conducted after the earth quack at Bhuj in Gujarat to find out the impact of earthquake on people's lives.

In undertaking surveys, the researcher defines the study population and draws the sample. The sample must be representative of the population. Researcher use different procedures of sampling. They can use random sampling in which every member of the population has an equal and independent chance of being included in the sample. Usually, the researcher use stratified random sampling in which two or more sub samples are represented according to some predetermined proportion as they exist in the population. Sometimes groups are selected by using clusters or groupings from a larger population. This is known as cluster sampling. The sample size is also determined because the ability to generalize depends on the sample size used in the survey.

Depending upon the ways of collecting data survey methods can be classified into different categories namely personal interview, mail questionnaire, telephone survey, internet survey, web survey, etc.

Advantages:

Survey methods have wide scope. In other words, through survey method a great deal of information can be obtained by studying the larger population

It is more accurate. As Kerlinger (1986) has put it." The accuracy of properly drawn samples is frequently surprising, even to experts in the field. A sample of 600 to 700 individuals or families can give a remarkably accurate portrait of a community its values attitudes and beliefs.

Survey methods have been frequently used in almost all the social sciences. Hence the method has inter-disciplinary value. In fact, such researches provide raw materials for vast increasing "gross disciplinary research" (Cambell & Katona,1953).

Survey method is considered a very important and indispensable tool for studying social attitudes, beliefs, values etc. with accuracy at the economic rate.

Survey methods remain at the surface and it does not penetrate into the depth of the problem being investigated.

Survey method is time consuming, and demand a good amount of expenditure.

Although it is true that survey research is accurate, it is still subject to sampling errors. In survey research there is always the probability of one chance in a twenty or hundred with an error, more serious than minor fluctuation of a chance, may occur and distort the validity of the result obtained.

Survey method demands expertise, research knowledge and sophistication on the part of the researcher. In other words, the researcher must know the techniques of sampling, questionnaire construction, interviewing and analysis of data.

2.4.3 Field Studies

Field studies are ex-post scientific inquiries aimed at discovering the relations and interactions among sociological, psychological and educational variables in real social structures. In scientific studies, large or small, they systematically pursue relations and test hypotheses, that are ex-post facto, that are made in actual life situations, will be considered field ex-post factor, that are made in actual life situations, will be considered field studies. The investigator in a field stud looks at the social or institutional situation and then studies the relations among the attitudes, values, perceptions, and behaviours of individuals and groups in the situation. He ordinarily manipulates no independent variables.

Katz (1953) has divided field studies into two board types – exploratory and hypothesis testing. The exploratory types seek what is, rather than predict relations to be found. They have three purposes: (1) to discover significant variables in the field situation, (2) to discover relations among variables (3) to lay a ground work for later, more systematic and

rigorous testing of hypothesis.

It is well to recognize though that there are activities preliminary to hypothesis testing in scientific research. In order to achieve the desirable aim of hypothesis testing, preliminary methodological and measurement investigation must often be done. The second subtype of exploratory field studies, research aimed at discovering or uncovering the relations, is indispensable to scientific advancement in the social sciences.

The field studies are strong in realism, significance, strength of variables, theory orientation and heuristic quality. The realism of field studies is obvious. They are highly heuristic. Any researcher knows that one of the research difficulties of the field studies is to keep himself contained within the limits of his problem. Hypothesis is frequently fling themselves at one. The field is rich in discovery potentiality. After starting to gather data, he might stumble upon many interesting notions that can reflect the course of investigation.

Despite these strengths, the field study is a scientific weakness of laboratory experiments. Its most serious weakness of course is its ex-post facto character. Anther methodological weakness is lack of precision in the measurement of field variables. Other weakness of field studies are practical problems: feasibility, cost, sampling, and time. The field researcher therefore, needs to be salesman, administrator and entrepreneur as well as investigator.

2.4.4 Case Study

The case study is one of the important types of non-experimental research. The case study is not a specific technique rather it is one way of organising social data for the purpose of viewing social reality. It tends to preserve the unitary character of a social object being studied. It tends to examine a social unit as a whole. The unit may be a person a family a social group a social institution or even a community (Goode & Hatt 1981, Best & Kahn 1992).

A case study may utilize interview, observation, and psychological tests. It is a valuable research strategy in the fields of clinical psychology and human development. Using case study a researcher is able to have an in-depth look at one person. Those unique aspects of a person's life which cannot be duplicated for practical or ethical reasons are captured by case study. With the help of case study, you can try to understand fantasies hopes fears traumatic experiences upbringing or anything that helps to understand a person's mind and behaviour.

Case studies provide a narrative or detailed description of the events that takes place in a person's life. Freud's insight that led to the development of psychoanalytic theory emerged from his observation and reflections on individual cases. It should be remembered that the person studied as a case is unique and our judgments are of unknown reliability. Case studies provide detailed in-depth depictions of people's lives but we need to exercise caution when generalizing from individual cases. They are like naturalistic observations and all one can do is to describe the course of events.

The problem of validity of single case study is very serious. It is therefore recommended that researchers should use objective measurement techniques multiple sources of information and frequent assessment of relevant variables. The uses of case study as a research strategy requires that the cases must be chosen that represent the variable in question and one must have sufficient access to the cases. Careful planning of data collection is very necessary. Throughout the data-collection process the investigator is required to maintain a chain of evidence linking the various data sources having bearing on the research questions.

2.5.1 Experimental Methods

2.5.1.1 Laboratory Experiments

As you know a laboratory experiment is one of the most powerful techniques for studying the relationships between variables under controlled condition. It may be defined as the study of a problem in a situation in which some variables are manipulated and some are controlled in order to have an effect upon the dependent variable. The variables which are manipulated are known as independent variables and the variables which are controlled, are known as extraneous or relevant variables. Thus, in a laboratory experiment the effect of manipulation of an independent variable upon the dependent variable is observed under controlled conditions. Festinger & Katz (1953:137) have defined a laboratory experiment as "one in which the investigator creates a situation with the exact conditions he wants to have and in which the controls some, and manipulates other variables".

Kerlinger (1986), there are three main purposes of the laboratory experiment. First, a laboratory experiment purports to discover a relationship between the dependent variable and the independent variable under pure, uncontaminated and controlled conditions. When a particular relationship is discovered, the experimenter is better able to predict the dependent variable. Second, a laboratory experiment helps in testing the accuracy of predictions derived from theses or researches. Third, a laboratory experiment helps building the theoretical systems by refining theories and hypotheses and thus, provides a breeding ground for scientific evaluation of those theories and hypotheses.

2.5.2 Field Experiment

A field experiment is very similar to a laboratory experiment. A field experiment may be defined as a study carried out in a more or less realistic situation or field where the experimenter successfully manipulates one or more independent variables under the maximum possible controlled conditions. Experimenter manipulates one or more independent variable in natural setting for determining their effect upon behaviour, the procedure is known as field experiment.

Field experiment has number of Strengths which are given below:

1. A field experiment deals with the realistic life situation. Hence it is more suited for studying social changes, social processes and social influence.
2. One principle of research is that the more realistic the situation, the stronger is effect

of the variables under study. In a field experiment this principle is fully satisfied. Thus, one can say that in the field experiment, since it deals with a realistic situation, the variables have stronger and more obvious effects.

3. Is derived from the above two points. When variables are stronger because of more realistic situations, an experimenter can make better and more sound generalizations on the basis of the obtained results. In other words, this tends to increase the external validity of the field experiment. For example, when one carried out a field experiment by taking small groups of workers from a factory, and reaches the conclusion that absenteeism among workers is primarily due to the poor financial incentive, this can be safely generalized with respect to the workers of other factories as well because the experiment has been carried on actual workers in a factory.
4. A field experiment is well-suited for testing a broad hypothesis and theories and for obtaining answers to practical questions.

The principles **weaknesses of field experiments** are as given below:

1. Since a field experiment is carried out in a realistic situation, there is always the possibility that the effect of independent variables is contaminated with uncontrolled environmental variables.
2. The unexpected noise and gathering may affect the dependent variable and thereby, contaminate the influence of the independent variable. In a laboratory experiment this problem does not arise because of the fully controlled laboratory situation. However, if the situation is somehow fully controlled in a field experiment, it would prove to be a more powerful tool than the laboratory experiment.
3. In many field situations the manipulation of independent variables may be difficult due to non-cooperation of subjects. Children are to be exposed to frustrating situations; they may not like it and may restrain their children from being exposed to field situation.
4. In a field experiment it is not possible to achieve a high degree of precision or accuracy because of some uncontrolled environment variables.
5. Field experiment requires that the investigator has high social skills to deal effectively with people in a field situation.

2.5 SUMMARY

This chapter provides a comprehensive discussion on *research methods*, *research methodology*, and the various categories of research techniques used in scientific studies.

The chapter begins by explaining **research methods**, which refer to practical tools, techniques, and procedures used to collect and analyze data. Examples include interviews, observations, experiments, content analysis, and surveys. Methods form the *heart* of data collection and help researchers decide how to gather and analyze information—whether qualitative or quantitative, descriptive or experimental. It then moves to **research methodology**, which deals with the overall logical framework and rationale of research. Methodology answers the "how" and "why" behind using specific methods. It includes the theories, strategies, and justification for selecting research methods. Methodology ensures that

the study's design, tools, and techniques align with its objectives and maintain validity, reliability, and credibility.

The chapter also clearly outlines the **differences between methods and methodology**, showing that methods are tools, while methodology is the blueprint guiding the tools. Chapter introduces **two broad categories of research methods**: Non-experimental methods and **Experimental methods**, Overall, the chapter helps learners understand how methods and methodology work together to shape high-quality, systematic, and valid scientific research.

2.6 KEY TERMS

- **Research Method** – Tools and techniques used to collect and analyze data.
- **Research Methodology** – The overall strategy and rationale guiding how research is conducted.
- **Naturalistic Observation** – Studying behavior in its natural environment without interference.
- **Participant Observation** – The researcher becomes part of the group being observed.
- **Observer Effect** – When individuals change behavior because they know they are being watched.
- **Observer Bias** – When a researcher's expectations influence what they record.
- **Archival Research** – Using existing records, documents, or datasets for analysis.
- **Content Analysis** – Systematically examining written, visual, or recorded content.

2.7 SELF-ASSESSMENT QUESTIONS:

Multiple Choice Questions (MCQs)

1. Which of the following are non-experimental research methods?
✓ Naturalistic Observation
✓ Archival Research
✓ Case Study
✗ Laboratory Experiment

2. Which elements are part of research methodology?
✓ Research design
✓ Theoretical framework
✓ Rationale for choosing methods
✗ Raw data collection tools only

3. Which are advantages of surveys?
✓ Wide scope
✓ Ability to estimate population characteristics
✓ Cost-effective for large populations

✗ Eliminates all sampling errors

4. Which are weaknesses of field experiments?
✓ Difficulty in controlling external variables
✓ Difficulty in manipulating variables ethically
✗ Lack of real-life applicability
✓ Lower precision due to environmental interference

SHORT ANSWER QUESTIONS

1. Define research method and give one example.
2. What is research methodology and why is it important?
3. Differentiate between research method and research methodology.
4. What is naturalistic observation?
5. Explain observer bias with an example.
6. What is archival research?

Questions

1. What do you understand by the term research method?
2. Define research methodology and explain its significance in research.
3. Explain how research methodology helps in ensuring the validity and reliability of research findings
4. Explain at least five key differences between research method and research methodology.
5. Describe the main advantages and disadvantages of naturalistic observation.
6. What are independent and dependent variables? Give examples.
7. Define content analysis and explain how it is conducted. Discuss its main data sources, merits and demerits.

CASE STUDY

Case Study: Classroom Behavior Research

A researcher wants to understand how 10th-grade students behave during free periods. Instead of conducting an experiment, she decides to sit quietly in the classroom without interfering and observe the interactions between students. She notices that when students realize they are being watched, some change their behavior. She also records her observations in detail, but later realizes that she may have interpreted some behaviors based on her own preconceived beliefs.

Questions:

1. Which research method is being used in this scenario?
2. Identify the observer effect in this case.

3. Identify the observer bias present here.
4. Would a field experiment be suitable here? Why or why not?
5. Suggest one additional method (survey, interview, or archival research) that could supplement the study

REFERENCES:

- Festinger and D Katz (Eds). Research Methods in Behaviour Sciences, New York: Holt Rinehart & Winston, Inc, Indian Edition 1970
- Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Age International Publishers.

2.8 SUGGESTED READINGS

- Kothari, C.R. 2004. *Research Methodology Methods and Techniques,* NewAge International (P) Limited: New Delhi.
- Rao K.V. 1993. *Research Methodology in Commerce and Management,* Sterling Publishers Private Limited: New Delhi.
- Sadhu, A.N. and A. Singh, 1980. *Research Methodology in Social Sciences,* Sterling Publishers Private Limited: New Delhi.

RANGANATHAM GANGINENI

LESSON-3

FORMULATION OF RESEARCH PROBLEM & HYPOTHESIS

OBJECTIVES OF THIS CHAPTER:

After studying this lesson, students will be able to:

- Define a research problem
- Identify various sources of research problems
- Explain the essentials of a good research problem
- Understand the meaning and purpose of a hypothesis
- Describe different types of hypotheses
- Identify the characteristics of a good hypothesis
- Explain the role of hypotheses in guiding research

STRUCTURE OF THE LESSON

3.1 FORMULATING RESEARCH PROBLEM

According to Bryman, Alan. ―The Research Question in Social Research: What is its Role?‖ International Journal of Social Research Methodology 10 (2007); A research problem is a definite or clear expression [statement] about an area of concern, a condition to be improved upon, a difficulty to be eliminated, or a troubling question that exists in scholarly literature, in theory, or within existing practice that points to a need for meaningful understanding and deliberate investigation. A research problem does not state how to do something, offer a vague or broad proposition, or present a value question. It is not always easy to formulate the research problem simply and clearly. It may take years to decide for some and just a few minutes for others to decide the research problem to be studied. The social issues may provide a broader prospect but it may not suggest a specific one. E.g. understanding economic background of society may not address the issues of unemployment in the same society therefore unemployment needs to studied differently and individually to assess the underlying problems.

The availability of resources like money, time, manpower, etc. also affects the selection of research problem.

Some sources of Research Problems may be identified as follows:

- Personal Experiences.
- Media: Documentation done on various issues, live coverage, panel discussions etc.
- Resources: Literature such as books, journals, news articles, periodicals etc may facilitate the researcher to identify a relevant problem based on the area of interest.
- Government / Official Records: The orders passed by government. The decisions given in various cases by courts, the petitions and surveys conducted become important sources to shortlist finger prints in abrader problem.
- People: A group of individuals may be studied to understand how they behave, how tiny respond to a particular situation do or what responses are generated when they are influenced from within or outside the group.
- Discussions: A researcher may be able to come to a conclusion to identify a research problem by discussing the perspectives with peers, colleagues, seniors in the field, guides etc.
- Problems: It may be decided to examine the existence of certain issues or problems relating to society, sciences or any subjects in reference.
- Programs: These may be used to evaluate the effectiveness of an interference, involvement or intrusions.
- Phenomena: To establish the existence of regularity and to understand if a procedure would yield similar results overtime when used repetitively. This includes causes and effects and relationships between variables.
- Ideas from external sources.
- Interdisciplinary Perspectives.

3.2 ESSENTIAL OF A GOOD RESEARCH PROBLEM

- **Question Mark?**
  The research problem can be in a declarative or in a question form. We recommend you to formulate your research problem as a question. This gives you (and the reader) something to hold on to during the rest of your thesis because it is simple: there is a question and, in the text, you look for an answer.

- **Possibility to Respond**
  Some questions are impossible to answer in a scientific way, for example: ‘how beautiful is the color yellow’. We don’t have the scientifically justified instruments to answer this question. It must also be possible to answer the question in a practical way so it must be researchable, meaning you have to be able to collect evidence that will answer the question.

- **Attainability**
  The problem must be one that can be solved during the amount of time you have. So it can't be too broad (ex: 'How can we have world peace?'). But it also can't be too narrow (ex: How does my neighbor think about Indians?').

- **Open Question**
  The research problem should be an open question. That means it cannot be answered by "yes" or "no". But also, with open questions you should watch out for the possibility of a shallow answer.

- **Unmistakability**
  Your research problem must be clear and there has to be only one way to interpret it. For example: The question 'What do Indians think about the West?' is un mistakable because it is not clear what is meant by 'the West', it can be a lot of things.

- **Punctuality**
  The problem must be clearly specified. For example: Don't write 'How can prejudices against Americans be combated?' if you mean: 'How can prejudices that live among Indian students for Americans be combated?'

- **Brevity**
  Although your research problem should be as punctual and specific as possible, not all fencings must be placed in your research problem.

R.S. Woodworth defines problem as "a situation for which we have no ready or successful response by instinct or by previously acquired habit we have to find the answer."

According to John Dewey- "the need of clearing up confusion, of straightening out an ambiguity, of overcoming obstacles, of covering the gap between things as they are and as they may be when transformed is a problem." Thus, research is an enquiry geared to the solution of problem. Hence, the first step in any research is to make the problem concrete and explicit. A researcher should identify some aspect of the topic which can be formulated into specific research questions. These should be capable of investigation with resources available to a researcher.

3.3 HYPOTHESIS

Hypothesis is usually considered as the principal instrument in research. Its main function is to suggest new experiments and observations. In fact, many experiments are carried out with the deliberate object of testing hypotheses. Decision makers often face situations wherein they are interested in testing hypotheses on the basis of available information and then take decisions on the basis of such testing. In social science, where direct knowledge of population parameter(s) is rare, hypotheses testing is the often-used strategy for deciding whether a sample data offer such support for a hypothesis that generalization can be made. Thus, hypothesis testing enables us to make probability.

Statements about population parameter(s). The hypothesis may not be proved absolutely. But in practice it is accepted if it has withstood a critical testing. Before we explain how hypotheses are tested through different tests meant for the purpose, it will be appropriate to explain clearly the meaning of a hypothesis and the related concepts for better understanding of the hypothesis testing techniques.

Ordinarily, when one talks about hypothesis, one simply means a mere assumption or some supposition to be proved or disproved. But for a researcher hypothesis is a formal question that he intends to resolve. Thus, a hypothesis may be defined as a proposition or a set of propositions set forth as an explanation for the occurrence of some specified group of phenomena either asserted merely as a provisional conjecture to guide some investigation or accepted as highly probable in the light of established facts. Quite often a research hypothesis is a predictive statement, capable of being tested by scientific methods, that relates an independent variable to some dependent variable. For example, consider statement like the following ones.

"Students who receive counseling will show a greater increase in creativity than students not receiving counseling" or

"The automobile A is performing as well as automobile B."

These are hypotheses capable of being objectively verified and tested. Thus, we may conclude that a hypothesis statement is what we are looking for and it is a proposition which can be put to a test to determine its validity.

3.4 CHARACTERISTICS OF GOOD HYPOTHESIS

Hypothesis must possess the following characteristics:

(i) Hypothesis should be clear and precise. If the hypothesis is not clear and precise, the inferences drawn on its basis cannot be taken as reliable.
(ii) Hypothesis should be capable of being tested. Some prior study may be done by researcher in order to make hypothesis a testable one. A hypothesis "is testable if other deductions can be made from it which, in turn, can be confirmed or disproved by observation."
(iii) Hypothesis should state relationship between variables, if it happens to be a relational hypothesis.
(iv) Hypothesis should be limited in scope and must be specific. A researcher must remember that narrower hypotheses are generally more testable and he should develop such hypotheses.
(v) Hypothesis should be stated as far as possible in most simple terms so that the same is easily understandable by all concerned. But one must remember that simplicity of hypothesis has nothing to do with its significance.
(vi) Hypothesis should be consistent with most known facts i.e. it must be consistent with a substantial body of established facts. In other words, it should be one which judges accept as being the most likely.

(vii) The hypotheses selected should be amenable to testing within a reasonable time. The researcher should not select a problem which involves hypotheses that are not agreeable to testing within a reasonable and specified time. He must know that there are problems that cannot be solved for a long time to come. These are problems of immense difficulty that cannot be profitably studied because of the lack of essential techniques or measures.

(viii) Hypothesis must explain the facts that gave rise to the need for explanation. This means that by using the hypothesis plus other known and accepted generalizations, one should be able to deduce the original problem condition. Thus, hypothesis must actually explain what it claims to explain, it should have empirical reference.

3.5TYPES OF HYPOTHESES

Social researchers have to work with many kinds of hypotheses. Hence, they can be classified in several ways.

(I) If hypotheses are classified in relation to their functions, we have two types: -

1. **Descriptive hypotheses:** These are propositions which describe the characteristics of a variable, e.g., income, expenditure, hours of work, size of public sector undertakings, etc. The variable may be an object, person, organization, situation or event. In these cases, the hypotheses assert a particular characteristic. Statements about the rate of inflation during a given period, the number of unemployed in a certain region, the rate of growth of agricultural products in a certain state, the income level of certain class of people, etc., describe particular characteristics.
2. **Relational hypotheses:** These are propositions which describe the relationship between two variables. This type of hypotheses state that something is greater or less than something or one variable occurs in a certain proposition of time of another variable. For example, families with small income spend large proportion of their income on necessities, increase in standard of living tends to reduce in income helps in increasing expenditure, etc.

This statement suggests that the relationship between two variables can be direct or inverse, i.e., it can be either positive or negative. Some of these statements also indicate causal relationships, i.e., change in one variable is a cause or effect of a change in the other variable. The variable that causes the change is called independent variable, the other variable or the affected variable is called the dependent variable. The researcher must be careful in determining, which is the cause and which is the effect.

(II) Another approach is to classify hypotheses into three types:

1. **Working Hypotheses:** Hypotheses are formulated while planning the study of a problem. They may not be specific in the initial stages. In such cases, they are called “working hypotheses.” Such hypotheses are subject to change or modification in the course of investigation.

2. **Null Hypotheses:** These are hypothetical statements and they deny what is stated in working hypotheses. They do not, and are not ever expected to, exist in reality. We discuss these types of hypotheses subsequently in the section on testing of hypothesis. There is some justification for using null hypotheses. They ensure detachment and objectivity in testing the hypothesis. Moreover, null hypotheses are more exact statements like "education does not increase the earning capacity of an individual" is a null hypothesis.
3. **Statistical Hypotheses:** These are statements about a statistical population. The statements are derived from a sample drawn from a given population. Statistical hypothesis is quantitative in nature as they are numerically measurable. These hypotheses can be hypotheses of difference or association, i.e., we can formulate them as null hypotheses or causal hypotheses.

(III) Classification of hypotheses on the basis of their level of abstraction is regarded as especially useful. Goode and Hatt have identified three broad levels of abstractions.

1. At the lowest level, we have simple description which gives rise to commonsense hypotheses. They state existence of certain empirical uniformities and hence expect verification of commonsense propositions, e.g., professionally qualified males marry at a late age, or a worker, in a small establishment lacks motivation. It is often said that such statements are not useful as hypotheses, since they merely state what everyone knows may be mistaken. Secondly, what everyone knows is not put in precise terms or is integrated in scientific terms. Commonsense statements are often a mixture of clichés and moral judgments. This is particularly true in social sciences. Scientists have to transform these statements and test them. It requires (a) removal of value judgments from the statements. (b) clarification of terms and (c) application of validity tests. According to Goode and Hatt, "what everybody knows is not known until it has been tested." Hence, such empirical generalizations play and important role in the growth of science.
2. At a relatively higher level of abstraction, we have logical derivations which give rise to complex hypotheses. These aim at testing the existence of logically derived relationships between empirical uniformities. They are purposeful distortions of empirical reality. Hence, they are also called "Ideal types." Such hypotheses try to create tools and problems for further research in complex areas of investigations.
3. The category of hypothesis at the highest level of abstraction is concerned with the relation of analytic variables. Hence, they are called analytical hypotheses. These are statements about how changes in one property will affect another property, e.g., statements about relation between level of education and migration, or level of income and social mobility are some such abstractions. These abstractions are highly sophisticated mode of formulation. They contribute to the development

of brilliant abstract theories. However, this does not mean that these types of hypotheses are superior or better than the other types. Each type of hypotheses has its own importance. Use of a particular type of hypotheses depends on the nature of investigation.

3.6 ROLE OF HYPOTHESIS

In social science research, hypothesis serves several important functions.

1. A hypothesis guides the direction of study - A hypothesis guides the direction of study or investigation. It states what we are looking for.
2. Its purpose is to include in the investigation - Its purpose is to include in the investigation all available and pertinent data either to prove or disprove the hypothesis.
3. Research becomes unfocussed or random - Research becomes unfocussed or random without a hypothesis and useless data may be collected in the hope that important data is not omitted.
4. Hypothesis specifies the sources of data - Thus, a hypothesis specifies the sources of data, which shall be studied and in what context they shall be studied.
5. Needs and prevents a blind search - It also determines the data needs, and prevents a blind search.
6. Study a given problem - A hypothesis can suggest the type of research which is likely to be appropriate to study a given problem.
7. Appropriate technique of data analysis - It determines the most appropriate technique of data analysis.
8. Various hypotheses relating to a stated theory - A hypotheses can contribute to the development of theory by testing various hypotheses relating to a stated theory. It is also likely, in some cases, that a hypothesis helps in constructing a theory.

3.7 SUMMARY:

This chapter discusses the fundamental concepts of research problems and hypotheses, the first essential steps in conducting scientific research. definition of a **research problem**, explaining that it is a clear statement about an issue or area of concern requiring investigation. The formulation of a research problem can be difficult and may depend on personal interest, literature availability, societal issues, and available resources like time, money, and manpower. Various sources for identifying research problems are outlined, including personal experiences, media, government records, literature, discussions, social problems, and interdisciplinary ideas. The **essentials of a good research problem**, stating that it must be clear, researchable, open-ended, specific, attainable, unmistakable, and concise. A well-formulated research problem helps guide the entire research process. **Hypotheses**, describing them as tentative, testable propositions that explain relationships between variables. Hypotheses are crucial in research because they guide investigation, suggest the type of data needed, and determine the techniques of

analysis. A good hypothesis must be clear, testable, specific, consistent with known facts, and empirically referent. Different **types of hypotheses** are presented, including descriptive, relational, working, null, statistical, commonsense, complex, and analytical hypotheses. These classifications depend on the function and level of abstraction. The chapter explains the **role of hypotheses** in social science research. Hypotheses direct the study, prevent random data collection, define the type of data required, guide analysis, and contribute to theory-building.

3.8 TECHNICAL TERMS

- **Research Problem** – A clear statement of an issue that requires investigation to generate understanding or solutions.
- **Sources of Research Problem** – Different ways through which a researcher identifies issues worth studying.
- **Attainability** – The research problem must be solvable within available time, resources, and skills.
- **Open Question** – A research question that cannot be answered with a simple yes/no response.
- **Unmistakability** – The research problem must be clear with only one interpretation.

3.9 SELF-ASSESSMENT EXERCISE

Multiple Choice Questions (MCQs)

1. **Which of the following are valid sources of research problems?**
   ✓ Personal Experiences
   ✓ Literature and Books
   ✓ Government Records
   ✓ Discussions with Experts
   ✗ Personal Opinions without Evidence
2. **A good research problem must be:**
   ✓ Clear and specific
   ✓ Researchable
   ✓ Attainable within available resources
   ✗ Answerable through yes/no responses
3. **Which of the following are types of hypotheses?**
   ✓ Descriptive Hypothesis
   ✓ Relational Hypothesis
   ✓ Null Hypothesis
   ✓ Statistical Hypothesis
   ✗ Imaginary Hypothesis

4. **Characteristics of a good hypothesis include:**
   ✓ Testability
   ✓ Simplicity
   ✓ Consistency with known facts
   ✗ Inability to be proven wrong

SHORT ANSWER-QUESTIONS

1. Define a research problem in your own words.
2. Mention any three sources of a research problem.
3. What is meant by attainability in research problem formulation?
4. Why should a research problem be an open question?
5. Define hypothesis and give one example.
6. Differentiate between independent and dependent variables.
7. What is the difference between descriptive and relational hypotheses?
8. Why is a null hypothesis used in research?
9. What are the characteristics of a good hypothesis?
10. What is a working hypothesis?

LONG ANSWER-QUESTIONS:

1. Explain the meaning and importance of formulating a research problem in social research
2. A good research problem should be clear, attainable, and open-ended." Discuss
3. Discuss the meaning and role of hypothesis in social research.
4. Compare null, working, and statistical hypotheses.
5. Discuss the characteristics of a good hypothesis in detail.

CASE STUDY

Case Study: Identifying a Research Problem

A researcher observes that college students increasingly prefer short video content over long educational videos. The researcher wants to study the impact of this preference on students' academic performance. She reads current literature, discusses the issue with faculty members, and identifies that the topic is manageable within her timeline and resources. She plans to formulate a hypothesis linking video preference and academic outcomes.

Questions:

1. Identify the research problem in this scenario.
2. Mention two sources the researcher used to identify the research problem.
3. Suggest a suitable research question for the study.
4. Propose one relational hypothesis for this study.
5. Name the independent and dependent variables.

REFERENCES:

- Kothari, C. R. (2004). *Research Methodology: Methods and Techniques* (2nd ed.). New Age International Publishers.
- Batra, S., & Kaushik, A. (n.d.). *The research process: Steps*. In *Social Work Research Process* (Unit 1). Department of Social Work, University of Delhi.
- Bhome, S., Jha, N., Chandwani, V., Iyer, S., Desai, S., Prabhudesai, A., & Koshti, S. D. (2013). *Research methodology in commerce*. Mumbai: Himalaya Publishing House Pvt. Ltd.

3.10 FURTHER READINGS:

- Kothari, C.R. 2004. *Research Methodology Methods and Techniques,* NewAge International (P) Limited: New Delhi.
- Rao K.V. 1993. *Research Methodology in Commerce and Management,* Sterling Publishers Private Limited: New Delhi.
- Sadhu, A.N. and A. Singh, 1980. *Research Methodology in Social Sciences,* Sterling Publishers Private Limited: New Delhi.

RANGANATHAM GANGINENI

LESSON-4

RESEARCH DESIGN

OBJECTIVES OF THE CHAPTER

After studying this chapter, you will be able to:

1. Understand the meaning and purpose of a research design.
2. Identify the need and importance of planning a research design in advance.
3. Explain the characteristic*s* of a good and scientific research design.
4. Differentiate between quantitative and qualitative research designs.
5. Explain types of quantitative designs—descriptive, correlational, experimental, quasi-experimental.
6. Understand the nature of qualitative research design.
7. Learn the steps involved in developing a research design.
8. Recognize how research design ensures objectivity, reliability, validity*,* and ethical rigor.
9. Apply the logic of research design to real research situations.

STRUCTURE OF THE LESSON

4.1 INTRODUCTION

The Chapter 'Research Design' brings forth various phases of research that could be carried out in a planned and phased manner by developing a research design. Once the researcher finalizes the research design, the entire research process will be under their control. The example below will make learners realize the significance of developing a research design. The architect prepares the design, keeping in mind that he needs to construct a building. The plan will contain the structure of the building, number of rooms, utilization of rooms, size of open space, size of garden space, building material, and length

and breadth of the building in square feet. The architect will decide and prepare a blueprint or plan before starting the building. I believe the above example will help you know the research design's significance. Once the design is in the hands of the researcher, they have the freedom to make necessary changes whenever the need arises.

Designing also helps the researcher reduce wasteful expenditure on finance, time and energy. The process of preparing a research design anticipates various difficulties, and the same difficulties will be addressed in the research design through appropriate strategies. Preparing the research design also helps the researcher to control the unexpected situation rationally. It also facilitates the researcher to articulate the research procedure in a better manner during the academic presentation. It helps them to prevent the possibility of failure. With the above introduction, the present unit facilitates you to learn the definition and need for research design, principles of design research design, and types of research designs. Finally, we discuss sampling techniques.

4.2 NEED FOR RESEARCH DESIGN

Let us start with defining “what is Research Design”. We define the research design as “such design as a symbolic construction or model”. “The research design is the arrangement of conditions for data collection and analysis in a manner that aims to combine relevance to the research purpose with the economy in procedure” (Selltiz et al. 1964). If we look at the above definition, we can see that the design talks explicitly about the research objectives, the rationale for doing specific research, and a methodology which includes methods and theory, sampling, data and data organization. Besides the core procedures mentioned above, scientific research design must explicitly describe the accepted methodology. The theory must be a core component of the social science research design. The research design should have the following aspects.

a. The details of the study
b. The rationale for choosing the study
c. Reasons for carrying out the study
d. Type of data required for the chosen study
e. Techniques of data collection
f. Area of Study
g. The period of study
h. Duration of the study
i. Requirements of different types of material for the research study
j. Tools and techniques required for the study
k. Selection of cases, number of cases required for the study
l. Methods of data analysis

4.3 PROPERTIES/CHARACTERISTICS OF GOOD RESEARCH DESIGN

In the view of various definition of research design, the following characteristics are found.

- **Objectivity:** Objective findings may be achieved by allowing more than one person to agree between the final scores/ conclusion of the research. E.g. a researcher studying the impact of a training program on employee performance uses standardized measurement tools, applies the same evaluation criteria to all participants, and reports the results exactly as obtained, even if they contradict the researcher's expectations. This demonstrates objectivity. E.g. A researcher analyzing survey data presents both positive and negative findings and avoids selective reporting of results to support a preconceived conclusion, thereby maintaining objectivity.
- **Reliability:** Researcher should ensure that research questions are framed judiciously to make it reliable and provide similar outcomes. Thus, the results obtained should be similar if the research is conducted in identical conditions and is repeated time and again. E.g. A Guest Comments Card contains the same set of questions and responses for all the guests staying at the hotel and it is suitably placed on the study table in a guest room allowing each guest to take time and fill the data.
- **Generalization:** The information collected from given sample must be utilized for providing a general application to the large group of which the sample is drawn. E.g. if a researcher studies the study habits of 300 undergraduate students selected randomly from a university and concludes that most students prefer digital learning resources, this conclusion may be generalized to the entire student population of that university, provided the sample is representative.
- **Ethical:** It should be acceptable and be free of practices or procedures that may not be honest or may give error /bias. E.g. A study involving human subjects ensures that participants are not harmed physically or psychologically, allows them to withdraw from the study at any time, and uses the collected data only for academic purposes. This reflects adherence to ethical principles
- It should be proficient in obtaining the most reliable and valid data;
- A good research design should be able to address any situation wherein any unexpected events can be accommodated.
- It also helps a researcher arriving at flawed / misunderstood conclusions;
- It can adequately control the various threats of validity, both internal and external.

4.4 TYPES OF RESEARCH DESIGN

Research design can be a **quantitative or qualitative research** with have extensive components. They can both be used or applied distinctively or together.

- **Quantitative Research design:** A quantitative research design shares similar characteristics with scientific research in the following ways:
- An outline question stating the problem that needs to be solved.
- Has a set order and procedure used to answer these questions?
- Analyses the data generated.
- Draws its conclusion after the data has been collated and analyzed so that the conclusion drawn from the findings are not predetermined.

A quantitative research design is used to examine the relationship between variable by using numbers and statistics to explain and analyze its findings and there are four types of quantitative research design. E.g. A study that measures the impact of study hours on students' academic performance using structured questionnaires and examination scores, and analyzes the data through statistical tests such as correlation and regression, is a quantitative research design.

1. **Descriptive design research**: As the name implies, it is intended to describe the present status of an issue or a problem which is analyzed based on the available data and so does not require hypothesis to begin with. E.g. If a guest is complaining about a faulty shower in the bathroom just because he may not have used a modular shower earlier has to be resolved delicately and not by pointing out to him that he is not aware of new technology. E. g A research study aimed at documenting the present working conditions, job satisfaction levels, and organizational structure of employees in an organization, based on existing records and survey responses, is an example of descriptive research. The study focuses only on what exists rather than why it exists.
2. **Co relational design research:** This seeks to discover if two variables are associated or related in some way, using statistical analysis, while observing the variable. E.g. If the heat is reduced or increased during cooking how does the food react to it. E. g. A study that analyzes the relationship between study hours and academic performance of students using correlation coefficients is a correlational study. The researcher only observes and measures the variables and applies statistical analysis to determine the degree and direction of association.
3. **Experimental design research:** This is a method used to establish a cause-and-effect relationship between two variables or among a group of variables. The independent variable is manipulated to observe the effect on the depended variable. E.g. The change in response to between groups of foreigners treated to welcome drinks and freshener tissues and the one that is simply welcomed and allocated rooms in a hurry due to peak hours of check in and check out. E. g. A study in which participants are randomly assigned to two groups—an experimental group and a control group—to test the effect of a new teaching method on students' academic performance is an experimental research design. The new method is applied only to the experimental group, and the outcomes are compared statistically
4. **Quasi-experimental design research:** As the name suggests such an experiment is designed replicating the true experimental design, except that it does not use randomized sample groups. Also, it is used when a typical research design is not practicable. Qualitative Research Design. E.g. A study evaluating the impact of a new curriculum on students' academic performance by comparing test scores of one school that adopts the curriculum with another school that does not, without random assignment of students, is a quasi-experimental study.

Qualitative research design, on the other hand, is exploratory in nature as it tries to discover not guess the conclusion. It seeks to answer the questions what and how. It is a process to identify or develop a hypothesis that is further tested using other techniques. E.g. A research study that examines students' experiences with online learning by conducting

semi-structured interviews and analyzing narratives to identify recurring themes represents a qualitative research design.

A qualitative research design is used to explore the meaning and understanding of complex social environments, like the nature of experiences gained by a tourist by reading about the texts and stories shared by them. It also intends to understand, describe or discover the findings. The researcher is usually the primary instrument that formulates the question and interprets the meaning of a data. The data used are mostly documented words from interview, newspapers videos etc. More than one type of data is collected during this research, from the field, where the participants are. In other words, the research goes beyond the intended scope, so making it emergent because the method of research changes and different types of data might be collected as the research goes on.

4.5 STEPS IN DEVELOPING RESEARCH DESIGN

a. Classify the intended outcome i.e. what needs to be understood.
b. Develop the research question.
c. Understand what needs to be measured.
d. Select the population as per the study taken up.
e. Identify the ideal data collection method.
f. Construct interconnected characteristics.
g. Use correct analysis tools.
h. Decide how the findings of the study shall be published.

a. Classify the Intended Outcome (What Needs to Be Understood)

The first step is to clearly define what the study aims to achieve. This involves identifying the purpose of the research, such as:

- Exploring a new phenomenon
- Describing characteristics of a population
- Examining relationships between variables
- Testing hypotheses or theories

At this stage, the researcher determines whether the study is exploratory, descriptive, explanatory, or causal. A clear understanding of the intended outcome ensures that all subsequent methodological choices align with the research goal.

b. Develop the Research Question

Once the intended outcome is identified, the researcher formulates precise and focused research questions. A good research question:

- Clearly states the problem
- Is specific, measurable, and researchable
- Defines the variables and population
- Guides data collection and analysis

Well-framed research questions help maintain focus and prevent scope creep, ensuring the study remains relevant and manageable.

c. Understand What Needs to Be Measured

This step involves identifying the key variables and constructs relevant to the research questions. The researcher must determine:

- Independent and dependent variables
- Mediating or moderating variables (if any)
- Abstract concepts that require operational definitions

For example, psychological or behavioral constructs such as investor sentiment, emotional bias, or financial risk perception must be translated into measurable indicators using scales or indices.

d. Select the Population as per the Study Taken Up

The researcher then defines the target population to which the findings will be generalized. This includes:

- Identifying who or what will be studied
- Determining inclusion and exclusion criteria
- Selecting an appropriate sampling frame

Decisions regarding sample size and sampling technique (probability or non-probability sampling) are made at this stage to ensure representativeness and validity.

e. Identify the Ideal Data Collection Method

Based on the nature of the research and type of data required, the researcher selects suitable data collection methods, such as:

- Surveys or questionnaires
- Interviews or focus groups
- Experiments or observations
- Secondary data sources (databases, reports, publications)

The chosen method should ensure accuracy, reliability, feasibility, and ethical compliance

f. Construct Interconnected Characteristics

This step focuses on establishing logical relationships among variables. The researcher:

- Develops a conceptual or theoretical framework
- Links variables based on existing theories and literature
- Proposes hypotheses or propositions

Constructing interconnected characteristics helps explain how and why variables influence each other, strengthening the theoretical foundation of the study.

g. Use Correct Analysis Tools

After data collection, appropriate statistical or qualitative analysis tools are selected based on:

- Nature of data (quantitative, qualitative, or mixed)
- Research objectives and hypotheses
- Measurement scales used

Examples include regression analysis, structural equation modeling, thematic analysis, or bibliometric tools. Using correct analysis techniques ensures valid interpretation and robust findings.

h. Decide How the Findings of the Study Shall Be Published

The final step involves planning how the research outcomes will be communicated and disseminated. This includes:

- Choosing suitable journals, conferences, or repositories
- Structuring the manuscript according to publication guidelines
- Identifying the target academic or practitioner audience

Effective dissemination enhances the visibility, impact, and practical relevance of the research.

4.6 SUMMARY

This chapter explains the concept and importance of research design in conducting systematic research. A research design is the structured blueprint that guides the entire research process, similar to an architect's plan for constructing a building. A well-developed design helps the researcher anticipate problems, use resources efficiently, and carry out the study in a planned, logical manner. Highlights the *need for research design*, stating that it should clearly specify the study details, rationale, methodology, sampling, data requirements, tools, techniques, and methods of analysis. The design connects the research objectives with appropriate strategies. outlines the *characteristics of a good research design*, including objectivity, reliability, generalizability, ethical considerations, and the ability to handle unexpected situations. Research designs are categorized into *quantitative* and *qualitative*. Chapter lists the *steps in developing a research design.*

4.7 TECHNICAL TERMS:

- **Research Design**: A blueprint that guides how a research study will be conducted.
- **Objectivity**: Keeping the research free from personal bias.
- **Reliability**: Producing consistent results when repeated.
- **Sampling**: Choosing a part of a population to study.
- **Data Collection**: The process of gathering information for research.
- **Validity**: The accuracy of measuring what the study intends to measure.
- **Research Question**: The central question the study aims to answer.

4.8 SELF-ASSESSMENT EXERCISE:

Multiple Choice Questions (MCQs)

1. Research design is similar to:

a) A legal document
b) An architect's blueprint
c) A mathematical formula
d) A random plan

2. Which of the following ensures findings are consistent?

a) Validity
b) Reliability
c) Ethics
d) Generalization

3. Experimental research involves:

a) Textual interpretation
b) Emotional analysis
c) Manipulation of variables
d) Undefined settings

4. Qualitative research design is:

a) Statistical and numerical
b) Predictive
c) Exploratory and descriptive
d) Concerned only with cause–effect

5. A good research design must be:

a) Biased
b) Rigid
c) Ethical
d) Ambiguous

6. Identifying independent and dependent variables comes under which step?

a) Selecting the population
b) Understanding what needs to be measured
c) Data collection
d) Publication planning

SHORT-ANSWER QUESTIONS:

- Define research design and explain its importance.
- List any four characteristics of a good research design.
- Differentiate between quantitative and qualitative research design.
- What is descriptive research design? Give an example.
- Explain the purpose of an experimental research design.
- What steps are involved in developing a research design?
- Why is objectivity important in research?
- State two differences between experimental and quasi-experimental research.
- What is meant by population in research design?
- Why is developing a clear research question important?

LONG ANSWER-QUESTIONS:

1. What is meant by research design? Explain its importance in the research process.
2. Explain the importance of reliability in research design with an example.
3. Differentiate between quantitative and qualitative research designs.
4. What are the key steps involved in developing a research design?
5. Discuss how a well-prepared research design helps to minimize waste of time, finance, and energy.
6. Elaborate on the significance of research design in producing reliable and publishable research findings.

Case Study:

A hotel chain wants to understand why customers give lower satisfaction scores during summer months. The management decides to conduct a research study. They plan to collect data from guests using questionnaires, interview housekeeping staff, compare satisfaction scores over different seasons, and test if higher temperature in rooms correlates with lower satisfaction. The researcher must choose a suitable research design and justify the method.

Questions:

1. Which type of research design is most appropriate for this study? Explain why.
2. Identify one quantitative and one qualitative data source used in this study.
3. How can the research design ensure reliability in findings?
4. Suggest one possible hypothesis for this situation.
5. Mention two steps the researcher should follow while developing this research design.

REFERENCES:

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

- Indira Gandhi National Open University (IGNOU). (n.d.). *Basics of research methodology: Unit 3 – Research design*. School of Social Sciences, Indira Gandhi National Open University, New Delhi.

4.9 FURTHER READINGS:

- Kothari, C.R. 2004. *Research Methodology Methods and Techniques,* NewAge International (P) Limited: New Delhi.
- Rao K.V. 1993. *Research Methodology in Commerce and Management,* Sterling Publishers Private Limited: New Delhi.
- Sadhu, A.N. and A. Singh, 1980. *Research Methodology in Social Sciences,* Sterling Publishers Private Limited: New Delhi.

RANGANATHAM GANGINENI

LESSON 5

UNDERSTANDING VARIABLES AND RESEARCH

5.0 OBJECTIVES OF THE LESSON

After studying this lesson, the student will be able to

1. Define and differentiate between independent and dependent variables.
2. Explain the conceptual meaning and role of endogenous and exogenous variables in research models.
3. Identify independent and dependent variables in a given research scenario.
4. Recognize sources of endogeneity and explain the importance of exogeneity in statistical modeling.
5. Differentiate between quantitative and qualitative research approaches.
6. Identify the purposes and characteristics of each type of research.
7. Describe common data collection methods used in both research types.
8. Recognize appropriate situations for using qualitative, quantitative, or mixed methods research.
9. Understand basic techniques for analyzing qualitative and quantitative data.

STRUCTURE OF THE LESSON

5.1 INTRODUCTION

In the context of research, variables refer to measurable characteristics or attributes that can assume different values, such as height, age, temperature, or test scores. Researcherssystematically manipulate or measure variables to examine cause-and-effect relationships and to draw valid conclusions based on empirical evidence.

5.2 INDEPENDENT VARIABLE

The independent variable is the variable that the researcher manipulates, controls, or categorizes to determine its influence on another variable. It represents the presumed cause in a study and is not affected by other variables within the research framework.

Alternative terms for the independent variable include:

- Explanatory variable – explains the occurrence of an event or outcome.
- Predictor variable – used to predict the value of another variable.
- Right-hand-side variable – appears on the right-hand side of a regression equation.

5.3 TYPES OF INDEPENDENT VARIABLES

(a) **1. Experimental Independent Variables**

These are variables that can be actively manipulated or controlled by the researcher in an experimental design to determine their effects on the dependent variable.

Example: A study examining the effect of a new medication on blood pressure among individuals with hypertension.

Independent Variable: Type of treatment (low-dose, high-dose, placebo) Dependent Variable: Blood pressure levels

Experimental designs often include multiple levels of the independent variable to assess both the presence and degree of its effect. Random assignment of participants to groups enhances internal validity by controlling for extraneous factors.

(b) **2. Subject (Participant) Variables**

Subject variables refer to pre-existing individual characteristics that cannot be manipulated by the researcher, such as gender, ethnicity, socioeconomic status, or educational level. When such variables are used to categorize participants, the study design becomes quasi-experimental, as random assignment is not possible.

Example: A study explores whether gender identity affects neural responses to infant cries. Independent Variable: Gender identity (men, women, others) dependent Variable: Brain activity measured through functional MRI.

In this case, gender identity serves as a naturally occurring grouping variable. Because there is no random assignment, such designs are more susceptible to research biases such as selection bias or sampling bias.

5.4 DEPENDENT VARIABLE

The dependent variable is the variable that is observed and measured to assess the impact of changes in the independent variable. It represents the effect or outcome of the manipulation.

Other terms used to describe the dependent variable include:

- • Response variable – reacts to alterations in another variable.
- • Outcome variable – indicates the result being measured.
- • Left-hand-side variable – appears on the left-hand side of a regression equation.

Illustrative Example

A researcher investigates whether variations in room temperature influence students' performance on a mathematics test

Independent Variable: Room temperature (cool vs. warm) Dependent Variable: Mathematics test scores

If test performance varies systematically with temperature changes, it may be inferred that room temperature exerts a causal influence on academic performance.

5.5 DISTINGUISHING INDEPENDENT AND DEPENDENT VARIABLES

When designing or interpreting a study, it is essential to accurately identify each type of variable. The following guiding questions can assist in this process:

- **For Independent Variables:**

1. • Is the variable manipulated, controlled, or used for participant grouping?
2. • Does it occur prior to the dependent variable in the sequence of events?
3. • Is the researcher examining whether this variable influences another variable?

- **For Dependent Variables:**

4. • Is the variable measured as the study's outcome?
5. • Does its value depend on changes in another variable
6. Is it assessed after manipulation of the independent variable?

Examples of Research Questions

Research Question	Independent Variable	Dependent Variable
Does the tomato plants grow faster under fluorescent, incandescent, or natural light?	Type of light	Growth rate of tomato plants
What is the effect of intermittent fasting on blood glucose levels?	Presence or absence of intermittent fasting	Blood glucose levels
Does remote work improve job satisfaction among employees?	Work setting (remote vs. in-office)	Self-reported job satisfaction

Analysis and Visualization of Variables

Following data collection, researchers employ descriptive and inferential statistical analyses to examine the relationship between independent and dependent variables. The choice of statistical test depends on the nature and level of measurement of the variables, and the number of levels of the independent variable. Common statistical techniques include the t-test and Analysis of Variance (ANOVA), which are used to determine whether differences between groups are statistically significant.

When presenting results graphically:

- The independent variable is plotted on the X-axis (horizontal axis).
- The dependent variable is plotted on the Y-axis (vertical axis).

Recommended visual formats**:**

- Bar charts – Suitable for categorical independent variables.
- Scatter plots or line graphs – Appropriate for continuous or quantitative variables.

5.6 ENDOGENOUS AND EXOGENOUS VARIABLES: A CONCEPTUAL AND ANALYTICAL OVERVIEW

In the domain of statistical modeling and econometrics, the accurate interpretation of relationships among variables depends critically on how these variables are classified. Distinguishing between endogenous and exogenous variables is far more than an academic formality—it forms the basis of sound regression modeling, valid causal inference, and reliable policy interpretation. Misclassification can produce biased estimators, undermine analytical credibility, and lead to erroneous conclusions.

Across disciplines such as economics, finance, social sciences, and environmental studies, understanding whether a variable is internally determined or externally imposed is essential. These classifications define the structure of a model and the causal pathways it represents. Before addressing advanced estimation techniques, it is necessary to establish precise and conceptually robust definitions for these two fundament types of variables.

Primary Classifications in models

In structural modeling, variables are categorized based on the source of their determination:

- Endogenous Variables: Variables whose values are determined within the model. They represent outcomes, effects, or dependent factors influenced by other elements of the system.
- Exogenous Variables: Variables whose values originate outside the model. They act as external inputs or drivers that affect the endogenous variables but are not influenced by the system's internal processes.

Nature and Implications of Endogenous Variables

An endogenous variable—often represented as the dependent variable (Y)—is determined jointly with other variables within the model. Its behavior is explained by the system's internal mechanisms or by simultaneous interactions with other endogenous variables. The term derives from the Greek for "originating within," reflecting its role as the key outcome or phenomenon under investigation. Endogeneity introduces methodological challenges because such variables are often correlated with the error term in a regression equation. This correlation may arise from reverse causality (where the dependent variable influences its predictors) or omitted variable bias, where an unobserved factor affects both the predictor and the outcome. Violating the assumption that regressors are uncorrelated with the error term leads to endogeneity bias, rendering ordinary least squares (OLS) estimates inconsistent.

Illustration: Suppose a researcher models a student's test score as a function of study hours. If an omitted factor—say, natural ability—influences both the amount of studying and the resulting test score, study hours become endogenous because they correlate with the error term. This correlation biases the estimated effect of study hours, overstating their true impact. Correcting this requires advanced econometric strategies that explicitly address endogeneity.

Definition and Role of Exogenous Variables In contrast, an exogenous variable (commonly denoted as X) is statistically independent of the error term and unaffected by the internal dynamics of the system. It is predetermined, fixed, or externally imposed relative to the model. Exogenous variables serve as external forces—they influence endogenous outcomes but are not influenced in return. The term "exogenous" thus means "originating externally."

Exogenous variables provide reliable explanatory power because their independence from the error term ensures unbiased and consistent coefficient estimates. Common examples include policy decisions, global commodity prices, or natural events that are outside the control of the system being analyzed. Correctly identifying such variables is crucial for establishing causal validity and preventing model misspecification.

In a properly specified model, the assumption of exogeneity justifies the use of

OLS estimation, as it guarantees that the explanatory variables are free from correlation with unobserved factors. A variable qualifies as exogenous only when it is demonstrably independent of all internal feedback mechanisms.

Consider a model in agricultural economics designed to predict total crop yield from a given plot of land:

Crop Yield = B0 + B1(Fertilizer) + B2 (Soil Type) + B3 (Rainfall)

The classification of each term depends on whether its value arises from within or outside the system:

- Crop Yield: As the primary output, crop yield is endogenous. It is determined by the specified inputs (fertilizer, soil type, and rainfall).

- Fertilizer Application: Typically endogenous, since the quantity used often depends on other factors like soil type or expected yield. It is a decision variable within the farmer's control.

- Soil Type: If the analysis considers soil improvements or amendments influenced by farmer decisions, soil type becomes endogenous. Otherwise, if it refers to natural soil characteristics, it may be exogenous.

- Rainfall: Completely exogenous, as weather conditions are external and unaffected by the farmer's actions or model variables. This example illustrates how variable classification is inherently tied to the model's boundaries and the scope of decision-making included within it.

Causality and Context Dependence The distinction between endogenous and exogenous variables fundamentally reflects the direction of causality and the boundaries of control within a model. When constructing regression models, researchers seek to determine whether an explanatory factor acts as a true cause or is merely an outcome jointly determined with the dependent variable.

Conceptually:

- Endogenous variables respond to internal changes and can be influenced or manipulated within the system. Their values are consequences of internal mechanisms.
- Exogenous variables are fixed and unaffected by the system's outcomes. They serve as external inputs that cannot be controlled by entities within the system.

Importantly, this classification is context-dependent. A factor may be exogenous in a localized model but become endogenous in a broader sy stem where feedback effects exist. Hence, researchers must clearly define model boundaries before classifying variables.

Example 1: Agricultural Economics and Crop Yield Analysis

5.7 REMEDIAL TECHNIQUES FOR ENDOGENEITY BIAS

The integrity of a regression model depends on correctly identifying endogenous and exogenous factors. When a variable is genuinely exogenous, OLS estimation produces unbiased and consistent results. However, misclassifying an endogenous variable as exogenous introduces endogeneity bias, where coefficients capture both the true causal effect and spurious correlations caused by omitted factors or reverse causality. To correct this, econometricians employ specialized estimation techniques collectively known as simultaneous equation methods:- Instrumental Variables (IV): An IV is a variable that influences the endogenous regressor but has no direct relationship with the dependent variable or the error term. It isolates the exogenous variation in the problematic predictor.

- Two-Stage Least Squares (2SLS): The standard practical implementation of IV estimation, often used when several equations or endogenous regressors exist.
- System Estimation Methods (e.g., 3SLS): Applied to complex systems of equations with correlated error terms, ensuring efficient and consistent estimation across multiple relationships.

QUANTITATIVE RESEARCH AND QUALITATIVE RESEARCH

5.8 QUANTITATIVE RESEARCH

Quantitative research involves the collection and analysis of numerical data. It is used to test or confirm hypotheses and theories, and to draw generalizable conclusions. The approach emphasizes objectivity, precision, and measurement.

Common Quantitative Methods:

1. Surveys: Structured questionnaires with closed-ended or multiple-choice questions.
2. Experiments: Controlled studies that manipulate variables to determine cause-and-effect relationships.
3. Observations: Systematic recording of behaviors or events in numerical form.

Common Biases: Information bias, sampling bias, omitted variable bias, and selection bias.

5.9 QUALITATIVE RESEARCH

Qualitative research seeks to explore human experiences, perceptions, and social contexts. It focuses on understanding meanings rather than quantifying variables. Data are typically collected in the form of words, images, or texts.

Common Qualitative Research Methods

1. **Interviews**: Use open-ended questions to obtain in-depth, personal insights from participants.
2. **Focus Groups**: Facilitated group discussions aimed at collecting diverse perspectives on a specific topic.
3. **Ethnography**: Involves prolonged, immersive observation of a community or culture to understand social practices and behaviors.
4. **Literature Review**: A systematic and critical analysis of existing scholarly publications related to the research topic.

Common Biases in Qualitative Research:
Hawthorne effect, observer bias, recall bias, and social desirability bias.

5.10 DIFFERENCES BETWEEN QUANTITATIVE AND QUALITATIVE RESEARCH

Feature	Quantitative Research	Qualitative Research
Nature of Data	Numerical data	Descriptive data (words, texts, visuals)
Purpose	To test or verify hypotheses	To explore and understand experiences and meanings
Approach	Deductive	Inductive
Outcome	Statistical results and measurable facts	Thematic findings and conceptual insights
Data Analysis	Statistical tools (SPSS, Excel, R)	Thematic or content analysis
Sample Size	Large and representative	Small and purposive
Researcher's Role	Objective and detached	Subjective and interpretive

---

5.11 DATA COLLECTION METHODS

Certain data collection methods may be applied in both quantitative and qualitative research, depending on the nature of data collection and analysis.

Quantitative Methods:

1. Surveys
2. Experiments
3. Structured observations

Qualitative Methods:

1. Interviews
2. Focus groups
3. Ethnography
4. Literature review

5.12 WHEN TO USE QUALITATIVE VS. QUANTITATIVE RESEARCH

A general guideline:

1. Quantitative research is used when the objective is to test or confirm theories or hypotheses.
2. Qualitative research is used when the goal is to explore ideas, experiences, or meanings.

Example:

Research Question: ***How satisfied are students with their studies?***

1. Quantitative: Survey 300 students using a 5-point satisfaction scale; analyze using statistics.
2. Qualitative: Interview 15 students to explore their views and experiences in depth.
3. Mixed Methods: Combine both approaches—use interviews for insights, then surveys to test those insights on a larger scale.

5.13 DATA ANALYSIS TECHNIQUES

Quantitative Analysis
Quantitative data are analyzed using mathematical and statistical procedures. Common analyses include:

- Mean, median, mode
- Frequency counts
- Correlation and regression analysis
- Tests of reliability and validity

Qualitative Analysis
Qualitative data require interpretive methods to derive meaning:

- Content Analysis: Identifying frequency and context of key words or phrases.
- Thematic Analysis: Finding recurring themes and patterns.
- Discourse Analysis: Examining how language and communication shape meaning.

5.14 SUMMARY

The decision to model a variable as endogenous or exogenous fundamentally shapes the analytical strategy and the credibility of causal conclusions. Only through precise classification can researchers ensure that regression results reflect the true causal mechanisms operating within a system, avoiding bias and enhancing the reliability of empirical findings. A clear understanding of independent and dependent variables is fundamental to sound research design and analysis. Proper identification and control of these variables ensure that observed relationships are valid, reliable, and interpretable within the context of the study. Variables are measurable characteristics that change or vary among individuals, groups, or situations. The independent variable is manipulated or controlled to examine its influence on another variable, while the dependent variable is observed to assess that influence. Experimental variables are directly manipulated by researchers, whereas subject variables are pre-existing characteristics of participants. Endogenous variables are determined within a model and are often correlated with error terms, leading to potential bias if not corrected. Exogenous variables originate outside the model and are assumed to be unaffected by internal system dynamics. Proper identification of variables is essential for achieving valid causal inference and reliable statistical conclusions. Quantitative and qualitative research represent two fundamental approaches to scientific inquiry. Quantitative research emphasizes measurement, objectivity, and generalizability, while qualitative research emphasizes depth, meaning, and context. The integration of both approaches, known as *mixed methods research*, offers a comprehensive understanding of complex research problems by combining numerical precision with interpretive insight.

5.15 TECHNICAL TERMS

Variable: Any characteristic or attribute that can vary among subjects or conditions.

Independent Variable (IV): The variable that is manipulated or controlled to test its effect on another variable.

Dependent Variable (DV): The variable that is measured to assess the effect of changes in the independent variable.

Experimental Variable: An independent variable that is directly manipulated by the researcher.

Subject Variable: A pre-existing characteristic of participants used to categorise groups (e.g., gender, age).

Endogenous Variable: A variable whose value is determined within the model and potentially correlated with the error term.

Exogenous Variable: A variable determined outside the model; it acts as an external input not influenced by the system.

Endogeneity Bias: The distortion in estimates caused by a correlation between explanatory variables and error terms.

Instrumental Variable (IV); A variable used in regression analysis to correct for endogeneity by isolating exogenous variation.

Two-Stage Least Squares (2SLS): An econometric method used to estimate parameters when endogeneity is present.

Quantitative Research: Research focused on numerical data and statistical analysis. Qualitative Research: Research that explores meanings, experiences, and perspectives. Mixed Methods: A combination of quantitative and qualitative approaches in one study. Survey: A structured data collection tool using questions or scales.

Experiment: Research involving the manipulation of variables to test causal relationships. Interview: A method of gathering in-depth information through verbal questioning.

Thematic Analysis: Identifying and analysing patterns or themes within qualitative data. Sampling Bias: Error arising from the non-representative selection of participants.

5.16 SELF-ASSESSMENT QUESTIONS

Section A – Short Answer Questions

1. Define independent and dependent variables and provide one example of each.
2. Distinguish between experimental and subject variables with suitable illustrations.
3. What are endogenous variables? Explain with one example.
4. What is meant by exogeneity in econometric models?
5. Write any two differences between endogenous and exogenous variables.
6. Define quantitative research and state its primary purpose.
7. List any three common methods of qualitative data collection.
8. What types of bias may affect quantitative research?
9. Distinguish between content analysis and thematic analysis.
10. What is a mixed methods approach?

Section B – Essay/Long Answer Questions

1. Explain the role of independent and dependent variables in experimental research.
2. Discuss how misclassification of endogenous and exogenous variables affects regression outcomes.
3. Describe the types of independent variables and their significance in research design.
4. Explain the econometric strategies used to address endogeneity bias.
5. Compare and contrast independent–dependent variables with endogenous–exogenous variables, using relevant examples.
6. Compare and contrast qualitative and quantitative research approaches with suitable examples.
7. Discuss the major techniques used in collecting and analyzing quantitative data.
8. Explain the importance of qualitative research in understanding human experiences.

9. Evaluate the advantages and limitations of mixed methods research.
10. Illustrate how a researcher can decide when to use qualitative or quantitative methods.

5.17 SUGGESTED READINGS

- Gujarati, D. N., & Porter, D. C. (2020). Basic Econometrics (6th ed.). McGraw-Hill Education.
- Field, A. (2018). Discovering Statistics Using IBM SPSS Statistics (5th ed.). Sage Publications.
- Kothari, C. R. (2014). Research Methodology: Methods and Techniques (3rd ed.). New Age International Publishers.
- Wooldridge, J. M. (2016). Introductory Econometrics: A Modern Approach (6th ed.). Cengage Learning.
- Creswell, J. W., & Creswell, J. D. (2018). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (5th ed.). Sage Publications.
- Creswell, J. W. (2014). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. Sage Publications.
- Flick, U. (2018). *An Introduction to Qualitative Research*. Sage Publications.
- Bryman, A. (2016). *Social Research Methods*. Oxford University Press.
- Neuman, W. L. (2014). *Social Research Methods: Qualitative and Quantitative Approaches*. Pearson Education.
- Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods*. Sage Publications.

Dr. S. ANITHA DEVI

LESSON 6

RELIABILITY, VALIDITY & LEVELS OF MEASUREMENT

OBJECTIVES OF THE LESSON

After studying this lesson, students will be able to:

1. Explain the concepts of validity and reliability and their significance in ensuring accurate and consistent research results.
2. Identify different types of validity (content, construct, criterion-related) and reliability (test-retest, inter-rater, internal consistency).
3. Describe the four levels of measurement — nominal, ordinal, interval, and ratio — and understand their distinct characteristics.
4. Classify research variables according to the appropriate level of measurement.
5. Apply the appropriate level of measurement to various examples in psychological, educational, or social research contexts.
6. Interpret how measurement levels influence the choice of statistical tools and data analysis techniques.

STRUCTURE OF THE LESSON

6.1 INTRODUCTION

In research, the processes of measurement and scaling are essential for the scientific study of abstract concepts. Numerous variables in fields such as social science, psychology, education, and management — including attitude, motivation, satisfaction, and performance — cannot be directly observed or quantified. To investigate these variables effectively, researchers are required to measure them using systematic methods and to represent their intensity or degree through scales. This is accomplished through measurement, which entails the systematic assignment of numbers or symbols to indicate the presence or degree of a specific attribute, and through scaling, which offers a structured approach to express and compare these measurements along a continuum. Consequently, measurement and scaling form the cornerstone of all empirical research, bridging the divide between theoretical concepts and observable data. They enable researchers to articulate human behavior, perceptions, and attitudes in numerical terms, facilitating precise analysis, comparison, and interpretation. In the absence of appropriate measurement and scaling techniques, it would be unfeasible to test hypotheses, validate theories, or generalize findings in a scientific manner.

6.2 MEASUREMENT

Measurement serves as the necessary process of systematically assigning numbers or symbols to various characteristics, traits, or behaviors in accordance with established rules. This process enables researchers to quantify abstract or intangible concepts such as intelligence, motivation, satisfaction, or performance, thereby rendering them suitable for scientific analysis. Measurement acts as a bridge between theory and observation by transforming qualitative attributes into quantitative data that can be compared, analyzed, and interpreted in an objective manner. Measurement entails the assignment of numerical values or symbols to the characteristics or attributes of objects, individuals, or events based on specific criteria. Within the context of research, measurement is vital for quantifying abstract concepts like intelligence, satisfaction, or motivation, which facilitates statistical analysis. It ensures that the data collected is objective, consistent, and comparable. The core components of measurement include defining the measurement target, selecting appropriate indicators, and ensuring both validity (accuracy) and reliability (consistency). Reliability and validity are pivotal concerns in all measurement practices. Both concepts relate to the connection between measures and constructs. The significance of reliability and validity arises from the fact that constructs are often ambiguous, diffuse, and not directly observable. Achieving perfect reliability and validity is exceedingly challenging. These two critical aspects of research design will be explored in this unit. All researchers aspire for their measures to be both reliable and valid, as these concepts contribute to establishing the truthfulness, credibility, and believability of research findings.

RELIABILITY

Meaning of Reliability

The concept of reliability asserts that any significant findings must be reproducible. Other researchers should be able to conduct the identical experiment under the same conditions and achieve the same outcomes. This replication will validate the findings and ensure that the hypothesis is accepted by the scientific community. In the absence of this replication of statistically significant results, the experiment and research do not meet all the criteria for testability. This condition is crucial for a hypothesis to establish itself as a recognized scientific truth. For instance, when conducting a time-sensitive experiment, a stopwatch is typically employed. It is generally reasonable to presume that the instruments are dependable and will accurately measure time. Nevertheless, scientists often take multiple measurements to reduce the likelihood of malfunction and to uphold validity and reliability. Conversely, any experiment that relies on human judgment is likely to face scrutiny. Human judgment can fluctuate, as individual observers may assess situations differently based on the time of day and their current emotional state. This variability implies that such experiments are more challenging to replicate and are inherently less reliable. Reliability is a vital component in assessing the overall validity of a scientific experiment and in strengthening the results. Reliability refers to the consistency of measurements, or the extent to which an instrument yields the same results each time it is utilized under identical conditions with the same subjects. In essence, it pertains to the repeatability of measurements. A measure is deemed reliable if an individual's score on the same test administered twice is comparable. It is crucial to note that reliability is not directly measured; rather, it is estimated. For example, if a test is designed to assess a specific trait, such as neuroticism, it should produce consistent results each time it is administered. A test is regarded as reliable if it consistently yields the same outcome.According to Anastasi (1957), test reliability pertains to the consistency of scores achieved by an individual across various occasions or with different sets of equivalent items. According to Stodola and Stordahl (1972), test reliability can be characterized as the correlation between two or more sets of scores from equivalent tests administered to the same group of individuals. According to Guilford (1954), reliability represents the ratio of the true variance in the scores obtained from tests. The concept of test reliability can also be viewed from a different perspective. Whenever a measurement is taken, it inherently involves some form of assessment. The error of measurement typically exists between the true scores and the observed scores. However, in psychological terms, the term 'error' does not necessarily indicate that a mistake has occurred. In other words, an error in psychological testing signifies that there is always a degree of inaccuracy in the measurement process. Therefore, the objective of psychological measurement is to ascertain the extent of such errors and to devise methods to reduce them.

6.4 METHODS OF ESTIMATING RELIABILITY

There are number of ways of estimating reliability of an instrument. Various procedures can be classified into two groups:

- External consistency procedures
- Internal consistency procedures

6.4.1 External Consistency Procedures

External consistency procedures compare findings from two independent process of data collection with each other as a means of verifying the reliability of the measure. Two methods are as beneath.

6.4.1.1 Test Re-test Reliability

The most frequently used method to find the reliability of a test is by repeating the same test on same sample, on two different time periods. The reliability coefficient in this case would be the correlation between the score obtained by the same person on two administrations of the test.

Test-Retest reliability is assessed when the same test is given to the same sample. Consequently, it pertains to the consistency of a test across two different time periods and various administrations. The underlying assumption of this method is that there will be no significant alterations in the measurement of the construct in question when administered on different occasions. The interval between measurements is of paramount importance; a shorter time gap results in a higher correlation value, and conversely, a longer gap leads to a lower correlation. If the test is deemed reliable, the scores obtained during the first administration should closely resemble those achieved during the second administration. The correlation between the two administrations should be strongly positive. Limitations of this approach There are several limitations to consider, which include the following: (i) Memory Effect/Carry Over Effect, (ii) Practice Effect, and (iii) Absence. These limitations are discussed in detail below:

i) **Memory effect /carry over effect:** One of the common problems with test- retest reliability is that of memory effect. This argument particularly holds true when, the two administrations takes place within short span of time, for example, when a memory related experiment including nonsense syllables is conducted whereby, the subjects are asked to remember a list in a serial wise order, and the next experiment is conducted within 15 minutes, most of the times, subject is bound to remember his/her responses, as a result of which there can be prevalence of artificial reliability coefficient since subjects give response from memory instead of the test. Same is the condition when pre-test and post-test for a particular experiment is being conducted.

ii) **Practice effect:** This happens when repeated tests are being taken for the improvement of test scores, as is typically seen in the case of classical IQ where there is improvement in the scores as we repeat these tests.

iii) **Absence:** People remaining absent for re-tests.

6.4.1.2 Parallel Forms Reliability

Parallel-Forms Reliability is known by the various names such as Alternate forms reliability, equivalent form reliability and comparable form reliability.

Parallel forms reliability compares two equivalent forms of a test that measure the same attribute. The two forms use different items. However, the rules used to select items of a particular

difficulty level are the same. When two forms of the test are available, one can compare performance on one form versus the other. Sometimes the two forms are administered to the same group of people on the same day.

The Pearson product moment correlation coefficient is used as an estimate of the reliability. When both forms of the test are given on the same day, the only sources of variation are random error and the difference between the forms of the test. Sometimes the two forms of the test are given at different times. In these cases, error associated with time sampling is also included in the estimate of reliability.

The method of parallel forms provides one of the most rigorous assessments of reliability commonly in use. Unfortunately the use of parallel forms occurs in practice less often than is desirable. Often test developers find it burdensome todevelop two forms of the same test, and practical constraints make it difficult to retest the same group of individuals. Instead many test developers prefer to base their estimate or reliability on a single form of a test.. More often they have only one test form and must estimate the reliability for this single group of items. You can asses the different sources of variation within a single test in many ways. One method is to evaluate the internal consistency of the test by dividing it into subcomponents.

6.4.2 Internal Consistency Procedures

The idea behind internal consistency procedures is that items measuring same phenomena should produce similar results. Following internal consistency procedures are commonly used for estimating reliability-

6.4.2.1 Split Half Reliability

In this method, as the name implies, we randomly divide all items that intends to measure same construct into two sets .The complete instrument is administered on sample of people and total scores are calculated for each randomly divided half; the split half reliability is then, the simply the correlation between these two scores.

Problem in this approach

A problem with this approach is that when the tests are shorter, they run the risk of losing reliability and it can most safely be used in case of long tests only. It is, hence, more useful in case of long tests as compared to shorter ones. However to rectify the defects of shortness, Spearman- Brown's formula can be employed, enabling correlation as if each part were full length:

$r = (2rhh)/(1 + rhh)$ (Where rhh is correlation between two halves)

6.4.2.2 Cronbach's Alpha (□)

As proposed by Cronbach (1951) and subsequently elaborated by others (Novick & Lewis, 1967; Kaiser & Michael, 1975), coefficient alpha may be thought of as the mean of

all possible split-half coefficients, corrected by the Spearman-Brown formula .The formula for coefficient alpha is

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right)$$

Where

- **α** = Cronbach's Alpha (coefficient of reliability or internal consistency)
- **k** = number of items (questions) in the test or scale
- σ_i^2 = variance of each individual item
- σ_t^2 = variance of the total test score (sum of all items)

Interpretation:

- **α ≥ 0.9** → Excellent reliability
- **0.8 ≤ α < 0.9** → Good reliability
- **0.7 ≤ α < 0.8** → Acceptable reliability
- **0.6 ≤ α < 0.7** → Questionable reliability
- **0.5 ≤ α < 0.6** → Poor reliability
- **α < 0.5** → Unacceptable reliability

Cronbach's Alpha measures **internal consistency**, that is, how closely related a set of items are as a group. It indicates whether the items in a scale measure the same underlying construct

Uses of Cronbach's Alpha

- Measures internal consistency: It helps determine how well items in a scale measure the same construct.
- Simple to calculate: Easy to compute using statistical software like SPSS, R, or Excel.
- Widely accepted: A standard and recognized method for assessing reliability in social sciences and psychology.
- Useful for scale development: Helps researchers refine questionnaires by identifying weak items.
- Quantitative measure: Provides a single numeric value to represent the reliability of a test.

Problems with Cronbach's Alpha

- Assumes equal item correlation: It assumes all items contribute equally, which may not be true in real data.
- Affected by number of items: Adding more items can artificially increase alpha, even if items are redundant.

- Does not confirm unidimensionality: A high alpha doesn't guarantee that all items measure only one construct.
- Sensitive to item variance: If item variances differ greatly, the value may be misleading.
- Not suitable for all scales: It works best for continuous data and may not be ideal for dichotomous (yes/no) items.

6.5 VALIDITY

As you know that the merit of the psychological test is determine first by its reliability but then ultimately by its validity. Validity refers to the degree to which a test measures, what it claims to measure. It is very necessary for a test to be valid for its proper administration and interpretation.

According to Standard for Educational and Psychological testing (AERA, APA & NCME 1985, 1999); a test is valid to the extent that inferences drawn from it are appropriate, meaningful and useful.

According to Cronbach (1951) validity is the extent to which a test measures what it purports to measure.

According to Freeman (1971) an index of validity shows the degree to which a test measures what it purports to measure when compared with accepted criteria.

According to Anastasi (1988) the validity of a test concerns what the test measures and how well it does so.

The above definitions pointed out that for determining the validity of the test, the test must be compared with some ideal independent measures or criteria. The correlation coefficients computed between the test and an ideal criterion is known as the validity coefficients. Independent criteria refer to some measure of the trait or group of the traits (out side the test) that the test itself claims to measure.

6.6 TYPES OF VALIDITY

There are six types of validity, viz.,

(i) Content validity (ii) Criterion-related validity (iii) Con current validity (iv) Predictive validity (v) Construct validity(vi) Convergent validity (vii) Discriminate validity and (viii) Face validity. These are being discussed below:

6.6.1 Content Validity

According to Mc Burney and White (2007); content validity is the notion that a test The sample should encompass a range of behaviors that align with the theoretical concept being assessed. This form of validity is non-statistical and involves evaluating the content of the test to determine if it adequately represents the behavior intended for measurement. When a test

exhibits content validity, its items reflect the complete spectrum of potential items that the test is designed to address. For example, if a researcher aims to create a spelling achievement test for third-grade students, they would identify nearly all the words that these children are expected to know. Individual test items may be selected from a vast pool that includes a wide variety of items. A test inherently possesses content validity. Items are chosen based on their alignment with the test's requirements following a thorough review of the subject matter. In some instances, when a test evaluates a trait that is challenging to define, an expert may assess the relevance of the items. Given that each judge has their own perspective on their evaluations, two independent judges will assess the test separately. Items deemed highly relevant by both judges will be included in the final version of the test

6.6.2 Criterion-related Validity

Criterion-related validity refers to the principle that a valid test should closely correlate with other measures of the same theoretical concept. A valid intelligence test should show a strong correlation with other intelligence assessments. If a test effectively predicts criteria or indicators of the construct, it is considered to have criterion-related validity. There are two distinct types of criterion validity.-

6.6.2.1 Concurrent Validity

Its occurrence is found when criterion measures are achieved at the same time as the test scores. It reflects the degree to which the test scores estimate the individual's present status with regards to criterion. For instance, if a test measures anxiety, it would be said to have concurrent validity if it rightly reflects the current level of anxiety experienced by an individual. Concurrent evidence of test validity is usually desirable for achievement tests and diagnostic clinical test.

6.6.2.2 Predictive Validity

Predictive validity occurs when criterion measures are obtained at a time after the test. For example, aptitude tests are useful in identifying who will be more likely to succeed or fail in a particular subject. Predictive validity is part curly relevant for entrance examination and occupational test.

6.6.3 Construct Validity

Construct validity approach is complex than other forms of validity. Mc Burney and White (2007) defined construct validity as the property of a test that the measurement actually measures the constructs they are designed to measure. There are several ways to determine whether a test generate data that have construct validity.

i) The test should actually measure whatever theoretical construct it supposedly tests, and not something else. For example a test of leadership ability should not actually test extraversion.
ii) A test that has construct validity should measure what it intends to measure but not measure theoretically unrelated constructs. For example, a test of musical aptitude should not require too much reading ability.

iii) A test should prove useful in predicting results related to the theoretical concepts it is measuring. For example, a test of musical ability should predict who will benefit from taking music lessons, should differentiate groups who have chosen music as a career from those who haven't should relate to other tests of musical ability and so on.

There are two types of construct validity— 'convergent validity' and 'divergent validity' (or discriminant validity).

6.6.3.1 Convergent Validity

It means the extent to which a measure is correlated with other measure which is theoretically predicted to correlate with.

6.6.3.2 Discriminant Validity

This explains the extent to which the operationalisation is not correlated with other operationalisations that it theoretically should not be correlated with.

6.6.4 Face Validity

Face validity refers to what appears to measure superficially. It depends on the judgment of the researcher. Each question is scrutinised and modified until the researcher is satisfied that it is an accurate measure of the desired construct. The determination of face validity is based on the subjective opinion of the researcher.

Exercise

Fill in the blanks

1) If a test measures what it purports to measure it is caleed …….
2) If a test is correlated against a criterion to be made available at the present time it is a type of validity known as…..... validity.
3) The property of a test that measurement actually measure the constructs they are design to measure are known as…..validity
4) A test should sample the range of behaviour represented by the theoretical concept being tested, is known as ..... validity.
5).....refers to what appears to measure superficially.

Answers: (1) Validity (2) Criterion Validity (3) Construct (4) Content

(5) Face Validity

6.6.5 Internal Validity

Internal validity is the most fundamental type of validity because it concerns the logic of the relationships between the independent variable and dependent variable. This type of

validity is an estimate of the degree to which inferences about causal relationship can be drawn, based on the measures employed and research design. Properly suited experimental techniques, where the effect of an independent variable upon the dependent one is observed under highly controlled conditions makes possible higher degree of internal validity.

6.6.5.1 Threats to Internal Validity

These include (i) confounding, (ii) selection bias, (iii) history, (iv) maturation, (v) repeated testing ,(vi) instrument change, (vii) regression toward the mean,(viii) mortality, (ix) diffusion, (x) compensatory rivalry, (xi) experimenter bias.

i) Confounding: A confounding error arises when the effects of two variables in an experiment cannot be distinguished, leading to a muddled interpretation of the findings. Confounding represents one of the most significant threats to the validity of experimental research. This issue is particularly pronounced in studies where the experimenter lacks control over the independent variable. When participants are chosen based on the presence or absence of a condition, subject variables can influence the outcomes. In situations where a spurious relationship cannot be avoided, an alternative hypothesis may be proposed in relation to the original cause and inference hypotheses.

ii) Selection bias: Any bias in the selection of a group can compromise internal validity. Selection bias refers to the issues that arise due to pre-existing differences between groups prior to testing, which may interact with the independent variable and subsequently affect the observed results, leading to complications. Examples of such biases include gender, personality traits, mental capabilities, physical abilities, motivation levels, and willingness to participate. If, at the time of selection, an unequal number of subjects with similar subject-related variables are tested, this could pose a threat to internal validity. For instance, if two groups are established, namely an experimental group and a control group, and the subjects in these groups differ concerning the independent variable but are similar in one or more subject-related variables, it becomes challenging for the researcher to ascertain whether the differences between the groups stem from the independent variable or from subject-related variables, as well as the randomization of group assignment. This randomization is not always feasible, as some significant variables may remain unnoticed.

iii) History: External events or occurrences between repeated measures of dependent variables may affect participants' responses, attitudes, and behaviors during the experimental process, such as natural disasters or political changes. Under these circumstances, it becomes impossible to determine whether change in dependent variable is caused by independent variable or historical event.

iv) *Maturation:* Usually, it happens that subjects change during the course of an experiment or between measurements. For instance, in longitudinal studies young kids might grow up as a result of their experience, abilities or attitudes which are intended to be measured. Permanent changes [such as physical growth] and temporary changes [like fatigue and illness] may alter the way a subject would react to the independent variable. Thus, researcher may have trouble in ascertaining if the difference is caused

by time or other variables.

v) *Repeated testing*: Participants may be driven to bias owing to repeated testing. Participants may remember correct answers or may be conditioned as a result of incessant administration of the test. Moreover, it also causes possibility of threat to internal validity.

vi) *Instrument change*: If any instrument is replaced/changed during process of experiment, then it may affect the internal validity as alternative explanation easily available.

vii) Regression toward the mean: During the course of the experiment, if participants are selected based on extreme scores, there is a possibility of encountering such an error. For instance, when individuals with the lowest mathematical abilities are chosen, any improvement observed at the conclusion of the study may likely be attributed to regression toward the mean rather than the effectiveness of the course.

viii) Mortality: It is important to consider that some participants may have withdrawn from the study prior to its completion. If the dropout of participants introduces significant bias between groups, an alternative explanation may account for the observed differences.

ix) Diffusion: It may be noted that a lack of differences between the experimental and control groups could occur if the treatment effects disseminate from the treatment groups to the control groups. However, this does not imply that the independent variable has no effect or that there is no relationship between the dependent and independent variables. Compensatory rivalry/resentful demoralisation: The behavior of subjects may change if the control groups are influenced as a result of the study. For example, participants in the control group may exert additional effort to ensure that the anticipated superiority of the experimental group is not demonstrated. Nevertheless, this does not suggest that the independent variable had no effect or that there is no relationship between the dependent and independent variables. Conversely, changes in the dependent variable may solely result from a demoralised control group that is less motivated or working less diligently.

x) Experimenter bias: Experimenter bias occurs when experimenters, whether intentionally or unintentionally, behave differently towards participants in the control and experimental groups, which subsequently affects the results of the experiment. This bias can be mitigated by ensuring that the experimenter is unaware of the experimental conditions or its purpose and by standardising the procedure to the greatest extent possible.

6.6.6.External Validity

According to McBurney and White (2007), external validity pertains to whether the findings of research can be applied to different situations, subjects, settings, times, and so forth. External validity is compromised by the fact that studies involving human participants frequently utilize small samples gathered from specific geographic areas or possess unique characteristics (e.g., volunteers). Consequently, it cannot be guaranteed that the conclusions regarding cause-and-effect relationships are relevant to individuals in other geographic regions or in the absence of these characteristics.

6.6.6.1 Threat to External Validity

One of the primary threats to external validity is the potential for errors in making generalizations. Typically, generalizations are constrained when the cause (i.e., independent variable) is influenced by other factors; thus, all threats to external validity interact with the independent variable. a) Aptitude-Treatment-Interaction: The sample may possess certain characteristics that interact with the independent variable, thereby restricting generalizability. For example, conclusions derived from comparative psychotherapy studies often rely on specific samples (e.g., volunteers, individuals with severe depression, hardcore criminals). b) Situations: Various situational factors, such as treatment conditions, lighting, noise, location, experimenter, timing, and the scope and degree of measurement, may restrict generalizations. c) Pre-Test Effects: When cause-and-effect relationships can only be identified after conducting pre-tests, this also tends to limit the generalizability of the findings. d) Post-Test Effects: When cause-and-effect relationships can only be examined after conducting post-tests, this can also serve as a limitation on the generalizability of the findings. e) Rosenthal Effects: When conclusions drawn from cause-and-effect relationships cannot be generalized to other investigators or researchers.

1) Results can not be generalised to another situation or population in external Validity. T/F
2) Dropping out of some subjects before an experiment is completed causing a threat to internal validity. T/F
3) Any bias in selecting the groups can enhance the internal validity. T/F
4) Internal Validity concern the logic of relationship between the independent variable and dependent variable. T/F
5) Confounding error occurs when the effects of to variable in an experiment can not be separated. T/F

Answers: (1) F, (2) T, (3) F, (4) T, (5) T

6.7. SCALING

Scaling pertains to the method of establishing a continuum on which the measured objects or individuals are positioned. It assesses the degree or intensity of a particular attribute. Various types of scales exist, including nominal, ordinal, interval, and ratio, each providing distinct levels of information and analytical opportunities. For example, nominal scales categorize data into distinct groups, whereas ratio scales offer precise measurements that include a true zero point.Critical to all research is the development and measurement of variables. The measurement of variables allows researchers to assess the nature of relationships (in nonexperimental research) and the effects of manipulations (in experimental research). These measured variables are a key in separating the sciences from approaches such as religion and philosophy that do not systematically measure outcomes.

Exactly how the measurement is carried out depends on the type of variable involved in the research study. Different types are measured differently. To measure the time taken to respond to a stimulus, you might use a stop watch. Stop watches are of no use, of course, when it comes to measuring someone's attitude towards a political candidate. A rating scale is more appropriate in this case (with labels like "very favorable," "somewhat favorable," etc.). For a

variable such as "favourite color," you can simply note the colour-word (like "red") that the subject offers.

Although procedures for measurement differ in many ways, they can be classified using a few fundamental categories. In a given category, all of the procedures share some properties that are important for you to know about. The categories are called "scale types," or just "scales," and are briefly described in this section. As an organizing principle, you should know that measurements can be grouped into four scales, from simplest to more sophisticated: Nominal, Ordinal, Interval, and Ratio. Each scale includes the characteristics of hte preceding scale plus one additional quality. Let's explore each scale now.

6.7.1 Nominal scales

When utilizing a nominal scale for measurement, responses are merely named or categorized. The values attributed to variables lack any intrinsic numerical significance; they serve solely as descriptive labels. Examples of variables assessed on a nominal scale include gender, handedness, favorite color, and religion. A key aspect of nominal scales is that they do not suggest any hierarchy among the responses. For instance, when sorting individuals based on their preferred color, there is no implication that green is ranked "before" blue. Responses are simply classified. Nominal scales represent the most basic level of measurement.

6.7.2. Ordinal scales

A researcher aiming to assess consumer satisfaction with their microwave ovens may request that they express their feelings using the options "very dissatisfied," "somewhat dissatisfied," "somewhat satisfied," or "very satisfied." The items within this scale are arranged in order, from least to most satisfied. Such scales are referred to as "Likert scales" and are extensively utilized across various research domains. This characteristic differentiates ordinal scales from nominal scales. Unlike nominal scales, ordinal scales facilitate comparisons regarding the extent to which two individuals evaluate the variable. For instance, the satisfaction hierarchy allows for a meaningful assertion that one individual is more satisfied than another with their microwave ovens. This assertion is based on the first individual's selection of a verbal label that appears later in the list compared to the label chosen by the second individual. Conversely, ordinal scales do not capture significant information that may be present in other scales we will explore. Specifically, the difference between two levels of an ordinal scale cannot be assumed to be equivalent to the difference between two other levels. In our satisfaction scale, for instance, the difference between the responses "very dissatisfied" and "somewhat dissatisfied" is likely not the same as the difference between "somewhat dissatisfied" and "somewhat satisfied." Our measurement procedure does not provide a means to ascertain whether these two differences reflect an identical difference in psychological satisfaction. Statisticians articulate this concept by stating that the differences between adjacent scale values do not necessarily signify equal intervals on the underlying scale that generates the measurements. (In this context, the underlying scale represents the true feeling of satisfaction that we are attempting to quantify.).

6.7.3 Interval scales

Interval scales are numerical scales where intervals maintain a consistent interpretation throughout. For instance, consider the Fahrenheit or Celsius temperature scales. The difference between 30 degrees and 40 degrees signifies the same temperature difference as that between 80 degrees and 90 degrees. This consistency arises because each 10-degree interval conveys the same physical meaning in terms of molecular kinetic energy. However, interval scales are not without flaws. Notably, they lack a true zero point, even if one of the values on the scale is labeled as 'zero.' The Fahrenheit scale exemplifies this problem. Zero degrees Fahrenheit does not indicate a complete absence of temperature, which would imply no molecular kinetic energy. In fact, the designation of 'zero' in this context is attributed to historical factors related to temperature measurement. Since an interval scale does not possess a true zero point, calculating temperature ratios becomes nonsensical. For instance, the ratio of 40 to 20 degrees Fahrenheit cannot be equated to the ratio of 100 to 50 degrees, as no significant physical property is maintained across these ratios. If the 'zero' label were assigned to the temperature that Fahrenheit designates as 10 degrees, the ratios would instead be 30 to 10 and 90 to 40, which are not equivalent! Consequently, it is illogical to assert that 80 degrees is 'twice as hot' as 40 degrees. Such a statement relies on an arbitrary choice regarding the starting point of the temperature scale, specifically which temperature is designated as zero, whereas the assertion aims to convey a more fundamental truth about the underlying physical reality

6.7.4 Ratio scales

The ratio scale of measurement is the most informative scale. It is an interval scale that possesses the additional characteristic of having a zero point which signifies the absence of the quantity being measured. One can conceptualize a ratio scale as a combination of the three preceding scales. Similar to a nominal scale, it assigns a name or category to each object (with the numbers acting as labels). In the same vein as an ordinal scale, the objects are arranged in order (based on the sequence of the numbers). Furthermore, akin to an interval scale, the difference between two points on the scale retains the same significance. However, additionally, the same ratio at two different points on the scale also conveys the same meaning. The Fahrenheit scale for temperature features an arbitrary zero point, thus it does not qualify as a ratio scale. Conversely, zero on the Kelvin scale represents absolute zero, which qualifies the Kelvin scale as a ratio scale. For instance, if one temperature is measured to be twice as high as another on the Kelvin scale, it indicates that it possesses twice the kinetic energy of the other temperature. Another illustration of a ratio scale is the amount of money one currently possesses (such as 25 cents, 50 cents, etc.). Money is quantified on a ratio scale because, in addition to exhibiting the characteristics of an interval scale, it includes a true zero point: having zero money genuinely signifies the absence of money. Given that money has a true zero point, it is logical to assert that an individual with 50 cents possesses twice the amount of money as someone with 25 cents.

Summary of the Levels of Measurement

SCALE	Names or	Order or	Measurements	Proportional

	Categories	Rankings	with arbitrary zero	measurements with absolute zero
NOMINAL	**X**			
ORDINAL	**X**	**X**		
INTERVAL	**X**	**X**	**X**	
RATIO	**X**	**X**	**X**	**X**

6.8 SUMMARY

Measurement in research is the systematic process of assigning numbers or symbols to characteristics or attributes of people, objects, or events according to defined rules. It bridges the gap between theory and observation, making abstract concepts like intelligence, satisfaction, or motivation quantifiable and analyzable.

Two vital properties of measurement are **reliability** and **validity**.

- **Reliability** refers to the consistency or stability of measurement results over time or across items. A reliable instrument yields similar results when repeated under identical conditions. It can be estimated through methods such as **test–retest**, **parallel forms**, **split-half reliability**, and **Cronbach's Alpha**. Cronbach's Alpha provides a measure of internal consistency, showing how closely related a set of items are in a scale.
- **Validity** indicates how accurately a test measures what it is intended to measure. Types include **content validity**, **criterion-related validity** (concurrent and predictive), **construct validity** (convergent and discriminant), and **face validity**. **Internal validity** ensures that the observed effects are due to the independent variable, not extraneous factors, while **external validity** concerns the generalizability of findings to other populations, settings, and times.

Measurement scales differ in the degree of precision and information they provide:

1. **Nominal scale** – categorizes data without order (e.g., gender, religion).
2. **Ordinal scale** – ranks data in order but without equal intervals (e.g., satisfaction levels).
3. **Interval scale** – allows comparison of equal intervals but lacks a true zero (e.g., temperature in °C).
4. **Ratio scale** – possesses all the properties of other scales and includes an absolute zero (e.g., weight, income).

Together, reliability, validity, and measurement levels form the backbone of scientific research, ensuring accuracy, consistency, and meaningful interpretation of data.

6.9 TECHNICAL TERMS

- **Measurement**: Process of assigning numbers or symbols to characteristics according to specific rules.
- **Reliability**: Degree of consistency or stability of a measurement instrument over time.
- **Validity**: The Extent to which a test measures what it claims to measure.
- **Test–Retest Reliability**: Correlation between test scores from the same group on two different occasions.
- **Parallel-Forms Reliability**: Reliability obtained by comparing results from two equivalent forms of a test.
- **Split-Half Reliability**: Reliability is measured by correlating scores from two halves of the same test.
- **Cronbach's Alpha (α)**: Coefficient of internal consistency among test items.
- **Content Validity**: The Degree to which test items represent the entire content area being measured.
- **Criterion-Related Validity**: Correlation between test scores and an external criterion (includes concurrent and predictive).
- **Construct Validity**: Extent to which a test truly measures the theoretical construct it is designed to measure.
- **Face Validity**: Degree to which a test appears, on the surface, to measure the intended concept.
- **Internal Validity**: Accuracy of causal inferences within the study design.
- **External Validity**: Extent to which research results can be generalized to other settings and populations.
- **Nominal Scale**: Scale that categorizes data without any order.
- **Ordinal Scale**: Scale that ranks data but does not specify equal intervals between ranks.
- **Interval Scale**: Scale with equal intervals but no actual zero point.
- **Ratio Scale**: Scale with equal intervals and an absolute zero, allowing ratio comparisons.

6.10 SELF-ASSESSMENT QUESTIONS

A. Short-Answer Questions

1. Define measurement and explain its importance in research.
2. Differentiate between reliability and validity.
3. What are the primary methods of estimating reliability?
4. Explain Cronbach's Alpha and its interpretation levels.
5. Define content validity and give one example.
6. What are the threats to internal validity?
7. How is external validity different from internal validity?
8. What are the four levels of measurement?
9. Distinguish between interval and ratio scales with examples.
10. Explain why reliability is necessary but not sufficient for validity.

B. Long-Answer Questions

1. Discuss in detail the various types of reliability and their estimation methods.

2. Explain the concept and types of validity in psychological measurement.
3. Describe the threats to internal and external validity in experimental research.
4. Elaborate on the four levels of measurement with suitable research examples.
5. How do reliability and validity contribute to the quality of research instruments?

6.11 SUGGESTED READINGS

- Anastasi, Anne. (1988). *Psychological Testing* (6th edition.) London: Mac-Millan.
- Freeman, F. S. (1971). *Theory and Practice of Psychological Testing*. New Delhi: Oxford (India).
- Guilford, J.P. (1954). *Psychometric Methods.* New Delhi: Tata McGraw Hill.
- Cronbach, L.(1951). *Coefficient Alpha and the Internal Structure of Tests*. Psychometrika, 16, 297-334.
- Kaiser, H.F., & Michael, W.B. (1975). *Domain Validity and Gernalisability. Educational and Psychological Measurement,* 35, 31-35.
- McBurney, D.H. & White, T. L. (2007) *Research Methods,* New Delhi; Akash Press.
- Novick, M.R., & Lewis, C. (1967). *Coefficient Alpha and the Reliability of Composite Measurements.* Psychometrika, 32, 1-13.
- Stodola, Q. and Stordahl, K. (1972*). Basic Educational Tests and Measurement*. New Delhi: Thomson (India).

Dr. S. ANITHA DEVI

LESSON- 7

SAMPLING- DEFINITIONS AND BASIC CONCEPTS

OBJECTIVES OF THE LESSON:

After completing this lesson, learners will be able to:

1. Understand the key concepts related to sampling in business research.
2. Explain the process and significance of sampling in research design.
3. Identify different types of probability and non-probability sampling techniques.
4. Evaluate the characteristics of a good sampling design.

STRUCTURE OF THE LESSON

7.1 BASIC DEFINITIONS AND CONCEPTS:

Sampling as an essential part of the business research process entails within itself a number of concepts. A thorough explanation of these concepts at this stage will help in better understanding the chapter in the subsequent stages.

7.1.1Population or Universe

The entire aggregation of items from which samples can be drawn is known as a population. In sampling, the population may refer to the units, from which the sample is drawn. Population or populations of interest are interchangeable terms. The term "unit" is used, as in a business research process; samples are not necessarily people all the time. A population of interest may be the universe of nations or cities. This is one of the first things the analyst needs to define properly while conducting a business research. Therefore, population, contrary to its general notion as a nation's entire population has a much broader meaning in sampling "N" represents the size of the population.

7.1.2 Census

A complete study of all the elements present in the population is known as a census. It is a time consuming and costly process and is therefore, seldom a popular method with researchers. The general notion that a census generates more accurate data than sampling is not always true. Limitations include failure in gathering a complete and accurate list of all the members of the population and refusal of the elements to provide information. The national population census is an example of censes survey.

7.1.3 Sample and Sampling

A sample is a part of the total population. It can be an individual element or a group of elements selected from the population. Although it is a subset, and representative of the population, it is ideal for research in terms of cost, convenience, and time. The sample group can be selected based on a probability or a non-probability approach. A sample usually consists of various units of the population. The size of the sample is represented by "n".

Sampling is the act, process, or technique of selecting a representative part of a population for the purpose of determining the characteristics of the whole population. In other words, the process of selecting a sample from a population using special sampling techniques is called sampling. It should be ensured in the sampling process itself that the sample selected is representative of the population.

7.1.4 Precision

Precision is a measure of how close an estimate is expected to be, to the true value of a parameter. Precision is a measure of similarity. Precision is usually expressed in terms of imprecision and related to the standard error of the estimate. Less precision is reflected by a large standard error.

7.1.5 Bias

Bias is a term that refers to how far the average statistic lies from the parameter it is estimating, that is, the error, which arises when estimating a quantity Errors from chance will cancel each other out in the long run those from bias, will not.

7.2 SAMPLING DESIGN PROCESS

The sampling design process consists of five steps that are intertwined and critical to every aspect of a research project as mentioned:

- Selecting the Population
- Selecting the Sampling Frame
- Specifying the Sampling Unit
- Choosing the Sampling Method
- Deciding the Sampling Size

Selecting the Population

The target population is the group of people that the researcher believes will provide the knowledge needed to complete the research study. The researcher, while developing a sample design, must choose the population according to his/her research study. Population can be finite or infinite. Population is finite if the number of elements in it are certain and countable. In the case of infinite population, no figure can be given about the number of elements in the population.

First and foremost, the researcher selects the target demographic from the entire population. The target population is the population from whom the researcher wishes to deduce the study's conclusions. The accessible population is the segment of the target population that the researcher can contact in order to do research. After determining the available population, a sampling frame consisting of all items or elements of the target population is created to extract a sample from it.

The list of all the uniquely identified elements/units in a population from which a sample will be taken is known as sampling frame. The frame aids in the identification of all items in the population, ensuring that everyone has an equal chance of being chosen for the study.

The sample is the unit(s) in which the researcher conducts his investigation. The term "target population" refers to the group of people or items to which researchers want to apply their results. The target population is the group of people or things from which a sample could be taken. A well-defined group decreases the chances of including items that are unsuitable for the research project's goal.

Selecting the Sampling Frame

In research, sampling frame refers to a list or database of all the items or elements or respondents in the population from which a sample can be chosen. Items or respondents can be selected from sampling frame to be included in the given research project. It is sometimes preferable to choose a list of the population from which the researcher selects units when selecting sample units from the population.

The sampling frame is a collection of people or items (for example, a list of all playgroups in the researcher's city) from which the researcher will select his or her sample. The sample is drawn from a list of all units in a study population. For example, in order to perform his study, the researcher may include all playgroups in his sampling frame located in his city.

Specifying the Sampling Unit

According to Organisation for Economic Cooperation and Development (OECD), a sampling unit is one of the units into which an aggregate is divided for the purpose of sampling where each unit is regarded as individual and indivisible when selection is made. For example, when a survey of a group of trees in a class is conducted, a single tree is a sampling unit. Each item or unit in the sampling frame is called as sampling unit.

Choosing the Sampling Method

Choosing a sampling technique might take some time and entail a number of options, such as whether to use a Bayesian or classical sampling approach, whether to sample with or without replacement, and whether to employ non-probability or probability sampling. Whether a researcher uses probability sampling technique or nonprobability sampling technique usually depends on the purpose of research.

If the sampling frame is almost identical to the target population, random selection can be employed to select the sample. If, on the other hand, the sampling frame does not accurately reflect the target population, the researcher may opt for a non-random selection method that will give him a rough notion of the population in his immediate vicinity.

Deciding the Sample Size

The number of units to be included in the sample is the sample size. Many factors influence the determination of sample size including time, cost, and facility. Larger samples are better in general, but they need more resources.

7.3 CHARACTERISTICS OF A GOOD SAMPLE DESIGN

Some of the important characteristics of a good sample design are:

- Sample design should produce a representative sample
- Sample design should produce a small sampling error
- Sample design should be feasible within the research study's budgetary limits
- Sample design should allow for the control of systematic bias

Sample size must be large enough for the conclusions of the sample study to be generalizable to the entire universe with a fair degree of confidence Provided the researcher wishes to generalize the results.

Apart from the above-mentioned characteristics, a good sample design must also have the following characteristics:

Goal Orientation

A sample design should be orientated to the research aims, adapted to the survey design, and fitted to the survey conditions. If this is done, it should have an impact on the population selection, measurement, and sample selection procedure.

Measurability

A sample design should allow meaningful estimates of sampling variability to be computed. In surveys, this variability is typically reported as standard error. However, this is only achievable with probability sampling. It is impossible to know the degree of precision of survey results in non-probability samples, such as a quota sample.

Practicality

This means that the sample design can be correctly followed in the survey, as planned. Complete, correct, practical, and unambiguous instructions must be provided to the interviewer so that no errors occur in sampling unit selection and the final selection in the field is consistent with the initial sample design. Practicality also relates to the design's simplicity, or its ability to be understood and followed in actual fieldwork operations.

Economy

Finally, economy means that the survey's goals should be met with the least amount of money and effort possible. Generally, survey objectives are stated in terms of precision, which isdefined as the inverse of the variation of survey estimates. The sample design should provide the lowest cost for a given degree of precision. Alternatively, the sample design should yield maximum precision for a given per unit cost (minimum variance).

7.4 APPROACHES TO SAMPLING

There are two basic approaches to sampling probabilistic and non-probabilistic sampling. If the purpose of a research is to arrive at conclusions or make predictions affecting the population as a whole, then the choice of a probabilistic sampling approach is desirable. On the other hand, if a research purpose is directed towards evaluating how a small representative group, is doing for purposes of illustration or explanation, then the use of a non-probabilistic sampling approach is deemed necessary.

Let us look at the various types of sampling under each category:

Probability Sampling

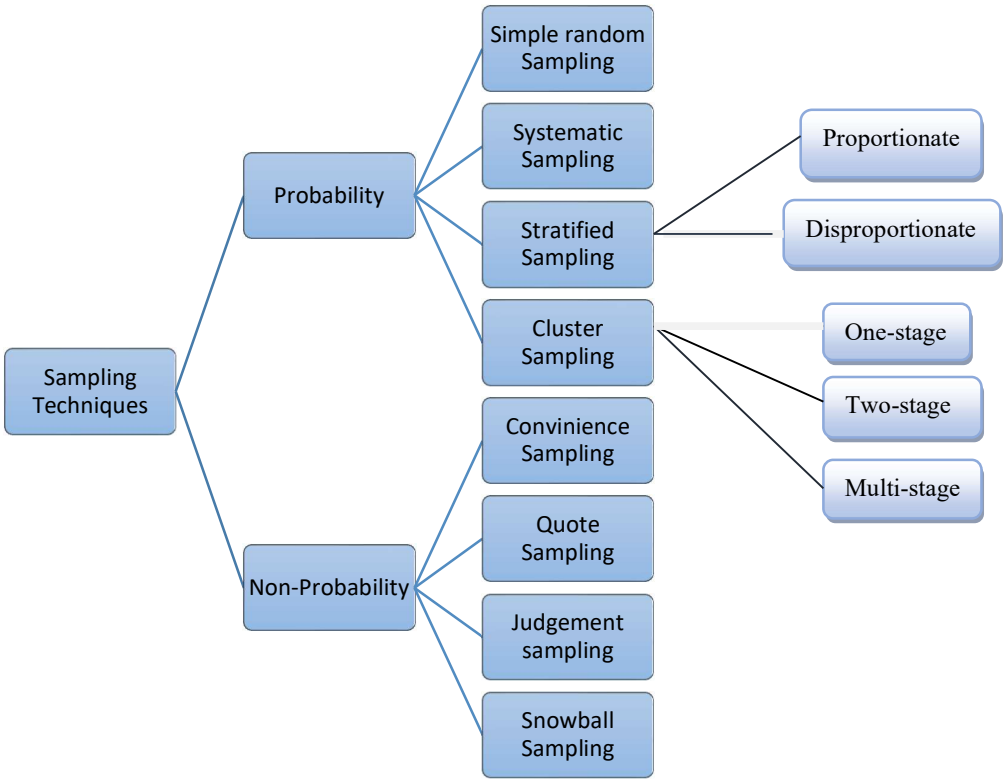
- Simple random sampling

- Systematic sampling
- Stratified random sampling
- Multistage cluster sampling

Non-Probability Sampling

- Convenience sampling
- Quota sampling
- Judgment sampling
- Snowball sampling

Classification of Sampling Techniques



7.4.1 Probability Sampling

A sampling in which every member of the population has a calculable and non-zero probability of being included in the sample is known as probability sampling. Some methods of random selection consistent with the probabilities and the known probabilities of inclusion are used in forming estimates from the sample The probability of selection need not be equal for members of the population.

Simple Random Sampling

A sampling process where each element in the target population has an equal chance or probability of inclusion in the sample is known as simple random sampling. For example, if a sample of 15,000 names is to be drawn from the telephone directory, then there is an equal chance for each number in the directory to be selected. These numbers (serial no of the name) could be randomly generated by the computer on picked out of a box. These numbers could be later matched with the corresponding names thus fulfilling the list. In small populations random sampling is done without replacement to avoid the instance of a unit being sampled more than once.

The benefits of simple random sampling can be reaped when the target population size is small, homogenous, sampling frame is clearly defined, and not much information is available regarding the population. It is advantageous in that it is free of classification error, and requires minimum advance knowledge of the population. Two striking features are the elimination of human bias and non-dependency on the availability of the element. It is seldom put into practice because of the application problem associated with it. This sampling method is generally not preferred as it becomes imperative to list every item in the population prior to the sampling and requires constructing a very large sampling frame, resulting in extensive sampling calculations and excessive costs.

Systematic Sampling

Systematic sampling involves the selection of every k element from a sampling frame. k represents the skip interval and is calculated using the following formulae.

$$\text{Skip Interval (k)} = \frac{\text{Population size}}{\text{Sample size}}$$

Often used as a substitute to simple random sampling, it involves the selection of unit from a list using a skip interval (k) so that every k^{th} element on the list, following a random start between 1 and k, is included in the sample. For example, if k were to equal 6, and the random start were 2, then the sample would consist of 2^{nd}, 8^{th}, 14^{th}, 20^{th} ...elements of the sampling frame.

It is to be noted here that if the skin interval is not a whole number then it is rounded off to the nearest whole number. This sampling method can be used in industrial operations where the equipment's and machinery in the production line are checked for proper functioning as per the specifications. The manufacturer can select every k^{th} item to ensure consistent quality or for detection of defects. Therefore, he requires the first item to be selected at random as the starting point and subsequently he can choose every k^{th} item for evaluation against specifications. It also finds its applicability while questioning people in a sample survey where the interviewer may catch hold of every 10^{th} person entering a particular shop. However, in every case, the researcher has to determine the skin interval and proceed thereafter. In both the cases, it is necessary to

select the first item in the population in a random manner and thereafter follow the skip interval. This method is more economical and less time consuming than simple random sampling.

Stratified Random Sampling

Stratification is the process of grouping members of the population into relatively homogeneous subgroups before sampling. It should be ensured that each element in the population is assigned a particular stratum only. The strata should also be collectively exhaustive to ensure that no population element is excluded. Then random sampling is applied within each stratum independently. This often improves the representativeness of the sample by reducing sampling error**.**

The number of units drawn for sampling from each the strata depends on the homogeneity (standard deviation of the elements present in it. A smaller sample can be drawn from that stratum known to have elements with the same value (proportional stratified sampling), whereas samples can be drawn in a much higher proportion from another stratum where the values are known to differ appreciably disproportional stratified sampling). This is because in the former case, the information from the smaller number of respondents can be enumerated to the whole sample stratum because of high representative sampling. However, in the latter case with much variability among the elements of the stratum, a higher number of respondents would keep the sampling errors to the minimal value Stratified sampling with its potential for greater statistical efficiency scores over simple random sampling. The smaller error may be because the groups are appreciably represented when the strata are combined. Even then, it cannot be used all the time, as there may be errors in designating buses for stratification and because of time and cost factors.

Multistage Cluster Sampling

Cluster sampling involves grouping the elements in a population into various clusters and then selecting a few clusters randomly for further study. The researcher should ensure that clusters are homogeneous (based on some characteristic of the units) in nature and the elements within each cluster are as heterogeneous as possible i.e. each cluster should be similar to the population. Cluster sampling is suitable for conducting research studies that cover large geographic areas with respondents scattered all over.

Once the clusters are formed, the researcher can either go for one-stage, two-stage, or a multi-stage cluster sampling. In single-stage cluster sampling, all the elements from each of the selected clusters are studied, where as in two-stage cluster sampling, the researcher uses random sampling to select a few elements from each of the selected clusters. Multistage sampling involves selecting a sample in two or more successive stages. Here, the cluster/unit selected in the first stage can be further divided into clusters/units.

For example, consider the case where a company decides to interview 400 households about the liability of its new detergent in a metropolitan city, It cannot afford to go for simple random sampling and spend huge resources and time to conduct the research. Therefore, the researcher virtually divides the city into separate blocks (each block represents a cluster), say 40,

ensuring that each of these blocks consists of heterogeneous units (based on some household characteristics) resembling the total population in all its varied features. At this stage he may collect data from 10 elements (households) of each block (one-stage cluster sampling) or move on further to first randomly select a sample of blocks, say 20, and then collect data from 20 elements from each of the selected blocks (two-stage cluster sampling). The researcher may opt for the two-stage cluster sampling if he finds that the individual clusters (blocks) have little heterogeneity within themselves as compared to other clusters, Similarly, a multi-stage cluster sampling involves three or more sampling steps, primarily seen in the case of national and international surveys. It differs from stratified sampling in that the sampling is done on clusters in contrast to elements within strata, as is the case in stratified sampling. Elements are randomly chosen from each stratum in case of stratified sampling whereas only selected clusters are studied in cluster sampling.

7.4.2 Non-Probability Sampling

Non-probability sample or non-random sampling involves the selection of units based on factors other than random chance. It is also known as deliberate sampling and purposive sampling. For example, a scheme whereby units are selected purposefully would yield a non-random sample in a general sense; it is an umbrella term, which includes any sample that does not conform to the requirements of a probability sampling. Convenience sampling, quota sampling, judgment sampling and snowball sampling are few examples of non-probability sampling

Convenience Sampling

The selection of units from the population based on their easy availability and accessibility to the researcher is known as convenience sampling. For example, imagine a company that surveys a sample of its employees to know the acceptance for a new flavour of potato chips that it plans to introduce into the market. This type of sampling is a typical example of convenience sampling as the criterion for selecting a sample is convenience and availability. Although this type of research is cash and cost effective, the findings of the sample survey cannot be generalized to the entire population, as the sample is not representative. As there is no set criterion for selecting the sample, there is a scope for the research being influenced by the bias of the researcher. The researcher may conduct a sample survey involving its own employees to find whether the market as in the above example, would accept the product.

Convenience sampling can be used as a part of a preliminary research that forms a basis for conducting the detailed research Convenience sampling is at its best in surveys dealing with an exploratory purpose for generating ideas and hypotheses.

Quota Sampling

In quota sampling, the entire population is segmented into mutually exclusive groups or categories. The number of respondents (quota) that are to be drawn from each of several categories is specified in advance and the final selection of respondents is left to the interviewer who proceeds until the quota for each category is filled. Quota sampling finds extensive use in

commercial research where the main objective is to ensure that the sample represents in relative proportion, the people in various categories in the population such as gender, age groups, social class, ethnicity, and region of residence. For example, it researcher wants to segment the entire population based on gender, and then he would have two categories of respondents, that is males and females. If he plans to collect a sample of 30, he may allot a quota of 15 for male 15 for female respondents (assuming that the population has an equal proportion of males and females). Therefore, the researcher will stop administering the questionnaire to females after he interviews the 15^{th} female respondent, that is, when the quota of 15 females is filled.

Although quota sampling is similar to stratified sampling, it differs from it, as the units from the individual strata are not drawn randomly. Here, the final selection of the elements is left to the judgment of the interviewer. As the selection of the individuals is left to the judgment of the interviewer, quota sampling is subject to interviewer bias that may result in:

- The quota reflecting the population in terms of superficial characteristics
- The researcher selecting the respondents based on availability rather than on their suitability to the study.

  However, quota sampling does have distinct advantages, such as speed in data collection, lower costs, and convenience. It is cheap and quicker as much time is not required in travelling from one place to another. This also saves money. Further, it is not necessary to keep a track of people to be re-contacted (calling back). Therefore, it is easier to manage.

Judgment Sampling

The selection of a unit, from the population based on the judgment of an experienced researcher or an expert, is known as judgment or purposive sampling. Here, the sample units are selected based on the population's parameters. It is often noticed that companies frequently select certain preferred cities during test marketing their products. This is because they consider the population of that particular city to be representative of the total population of the country. The same is the case with the selection of specific shopping malls that according to the researcher's judgment attract a reasonable number of customers from different sections of the society. Polling results predicted on television is also a result of judgment sampling. Researchers select those districts that have voting patterns close to the overall state or the country in the previous year. The judgment of the researcher is based on the assumption that the past voting trends of selected sample districts are still representative of the political behaviour of the state's population. For example, certain companies test market their new product launches in cities like Mumbai and Indore, because the profile of these cities is representative of the total Indian population.

Snowball Sampling

Sampling procedures that involve the selection of additional respondents based on referrals of initial respondents are known as snowball sampling. This sampling technique is used against low incidence or rare populations. Sampling is a big problem in this case, as the defined population from which the sample can be drawn is not available. Therefore, the process of sampling depends on the chain system of referrals. Suppose, SG sports Ltd., a manufacturer of

sports equipment plans to survey 100 senior squash player through its new website for getting their feedback on the quality of its products.

However, getting a track of such senior squash players can be very difficult, as their presence may be very rare or low. Therefore, it collects the details of the first 200 visitors to its website, to list if any of them is a squash player or knows a squash player. If the visitor is a squash player, then he is requested to refer the names of at least 3 other players known to him. The referred names of the squash players are then called upon for further referrals and this goes on until the sample size of 100 adult players is reached. Although small sample sizes and low costs are the clear advantages of snowball sampling, bias is one of its disadvantages. The referral names obtained from those sampled in the initial stages may be similar to those initially sampled. Therefore, the sample may not represent a cross-section of the total population. It may also happen that visitors to the site or interviewees may refuse to disclose the names of those whom they know.

7.5 SUMMARY

Sampling forms the foundation of business research methodology by allowing researchers to draw inferences about an entire population based on a subset of it. A **population** refers to the total group under study, while a **sample** is a representative part of that group. The **sampling process** includes defining the population, developing a sampling frame, determining sampling units, choosing sampling methods, and deciding the sample size. Researchers can adopt **probability** or **non-probability** techniques depending on their objectives. A good sample design ensures representativeness, reduces bias, and maintains economy and practicality. Proper understanding of **precision** and **bias** enables accurate data collection and valid generalizations for effective business decision-making.

7.6 TECHNICAL TERMS

1. **Population (N):** The entire group of items, individuals, or elements under study.
2. **Sample (n):** A subset of the population selected for analysis.
3. **Census:** A complete enumeration of all elements in the population.
4. **Sampling Frame:** A list of all units or elements from which a sample is drawn.
5. **Precision:** The degree to which repeated measurements under unchanged conditions show the same results.
6. **Bias:** Systematic error that distorts estimates and prevents representativeness.
7. **Probability Sampling:** Sampling method where every element has a known chance of being selected.
8. **Non-Probability Sampling:** Sampling based on researcher's judgment or convenience, not random chance.

7.7 SELF – ASSESSMENT QUESTIONS

1. Define *population* and *sample* and explain their relationship in research.
2. What are the key differences between *census* and *sampling*?
3. List and explain the five main steps in the sampling process.

4. Differentiate between *precision* and *bias* with suitable examples.
5. Identify any four characteristics of a good sample design.
6. Explain the differences between *probability* and *non-probability* sampling techniques.
7. Describe one real-life example where *systematic sampling* is useful.
8. Discuss the advantages and disadvantages of *convenience sampling*.

7.8 SUGGESTED READINGS

1. Kothari, C. R. (2014). *Research Methodology: Methods and Techniques.* New Age International Publishers.
2. Malhotra, N. K. (2020). *Marketing Research: An Applied Orientation.* Pearson Education.
3. Cooper, D. R., & Schindler, P. S. (2017). *Business Research Methods.* McGraw-Hill Education.
4. Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students.* Pearson.
5. Tashakkori, A., & Teddlie, C. (2010). *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Sage Publications.

Dr. K. NAGA SUNDARI

LESSON- 8

SAMPLING PROCESS

OBJECTIVES OF THE LESSON

By the end of this lesson, learners will be able to:

1. Explain the concept and steps involved in the sampling process in business research.
2. Differentiate between various sampling methods and evaluate their suitability for specific research contexts.

STRUCTURE OF THE LESSON

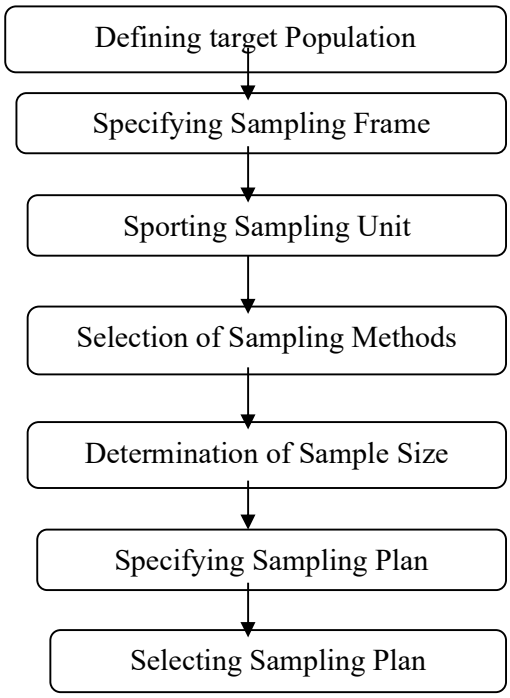
The sampling process refers to the systematic procedure of selecting a subset (sample) from a larger group (population) to obtain information and draw conclusions about the entire population. It ensures that research findings are representative, reliable, and cost-effective.

8.1 STEPS IN A SAMPLING PROCESS

An operational sampling process can be divided into seven steps as given below:

- Defining the target population
- Specifying the sampling frame
- Specifying the sampling unit
- Selection of the sampling method
- Determination of sample size
- Specifying the sampling plan
- Selecting the sample

Steps in Sampling Process



Let us study each of these steps in detail.

An operational sampling process can be divided into seven steps as given below:

- Defining the target population
- Specifying the sampling frame
- Specifying the sampling unit
- Selection of the sampling method
- Determination of sample size
- Specifying the sampling plan
- Selecting the sample

Let us study each of these steps in detail.

Defining the Target Population

Defining the population of interest, for business research is the first step in the sampling process. A clear definition of the target population simplifies the later steps in the sampling process and enhances the quality of the research findings. In general, target population is defined in terms of element, sampling unit, extent, and time frame. The definition should be in line with

the objectives of the research study. For example, if a kitchen Appliances firm wants to conduct demand for its micro ovens, 1 may define the population as 'all women above the age survey to ascertain the age of 20 who cook (assuming that very few males cook). However, this definition is too broad and will include every household in the country, in the population that is to be covered by the survey. Therefore, the definition can be further refined and defined at the sampling unit level, that all women above the age of 20, who cook and whose monthly household income exceeds Rs. 20.000. This reduces the target population size and makes the research more focused. The population definition can he refined further by specifying the area from where the researcher has to draw his sample, that is, households located in Hyderabad.

A well-defined population reduces the probability of including the respondents who do not fit the research objective of the company. For example, if the population is defined as all women above the age of 20, the researcher may end up taking the opinions of a large number of women who cannot afford to buy a micro oven.

Specifying the Sampling Frame

Once the definition of the population is clear a researcher should decide on the sampling frame. A sampling frame is the list of elements from which the sample may be drawn. Continuing with the micro oven example, an ideal sampling frame would be a database that contains all the households that have a monthly income above Rs. 20,000. However, in practice it is difficult to get an exhaustive sampling frame that exactly fits the requirements of a particular research. In general, researchers use easily available sampling frames like telephone directories and list of credit card and mobile phone users. Various private players provide databases developed along various demographic and economic variables. Sometimes, maps and aerial pictures are also used as sampling frames. Whatever may be the case, an ideal sampling frame is one that represents the entire population and lists the names of its elements only once.

A sampling frame error pops up when the sampling frame does not accurately represent the total population or when some elements of the population are missing, Another drawback in the sampling frame is over-representation. If SEBI would like to define its population for the NSE, then it has to do it in terms of listing the names of those individuals who operate in the NSE through different registered brokers. A wrong approach that would lead to over-representation would be to list down all the accounts from the different brokers, as it would also contain the repeated names of those people registered with different brokers. Similarly, a telephone directory can be over-represented by names/households that have two or more connections.

Specifying the Sampling Unit

A sampling unit is a basic unit that contains a single element or a group of elements of the population to be sampled. In this case, a household becomes a sampling unit and all women above the age of 20 years living in that particular house become the sampling elements. If it is possible to identify the exact target audience business research, every individual element would be a sampling unit. This would present a case primary sampling unit. However, a convenient and

better means of sampling would be to select households as the sampling unit and interview all females above 20 years who cook. This would present secondary sampling unit.

Selection Sampling Method

The sampling method outlines the way in which sample units are to be selected. The choice of the sampling method influenced by the objectives of the business research, availability of financial resources, time constraints, and the nature of the problem to be investigated. All sampling methods grouped under two distinct heads, probability non-probability sampling.

Types of Sampling Methods

Sampling methods can be broadly classified into two categories:

1. Probability sampling

2. Non-probability sampling**.**

1. Probability Sampling

In probability sampling, every individual or item in the population has a known, non-zero chance of being selected. This type of sampling is often used when researchers aim for unbiased, generalizable results.

Examples of Probability Sampling:

- Simple random sampling
- Stratified sampling
- Systematic sampling
- Cluster sampling

2. Non-Probability Sampling

In non-probability sampling, individuals are selected based on specific characteristics or convenience rather than random selection. This method is suitable for exploratory research where generalizability is less critical.

Examples of Non-Probability Sampling:

- Convenience sampling
- Quota sampling
- Snowball sampling
- Purposive sampling

Techniques and Examples for Each Sampling Method

Probability Sampling Techniques

1. **Simple Random Sampling**

   Technique: Each individual in the population has an equal chance of being selected. Researchers use random number generators or random selection tools to choose participants.

   Example: A school administrator randomly selects 50 students from a list of all students to survey about cafeteria satisfaction.

2. **Stratified Sampling**

   Technique: The population is divided into subgroups (strata) based on a characteristic (e.g., age, gender), and random samples are taken from each subgroup.

   Example: In a study on employee satisfaction, researchers divide employees into departments (e.g., sales, HR, finance) and randomly select employees from each department.

3. **Systematic Sampling**

   Technique: A starting point is randomly selected, and then every kth individual is chosen from a list. This method is often used when there's a fixed pattern or order in the population list.

   Example: A researcher wants to survey a population of 1,000 people and decides to select every 10th person on a sorted list after a random start.

4. **Cluster Sampling**

   Technique: The population is divided into clusters (groups) that are randomly selected. All individuals within selected clusters are then included in the sample.

   Example: In a national health study, a researcher randomly selects specific cities (clusters) and surveys all residents within those cities.

Non-Probability Sampling Techniques

1. **Convenience Sampling**

   Technique: Participants are selected based on availability or ease of access, making it a fast and easy sampling method.

   Example: A psychology student surveys classmates because they are easily accessible and available for quick data collection.

2. **Quota Sampling**

   Technique: The population is divided into categories (e.g., age, gender), and a specified number of participants from each category is chosen non-randomly.

   Example: A researcher studying consumer preferences might set a quota to survey 50 men and 50 women in a shopping mall.

3. **Snowball Sampling**

   Technique: Participants recruit other participants, making it useful for studying hard-to-reach populations.

   Example: In a study on experiences of ex-convicts, initial participants refer other ex-convicts they know, expanding the sample.

4. **Purposive Sampling**

   Technique: Participants are selected based on specific criteria or characteristics relevant to the study's purpose.

   Example: In a study on the effects of leadership training, a researcher selects participants who hold managerial positions to gain insights specific to leaders.

When to Use Each Sampling Method

1. **Simple Random Sampling**: Use when you need a fully representative sample, especially if the population is homogeneous and a sampling frame is available.
2. **Stratified Sampling**: Best when studying specific subgroups within a population, as it ensures representation across key characteristics.
3. **Systematic Sampling**: Suitable when you have a large population list and need a simple yet systematic approach, especially if the list has no inherent order.
4. **Cluster Sampling**: Useful for large, geographically dispersed populations; ideal when it's impractical to survey individuals directly.
5. **Convenience Sampling**: Ideal for exploratory studies, pilot tests, or when time and resources are limited.
6. **Quota Sampling**: Use when studying demographic or categorical diversity, especially when you need specific representation within the sample.
7. **Snowball Sampling**: Ideal for reaching hidden, hard-to-reach, or marginalized populations.
8. **Purposive Sampling**: Best when studying a specific, well-defined population or a unique group that directly relates to the research question.

Examples of Sampling in Research Studies

1. **Education Study**
   Objective: Investigate student study habits across grade levels.
   Sampling Method: Stratified sampling, where students are divided into grades (strata) and randomly sampled from each grade.
2. **Healthcare Study**
   Objective: Examine patient satisfaction in a hospital network.
   Sampling Method: Cluster sampling, where hospitals (clusters) are selected, and all patients within selected hospitals are surveyed.
3. **Consumer Research**
   Objective: Understand shopping preferences among young adults.
   Sampling Method: Convenience sampling, where young adults at a popular mall are surveyed.
4. **Social Science Study**

Objective: Study the experiences of refugees in a new country.
Sampling Method: Snowball sampling, where initial participants (refugees) refer others in their community.

Method	Advantages	Disadvantages
Simple Random	Representative, unbiased, straightforward.	Can be time-consuming, requires a complete list of the population.
Stratified	Ensures all subgroups are represented, good for diverse populations.	More complex, requires accurate subgroup identification.
Systematic	Simple to implement, evenly spaced selection.	Risk of hidden bias if population has patterns.
Cluster	Cost-effective for large, dispersed populations.	Less precision, higher margin of error due to grouped sampling.
Convenience	Quick, easy, low-cost.	Limited generalizability, high risk of sampling bias.
Quota	Ensures representation across groups, efficient.	Non-random, may introduce selection bias.
Snowball	Effective for hard-to-reach populations, participant-driven expansion.	Risk of network bias, limited generalizability.
Purposive	Targeted, suitable for specific populations with unique characteristics.	Non-random, subject to researcher bias, not generalizable.

Tips for Choosing the Right Sampling Method

1. **Define Your Research Goals**: Clarify whether you need a representative sample or a specific target group to meet the objectives.
2. **Consider Resources**: Time, budget, and accessibility influence the feasibility of sampling methods.
3. **Evaluate Population Characteristics**: Large, diverse populations may require stratified or cluster sampling, while homogeneous populations might benefit from simple random sampling.
4. **Assess Generalizability**: If generalizing results to a larger population is important, prioritize probability sampling methods.
5. **Address Ethical Concerns**: Ensure ethical considerations for sensitive populations, especially when using snowball or purposive sampling.

Determination of Sample Size

The sample size plays crucial role in the sampling process. There various ways classifying the techniques used in determining sample size. A couple those hold primary importance and are worth mentioning are whether technique deals with fixed or sequential sampling and whether logic is based on traditional or Bayesian methods. In non-probability sampling procedures, the allocation of budget, thumb rules and number of subgroups to be analyzed, importance of the decision, number of variables, and nature of analysis, incidence rates and completion rates play a major role in sample determination. In the case of probability sampling, however, formulas are used to calculate sample after the levels of acceptable error and level confidence are specified. The details of the various techniques used to determine the sample size will be explained at the end of the chapter.

Specifying the Sampling Plan

In this step, the specifications and decisions regarding the implementation of the research process are outlined. Suppose, blocks in a city are the sampling units and the households are the sampling elements. This step outlines modus operandi of the sampling plan in identifying houses based on specified characteristics. It includes issues like how is the interviewer going to take a systematic sample of the houses. What should the interviewer do when a house is vacant? What is the re-contact procedure for respondents who were unavailable? All these and many other questions need to be answered for the smooth functioning of the research process. These are guidelines that would help the researcher in every process. As the interviewers and their co-workers will be on field duty most of the time, a proper specification of the sampling plans would make their work easy and they would not have to revert to their seniors when faced with operational problems.

What is a Sampling Plan?

A sampling plan is a structured approach used in statistics and data analysis to select a subset of individuals, items, or observations from a larger population. This process is essential for making inferences about the entire population without the need to examine every single member. The sampling plan outlines the methodology for selecting samples, ensuring that they are representative of the population, which is critical for the validity of statistical conclusions**.**

Types of Sampling Plans

There are several types of sampling plans, each suited for different research objectives and population characteristics. The most common types include simple random sampling, stratified sampling, systematic sampling, and cluster sampling. Simple random sampling involves selecting individuals randomly, ensuring that every member has an equal chance of being chosen. Stratified sampling divides the population into distinct subgroups and samples from each, which helps in achieving a more representative sample. Systematic sampling selects

every nth individual from a list, while cluster sampling involves dividing the population into clusters and randomly selecting entire clusters for analysis.

Importance of a Sampling Plan

The importance of a sampling plan cannot be overstated in the fields of statistics and data science. A well-designed sampling plan enhances the accuracy and reliability of the results obtained from the sample. It minimizes bias and ensures that the sample reflects the diversity of the population. This is particularly crucial in fields such as market research, quality control, and social sciences, where decisions are often based on sample data. By employing a robust sampling plan, researchers can draw valid conclusions and make informed decisions based on their findings.

Components of a Sampling Plan

A comprehensive sampling plan typically includes several key components: the target population, the sampling frame, the sample size, and the sampling method. The target population defines the group from which the sample will be drawn, while the sampling frame is a list or database that includes all members of the target population. The sample size is determined based on the objectives of the study, the desired level of precision, and the available resources. Lastly, the sampling method specifies how the sample will be selected, which can significantly impact the study's outcomes.

Selecting the Sample

This is the final step in the sampling process, where the actual selection of the sample elements is carried out. At this stage it is necessary that the interviewers stick to the rules outlined for the smooth implementation of the business research. This step involves implementing the sampling plan to select a sample required for the survey.

8.2 CRITERIA FOR SELECTING AN APPROPRIATE SAMPLING DESIGN

A business researcher usually decides upon the sampling design. Therefore, each researcher has the leeway to evaluate the sampling design against his preferred criteria. In the following section, we highlight certain common criteria for evaluating and selecting the appropriate sampling design.

Degree of Accuracy

Degree of accuracy is one of the main factors that all business researchers look forward to. Therefore, while drawing upon a sample it is ensured that the sample is representative of the target population. If this is not the case, sampling errors may arise that would lead to errors in the subsequent steps. However, the degree of accuracy sought by researcher varies from one research to another. For instance, an exploratory survey may not demand a highly accurate sampling design, but the same is required for research that is more conclusive and where the

researcher is willing to invest a lot of money and time. The need for accuracy also depends on the decision it is going to support. If the stakes attached with decision area are high, a high degree of accuracy is sought.

Resources

Resources in the form of budget allocation and manpower also influence a researcher's choice of sampling design. For example, if limited resources are allocated for a business research program, the researcher may choose a non-probability sampling design that can be implemented within the time and budgetary constraints.

Time

Time is another important criterion for research. Researchers are likely to opt for simple, less time-consuming sampling designs in the case of a time constraint in the research process. A researcher would prefer telephone survey or convenience sampling to cluster and stratifies sampling.

Prior Knowledge of the Population

Prior knowledge of the population in terms of characteristics, availability, and lists is imperative for a researcher. A population may not be accessible for sampling in the case of researches where population is defined in terms of ownership, experience, or some other qualitative dimension. The lack or unavailability of data may rule out the use of better sampling designs, and a researcher may have to engage in telephonic survey, convenience sampling or snowball sampling to gather relevant data for further progress.

Apart from the above-mentioned factors, the geographical spread of the elements in the population also influences the selection of a sampling design. For example, if the scope of the research covers the whole of India, and the elements in the population are scattered across the country, the researcher has to opt for cluster sampling.

8.3 SAMPLING ERRORS

Sampling errors occur when the sample does not accurately represent the population, leading to incorrect conclusions. These errors can arise from various sources, including selection bias, non-response bias, and measurement errors. Understanding and mitigating sampling errors is crucial for maintaining the integrity of research findings. Researchers can employ techniques such as randomization, stratification, and careful survey design to minimize the risk of sampling errors and enhance the validity of their results.

Applications of Sampling Plans

Sampling plans are widely used across various fields, including market research, public health, social sciences, and quality assurance. In market research, sampling plans help businesses understand consumer preferences and behaviors by analyzing a representative group of

customers. In public health, sampling plans are essential for conducting surveys and studies that inform policy decisions and health interventions. Similarly, in quality assurance, sampling plans are used to assess product quality and ensure compliance with industry standards.

Challenges in Developing a Sampling Plan

Developing an effective sampling plan presents several challenges, including defining the target population, selecting an appropriate sampling method, and ensuring adequate sample size. Researchers must also consider logistical constraints, such as time and budget limitations, which can impact the feasibility of the sampling plan. Additionally, maintaining ethical standards and ensuring participant confidentiality are critical considerations that must be addressed throughout the sampling process.

Best Practices for Sampling Plans

To create an effective sampling plan, researchers should adhere to best practices that enhance the quality and reliability of their findings. These practices include clearly defining the research objectives, selecting an appropriate sampling method, calculating the required sample size, and conducting pilot studies to test the sampling process. Additionally, researchers should document the sampling plan thoroughly to ensure transparency and reproducibility, allowing others to evaluate and replicate the study if necessary.

8.4 SUMMARY

The sampling process is a systematic procedure that guides researchers in selecting representative units from a population for study. It begins with clearly defining the target population—the group that the researcher aims to understand—and progresses through developing a sampling frame**,** determining the sampling unit, and selecting an appropriate sampling method (probability or non-probability). Deciding on the sample size ensures a balance between accuracy and resource efficiency, while the sampling plan provides operational guidelines for data collection. The final step, selecting the sample, involves executing the plan as per the established criteria. Effective sampling design ensures accuracy, minimizes bias, and provides valid results that can be generalized to the broader population.

8.5 TECHNICAL TERMS

1. **Target Population:** The specific group of elements about which the researcher wishes to draw conclusions.
2. **Sampling Frame:** The complete list of all elements or units from which the sample is drawn.
3. **Sampling Unit:** The basic unit containing one or more elements to be sampled.
4. **Sampling Method:** The procedure used to select elements—either probability or non-probability.
5. **Sample Size (n):** The number of elements selected from the population.
6. **Sampling Plan:** A detailed set of operational guidelines for how the sample will be identified, contacted, and surveyed.

7. **Sampling Error:** The difference between results derived from the sample and those that would be obtained from the entire population.
8. **Representativeness:** The degree to which a sample reflects the characteristics of the population.

8.6 SELF – ASSESSMENT QUESTIONS

1. What are the seven steps involved in the sampling process?
2. Why is defining the target population considered the most critical step?
3. Explain the difference between a **sampling frame** and a **sampling unit** with examples.
4. Discuss how a researcher decides on the appropriate **sampling method**.
5. What factors influence the determination of **sample size** in a study?
6. Why is a detailed **sampling plan** important for field researchers?
7. List any four criteria used in selecting an appropriate sampling design.

8.7 SUGGESTED READINGS

1. Kothari, C. R. (2014). *Research Methodology: Methods and Techniques.* New Age International Publishers.
2. Cooper, D. R., & Schindler, P. S. (2017). *Business Research Methods.* McGraw-Hill Education.
3. Malhotra, N. K. (2020). *Marketing Research: An Applied Orientation.* Pearson Education.
4. Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students.* Pearson Education.
5. Sekaran, U., & Bougie, R. (2016). *Research Methods for Business: A Skill Building Approach.* Wiley India.

Dr. K. NAGA SUNDARI

LESSON-9

DEVELOPMENT OF MEASUREMENT SCALES

OBJECTIVES OF THE LESSON

By the end of this lesson, learners should be able to:

1. Understand the concept, significance, and types of measurement scales in business research, including the distinction between qualitative and quantitative parameters.
2. Analyze challenges in measuring complex variables such as motivation and satisfaction
3. Apply appropriate measurement scales in various research contexts

STRUCTURE OF THE LESSON

9.1 INTRODUCTION TO MEASUREMENT IN BUSINESS RESEARCH

Developing measurement scales is a critical dimension of business research. A scale can be defined as a set of numbers or symbols developed in a manner so as to facilitate the assigning of these numbers or symbols to the units under research following certain rules. Generally, it is very easy to measure certain parameters such as sales of a particular product or the profitability of a firm, or the productivity of the employees in an organization, and so on. These are relatively easier because they can be measured quantitatively by applying different scales for measurement. On the other hand, it is relatively difficult to measure some aspects like the motivational levels of

employees in an organization, the attitude of customers towards a particular product, or the customer acceptance levels of a new design of a product, and so on. Measurement of such concepts is very difficult because the respondents may be unable to put their feelings across exactly in words, and sometimes the scales may not be capable of drawing the right response from the respondent.

At times, the respondents might not be willing to reveal their opinions to the researcher. To overcome such difficulties, a researcher's primary objective is to seek the cooperation of the respondent and create an environment of trust and mutual understanding. The interviewer should try to reduce all the negative feelings of the respondent and develop a situation wherein the respondent feels free to share all his feelings relevant to the research with the interviewer. It is also important for the researcher to clearly specify what information he needs and why, if the research is design permits.

9.2 TYPES OF MEASUREMENT SCALES

The design of a measurement scale depends on the objective of the research study, and the mathematical or statistical calculations that a researcher expects to perform on the data collected using the scales. The objective of the research study may be as simple as classifying the population into various categories, or as complex as ranking the units under study and comparing them to predict some trends. Different types of measurement scales are given below.

- Nominal scale
- Ordinal scale
- Interval scale
- Ratio scale

9.2.1 Nominal scale:

A nominal scale uses numbers or letters so as to identify different objects. The scale helps segregate data into categories that are mutually exclusive and collectively exhaustive. This scale assigns numbers to each of these categories and these numbers do not stand for any quantitative value, and hence they cannot be added subtracted or divided. For example, a nominal scale designed to measure the nature of occupation (employment status) may be given as below:

Occupation**:** 1) Public sector; 2) Private sector; 3) Self employed; 4) Unemployed; 5) Others.

In the above example, the numbers 1, 2, 3, 4 and 5 only serve as labels to the various categories of employment status, and hence a researcher cannot use those numbers to perform any type of mathematical or statistical operations on those numbers. A nominal scale does not give any relationship between the variables, and the only quantitative measure is the frequency of items appearing under each category i.e., the number of people in public sector jobs, etc. One can only calculate the mode for the data collected using nominal scale.

9.2.2 Ordinal Scale:

An ordinal scale is used to arrange objects according to some particular order. Thus, the variables in the ordinal scale can be ranked. For instance, if someone says that person came second in the exam, then we can understand that there was another person who came first and some others were there who were ranked after him. This type of scale that gives ranks is called an ordinal measurement scale. Ordinal variables can only give us the information regarding relative position of the participants in the observation, but they do not give any information regarding the absolute magnitude of the difference between the first and the second position, or second and third position and so on.

For example, an ordinal scale used to measure the preference of customers (in Andhra Pradesh) for various mobile telephone service providers would ask a question like

Please rank the following mobile telephone service providers from 1 to 5 with 1 representing the most preferred and the least preferred.

Airtel -------

Hutch -------

Idea -------

BSNL -------

Reliance -------

A respondent may rank these players depending on his experience/perception of them. If a respondent ranks Airtel as 1 and Idea as 2, a researcher can know that the respondent prefers Airtel. However, the limitation is that the researcher cannot be sure as to how strong the respondents' liking is for Airtel when compared to Idea.

9.2.3 Interval Scale:

Interval scales are similar to ordinal scales to the extent that they also arrange objects in a particular order. However, in an interval scale the intervals between the points on the scale are equal. This is the scale where there is equal distance between the two points on the scale. Examples of interval scales are Fahrenheit and Celsius scales used to measure temperature. In these scale the difference between the intervals is the same i.e., the difference between 40º and 60º is the same as the difference between 25º and 45°. But the base point, freezing of water is represented by 32°F and 0ºC. Thus there is no natural zero (base) for these scales.

Similarly we can design an interval scale with points placed at an interval of 1 point [10] --- [9] --- [8] --- [7] --- [6] --- [5] --- [4] --- [3] --- [2] --- [1] and ask the respondents to place the mobile telephone service providers on this scale of 10 to 1. If Idea is assigned 8 and BSNL 4 we can say that the value of difference in preference is 4. But we cannot say that the liking for Idea

is twice that for BSNL because we did not define a point of no liking i.e., 0. The only statement we can make about a respondent's preference for Idea and BSNL is 'he likes Idea more than BSNL' but we can't give a ratio of the likings as there is no base zero.

Interval scales are suitable for the calculation of an arithmetic mean, standard deviation, and correlation coefficient.

9.2.4 Ratio Scale:

Ratio scales have a fixed zero point and also have equal intervals. Unlike the ordinal scale the ratio scale allows for the comparison of two variables measured on the scale. This is possible because the numbers or units on the scale are equal at all levels of the scale. A very good example of ratio scale is distance; for instance, not only can we say that the difference between four miles and six miles is the same as the difference between six miles and eight miles but we can also say that eight miles is twice as long as four miles. In other words, a ratio scale can be defined as a scale that measures in terms of equal intervals and an absolute zero point of origin exists. This zero is common to a distance scale using yards, meters, etc. Age, height, weight, money scales are other common examples of ratio scales. Since there exist an absolute zero on the ratio scale the data collected can be subjected to any type of mathematical operation say addition, subtraction, multiplication, and division.

9.3 CRITERIA FOR GOOD MEASUREMENT

Researchers normally develop their own scales for measuring variables for different attributes as it is very difficult to find readily available scales. It is in this process of developing scales that researchers have to be very careful, since the scales that they develop should primarily stand the tests of reliability, validity, sensitivity and so on. In the following sections, we will discuss the criteria for good measurement. There are five major criteria for analyzing the goodness of a measurement, namely, reliability, validity, sensitivity, generalizability and relevance.

Reliability

It is considered that, when the outcome of a measuring process is reproducible, then the measuring instrument is reliable. Reliable measuring scales provide stable measures at different times under different conditions. For example, if a coffee vending machine gives the same quantity of coffee every time, then it can be concluded that the measurement of the coffee vending machine is reliable. Thus reliability can be defined as the degree to which the measurements of a particular instrument are free from errors and as a result produce consistent results. However in certain situations, poor data collection methods give rise to low reliability. The quality of the data collected can become poor if the respondents do not understand the questions properly and give irrelevant answers to them. There are three methods that can be used to evaluate the reliability of a measure. They are test-retest reliability, equivalent forms and internal consistency.

Test-retest reliability:

If the result of a research is the same, even when it is conducted for the second or third time, it confirms the repeatability aspect. For example, if 40 percent of a sample say that they do not watch movies, and when the research is repeated after sometime and the result is same (or almost the same) again, then the measurement process is said to be reliable. However, there are certain problems regarding the test-retest method of testing reliability, the first and foremost issue is that it is very difficult to obtain the cooperation and locate all the respondents for a second round of research. Apart from this, the responses of these people may have changed on the second occasion, and sometimes environmental factors may also influence the responses.

Equivalent form reliability:

Some of the shortcomings of test-retest reliability can be overcome in this method. In equivalent form reliability, two measurement scales of a similar nature are to be developed. For instance, if the researcher is interested in finding out the perceptions of consumers on recent technologically advanced products, then he can develop two questionnaires. Each questionnaire contains different questions to measure their perceptions, but both the questionnaire should have an approximately equal number of questions. The two questionnaires can be administered with a time gap of about two weeks. The reliability in this method is tested by measuring the correlation of the scores generated by the two instruments. The major problem with equivalent form reliability is that it is almost impossible to frame two totally equivalent questionnaires.

Internal consistency:

Internal consistency of data can be established when the data give the same results even after some manipulation. For example, after a research result is obtained for a particular study, the result can be split into two parts and the result of one part can be tested against the result of the other, if they are consistent, then the measure is said to be reliable. The problem with internal consistency is that the reliability of this method is completely dependent on the way the data is divided up or manipulated. Sometimes it so happens that different splits give different results. To overcome such problems with split halves, many researchers adopt a technique called as Cronbach Alpha which needs the scale items to be at equal intervals. In case of difficulty in obtaining the data at equal intervals of time then an alternate method called KR-20 (Kuder Richardson Formula 20) is used to calculate how consistent subject responses are among the questions on an instrument. Items on the instrument must be dichotomously scored (0 for incorrect and 1 for correct). All items are compared with each other, rather than half of the items with the other half of the items. It can be shown mathematically that the Kuder-Richardson reliability coefficient is actually the mean of all split-half coefficients.

Validity

The ability of a scale or a measuring instrument to measure what it is intended to measure can be termed as the validity of the measurement. For instance, students may complain about the validity of an exam, stating that it did not measure their understanding of the topic, but only their memorizing ability. Another example may be of a researcher who tries to measure the morale of

employees based on their absenteeism alone; in this case too, the validity of the research may be questioned, as absenteeism cannot be purely attributed to low morale, but also to other conditions like prolonged illness, family reasons and so on.

Validity can be measured through several methods like face validity, content validity, criterion-related validity and construct validity.

Face Validity:

Face validity refers to the collective agreement of the experts and researchers on the validity of the measurement scale. However, this form of validity is considered the weakest form of validity. Here, experts determine whether the scale is measuring what it is expected to measure or not.

Content Validity:

Content validity refers to the adequacy in the selection of relevant variables for measurement. The scale that is selected should have the required number of variables for measurement. For instance, if the state education department wants to measure whether all the schools in the city have adequate facilities, and for measuring this, it develops a scale to measure the attributes like the attractiveness of schools names, the frequency of old students meet, the different varieties of eatables that are prepared in the school canteen and so on. Here, it is clear that these variables considered for measurement do not possess any content validity as they will not serve the purpose of the research. The scale should instead be developed to measure aspects such as the number of classrooms, the number of qualified teachers on roll, the capacity of the playground and so on. It is often difficult to identify and include all the relevant Variables that need to be studied for any research process.

Criterion-related Validity:

The criterion related validity refers to the degree to which a measurement instrument can analyze a variable that is said to have a criterion. If a new measure is developed, one has to ensure that it correlates with other measures of the same construct. For instance, length of an object can be measured with the help of tape measure, callipers, odometer and also with a ruler and if a new technique of measure is developed then one has to ensure that this new measure correlates with other measures of length. If a researcher wants to establish criterion validity for a new measure for payment of wages, then he may want to ensure that this measure correlates with other traditional measures of wage payment such as total number of days worked.

Criterion validity may be categorized as predictive validity and concurrent validity. Predictive validity is the extent to which a future level of a criterion variable can be predicted by a current measurement on a scale. A scale for measuring the future occupancy of an apartment complex for example may use this scale. A builder may give preference to only those repairs that may attract new tenants in the future rather than focusing on all the areas that need repair. Concurrent validity is related with the relationship between the predictor variable and the

criterion variable. Both the predictor variable and the criterion variable are evaluated at the same point in time.

Construct validity:

Construct validity refers to the degree to which a measurement instrument represents and logically connects through the underlying theory. Construct validity, although it is not directly addressed by the researcher, is extremely important. It assesses the underlying aspects relating to behaviour; it measures why a person behaved in a certain way rather than how he has behaved. For instance, whether a particular product was purchased by a consumer, is not the consideration, but why he has/has not purchased the product is taken into account to judge construct validity. This helps to remove any extraneous factors that may lead to incorrect research conclusions. For example, for a particular product, price may not be the factor that affects a person deciding whether to buy it. If this product is used in the measurement of a general relationship of price and quantity demanded, it does not have construct validity, as it does not connect with the underlying theory.

There are two statistical methods for analyzing construct validity - convergent validity and discriminant validity. Convergent validity is the extent of correlation among different measures that are intended to measure the same concept. Discriminate validity denotes the lack of or low correlation among the constructs that are supposed to be different. Consider a multi-item scale that is being developed to measure the tendency to stay in low cost hotels. This tendency has four personality variables; high level of self-confidence, low need for status, low need for distinctiveness, and high level of adaptability. Additionally, this tendency to stay in low cost hotels is not related to brand loyalty or high level aggressiveness. The scale can be said to have construct, if it correlates highly with other measures of tendency to stay in low cost hotels such as reported hotels patronized and social class (convergent validity). Has a low correlation with the unrelated constructs of brand loyalty and a high level of aggressiveness (discriminant validity).

Sensitivity

Sensitivity refers to an instrument's ability to accurately measure variability in stimuli or responses. Sensitivity is not high in instrument's involving 'Agree' or 'disagree types of response. When there is a need to be more sensitive to subtle changes, the instrument is altered appropriately. For example, strongly agree, mildly agree, mildly disagree, strongly disagree, none of the above are categories whose inclusion increases the scale's sensitivity.

Generalizability

Generalizability refers to the amount of flexibility in interpreting the data in different research designs. The generalizability of a multiple item scale can be analyzed by its ability to collect data from a wide variety of respondents and with a reasonable flexibility to interpret such data.

Relevance

Relevance, as the name itself suggests, refers to the appropriateness of using a particular scale for measuring a variable. It can be represented as,

Relevance = reliability x validity

If correlation coefficient is used to analyze both reliability and validity, then the scale can have relevance from 0 to 1, where 0 is the low or no relevance level to 1 which is the high relevance level. Here if either of reliability or validity is low then the scale will have little relevance.

9.4 TYPES OF ATTITUDE SCALES:

There are two major types of scales used to measure the attitudes of respondents. They are single item scales and multi- item scales. The different types are shown below:

Exhibit 1

Itemized Category Scale

Given below is an itemized category scale ranging from highly satisfied to highly unsatisfied. Please select one of the following options based on your satisfaction levels of hotel service.

☐	☐	☐	☐	☐
Highly Satisfied	Considerably Satisfied	Reasonably Satisfied	Unsatisfied	Highly Unsatisfied

Single Item Scales:

Single item scales are those with only one item to measure. Itemized category scales are most commonly used under single item scales. Besides, itemized category scales, there are several other scales such as comparative scales, rank order scales and so on, which are used for attitude measurement. We will discuss each of these in the following sections.

- **Itemized category scales**
  Itemized category scales are those in which respondents have to select an answer from a limited number of ordered categories. Respondents are given the scale that contains a number or brief description about a particular category. These categories are ordered in terms of position of the scale and respondents have to select one category that they feel best describes the object. It is easy to develop itemized category scales. Exhibit 1 gives an

itemized category scale where a hotel customer is asked to indicate the level of satisfaction for service provided.

- **Rank - order scales**
  Rank order scales are comparative scales, where the respondent is asked to rate an item in comparison with another item or a group of items against each other based on a common criterion. For instance, a respondent may be asked to rank three motorcycle brands on attributes such as cost, mileage, style, pick-up and so on. Although it is easy to develop a rank-order scale, it has some disadvantages. It is very difficult to include every possible brand or attribute on a scale. Therefore, a respondent may rate a brand as number one, but it might not be his first choice as the brand be prefers may not have been included in the list at all. Sometimes, respondents may feel that the attributes used to construct the scale are not relevant to judging the subject under research. One major shortcoming is that the researcher will not have any clue about why the respondent has given a particular rating for items listed on the scale. Exhibit 2 shows a rank order scale for ranking different brands of motorcycles on specified attributes.

Exhibit 2

Rank Order Scale used for Analyzing Motorcycles

Please rank the following brands of motor cycles with 1 being the brand that best meets the characteristic being evaluated and 7 being the worst on the characteristic being evaluated. Let us now start rating these brands basing on their affordability, first. Which brand has the highest affordability? Which is second? (Record the answers below).

Brand of Motorcycle	Affordable Cost	High Mileage	Stylish	Great Pickup
Hero Honda				
TVS				

Exhibit 3

Comparative Scales

Given below is the scale ranging from excellent to very poor. If you were asked to rate the sweet shop ‘X’ in comparison to sweet shop ‘Y’ in Hyderabad. Which one will you choose? If you choose excellent then select the first option:

☐ ☐ ☐ ☐ ☐ ☐

Excellent Very Good Good Both are same Poor Very poor

- **Q-sort scales:**
  When the number of objects or characteristics to be rated is very large in number, it becomes difficult and tedious for respondents to rank order. In such cases, Q-sort scaling is used. Here, respondents are asked to sort out various characteristics or objects that are being compared into various groups so that the distribution of the number of objects or characteristics in each group follows a normal pattern. For instance, let us consider that the designing team of a toy manufacturing company has come out with hundreds of new product ideas with slight variations. The research team's task is to find out from customers which combination of features is the best and will generate maximum sales. To accomplish this, Q-sort scaling is the best method. The procedure followed is:
  Respondents are given a set of cards, usually varying from 80 to 120 cards containing different categories of items to be selected from for instance, if respondents have to rate 100 different products according to their tastes and preferences, each respondent will be given about 100 cards containing a product and its features. Respondents are then asked to segregate the cards into 10 stacks so that the 1st stack contains a set of cards that are highly preferred by respondents. The 10th stack will contain a set of cards that are least preferred by them. The individual stacks in between (2nd to 9th) should be prepared by the respondent in such a way that they range from higher preference to lower preference. Once the stacks are ready, the cards in each stack should be arranged in the respondents' order of preference, based on criteria like features of a product, communication processes and customer service. This gives the best and the worst product in each stack. The disadvantage of this process is that it asks a lot of time and effort on the part of respondents.
- **Comparative scales:**
  In the itemized category scale, respondents select a category that they feel best describes a product. The problem here is that respondents may select a category based on their own perceptions. For instance, respondent A might select a category based on his or her view of an ideal brand, respondent B may pick a brand based on knowledge of an existing brand and respondent C might choose based on some other criteria. Ultimately, the selection process lacks uniformity. To overcome this, comparative scales have been developed, where the researcher provides a point of comparison for respondents to provide answers. Therefore, all respondents will have a uniform point of comparison for selecting answers. For instance, rather than asking a person to evaluate the quality of sweets in one sweet shop in Hyderabad, the respondents will be asked to evaluate the quality of that sweet shop in comparison to another sweet shop in Hyderabad. Exhibit 3 gives a comparative rating scale.

Exhibit 4

Paired Comparison scale for a Toothpaste

Please select one item each from the following pairs that is most important to you for selecting a toothpaste.

a. Fights decay	b. Affordable
a. Affordable	b. Longer germ protection
a. Longer germ protection	b. Fights decay

In paired comparison scales, respondents are asked to select one of two items in a pair based on pre-set criteria. As each item is compared with all other items, the number of times an item is selected from a pair gives its rank. The higher the number, the better is the rank. In this method, the shortcoming of rank order scaling is overcome, as it is easy for respondents to select one item from two rather than ranking a long list of items. Another advantage is that the problem of order bias is eliminated as no set pattern is followed while providing respondents the pairs. A typical paired comparison scale for toothpaste is shown in Exhibit 4.

- **Constant sum scales**
  In constant sum scales, respondents are asked to divide a given number of points, usually 100, among two or more attributes based on the importance they attach to each attribute. These scales are often used in place of paired comparison scales to eliminate the long lists in paired comparisons. Here, respondents have to rate an item in relation with all other items. Ranking for each item is based on the points assigned by the respondent to the items. The disadvantage of this approach is that the researcher is limited to giving 10 items for the respondent as a higher number of items will confuse the respondent. Exhibit 5 shows the constant sum scale where respondents are asked to rate 10 characteristics of a supermarket for a total sum of 100 points.

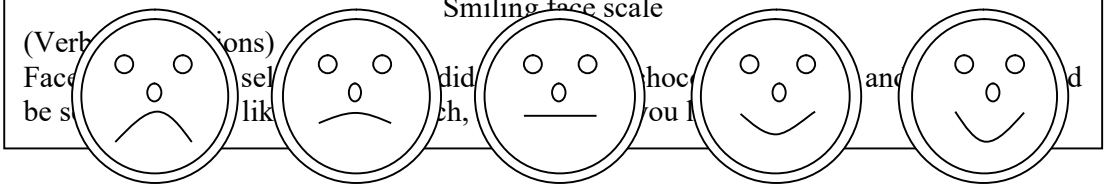
Exhibit 5

Constant Sum Scale used for a Supermarket

Given below are the ten characteristics of a supermarket. Please give each characteristic some point(s) based on your assessment, so that the total points add up to 100. The higher number of points allocated to a particular characteristic, the higher its importance to you, and then you need not assign any points to it. However, it is essential that all points given add up to 100.

Characteristics of a supermarket	**Number of points**
The supermarket is conveniently located	_____
The supermarket has enough range of products	_____
All the items in the store are conveniently located	_____
Sales persons are cooperative	_____
Aisle space is comfortable	_____
Prices are very much affordable	_____
The ambiance in the store is pleasing	_____
Soft music played in the store is entertaining	_____
Billing counters are sufficient	_____
Parking facility is adequate	_____
	100 Points

Exhibit 6
Smiling face scale



- **Pictorial scales:**
  Here, the different types of scales are represented pictorially. The respondents are asked to rate a concept or statement based on their intensity of agreement or disagreement, on a pictorial scale. Pictorial scales have to be developed carefully so that respondents will not have problems selecting appropriate responses. These scales are generally used for respondents who cannot analyze complex scales, such as young children or illiterates. Typical pictorial scales are a thermometer scale or a scale depicting a smiling face. Exhibit 6 shows the smiling face scale for measuring the effectiveness of an advertisement campaign for a chocolate.
- **Continuous scales:**
  Continuous scales are those where respondents are asked to rate items being studied by marking at an appropriate place on a line drawn from one extreme of the scale to the other. These scales are rarely used in marketing research as they do not give accurate results and the scoring process is complicated. This scale's only advantage is that it is very easy to develop. For instance, if a fast food outlet such as Pizza Hut wants to find out whether customers are satisfied with its overall service, then a continuous scale can be developed as shown in Exhibit 7.

Exhibit 7

Continuous Rating Scale

Given below is a continuous scale ranging from 0 to 100 points. You have to indicate a point that best describes how you rate the overall service. If you rate it the best then it would be 100.

How would you rate the overall service of pizza hut?

Best I Worst

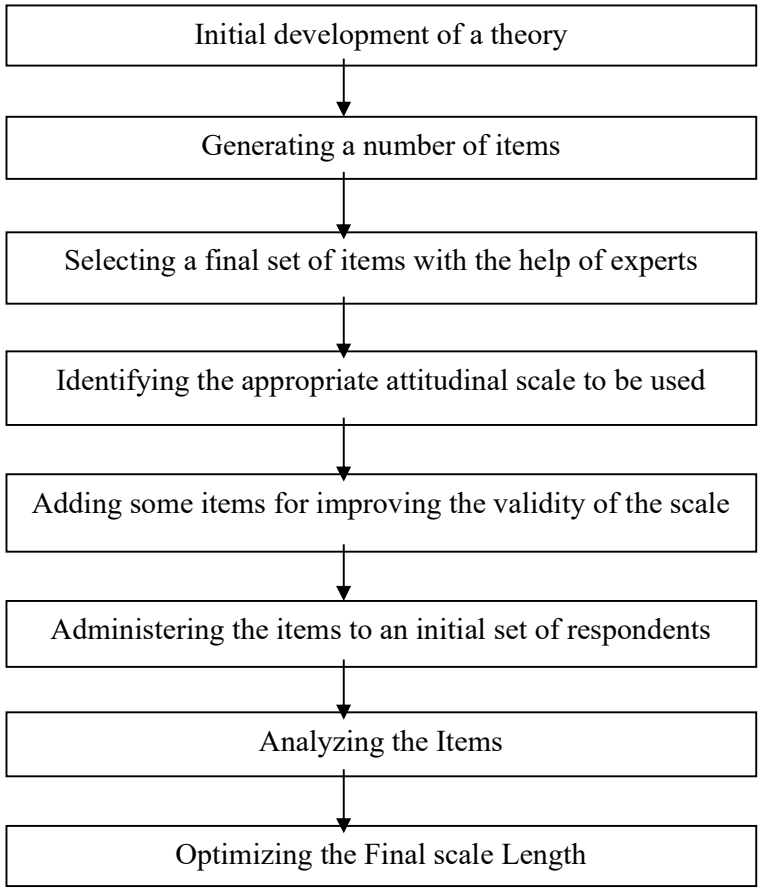
100 90 80 70 60 50 40 30 20 10 0

Multi- Item Scales:

Let us move now to multi-item scales. These scales are used when it is difficult to measure people's attitudes based only on one attribute. For instance, to measure respondents' attitudes towards the Indian Railways, if you ask them only whether they are satisfied with Indian Railways or not, it will not suffice. People may say that they are satisfied on an overall basis, but there might be number of factors that they find unsatisfactory. Thus, it is impossible to capture the complete picture with one overall attitude-scale question. To measure individual attributes, a number of scales have been developed that can measure a respondent's attitude on several issues on a scale ranging from most favourable to least favourable. The Semantic, Likert, Thurstone and differential scales are some examples that follow such measurement techniques. Developing multi-item scale involves certain crucial steps that have been discussed in exhibit 8.

Exhibit 8

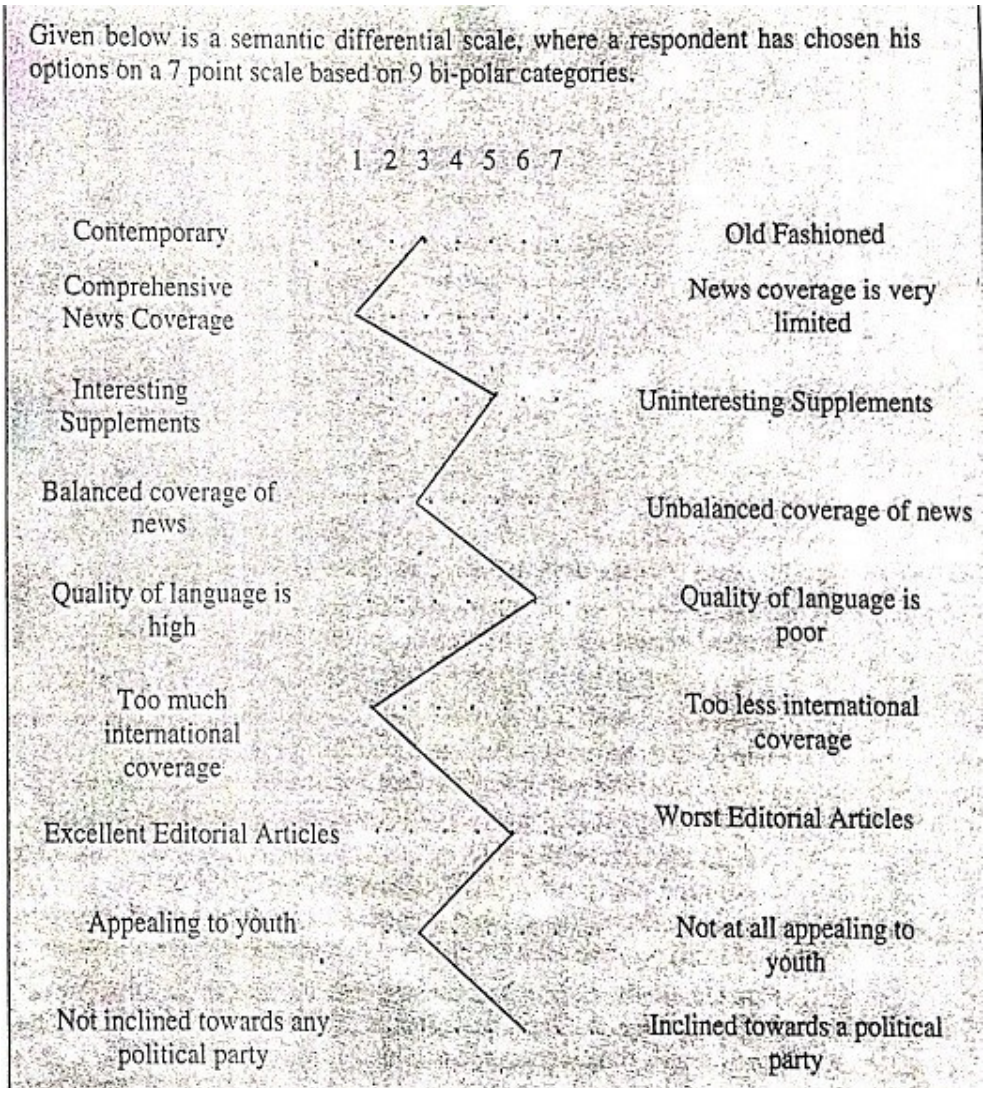
Developing a Multi-item Scale



- **Semantic differential scales:**
  Semantic differential scales are used to describe a set of beliefs that underline a person's attitude towards an organization, product or brand. This scale is based on the principle that individuals think dichotomously or in terms of polar opposites such as reliable-unreliable, modern-old fashioned, cold-warm.
  The respondents are asked to rate an attitude object on a set of itemized, seven- point rating scale, bounded by bipolar phrases or adjectives. The initial process of developing a semantic differential scale starts with determining the object to be rated.

Exhibit 9

A Semantic Differential Scale for Measuring the Attitudes of Respondents for a Newspaper



Once this object is determined, respondents are asked to express their attitudes towards the object, using the dichotomous pair on a scale. Such points are then plotted on a graph. This is the most efficient technique for determining the strengths and shortcomings of a product/service or a company in the market.

While designing the scale, care should be taken that all negative or positive adjectives or phrases do not appear on one side. This avoids a person from picking either only positive or negative phrases.

Another problem should be addressed while developing a seven- point semantic scale is the response of 4. If the respondent selects 4 for all items, then it becomes neutral without indicating any specific direction. Exhibit 9 represents the semantic differential scale administered for measuring the attitudes of respondents towards a newspaper. It can also be used for comparing the products with that of the competition. Consider four brands of cars being rated on the same scales as shown in the Exhibit 10.

Exhibit 10

Semantic Differential Scale for Comparing Four Brands of Cars

Below given is the semantic differential scale rated by a responding by comparing 4 brands of cars.

Mitsubishi (L) Lancer
Skoda Octavia (O)
Hyundai (E) Elantra
Honda (C) City

Fast	EL		O		C			Slow
Large		L	E	O	C			Small
Plain			E	L	O	C		Stylish
Inexpensive			C	L	E	O		Expensive

- **Stapel scales:**
A Stapel scale is an attitude measure that places a single adjective or an attribute describing an object in the centre of an even number of numerical values. In general, staple scales are constructed on a scale of 10 ranging from -5 to +5, without a neutral point (zero). The respondent is asked to rate attributes on this scale.
Stapel scales are similar to semantic differential scales but here there is only one pole (single adjective) rather than bipolar adjectives. This scale is useful for researchers to

understand the positive and negative intensity of attitudes of respondents. The numeric value assigned to an adjective show how well it describes the object. The higher the positive value, the better it describes the object. One big disadvantage is that the respondent might select all attributes on a positive or negative range. A Stapel scale that is designed to measure the attitude of passengers towards an airline is shown in exhibit 11.

- **Likert scales:**
  Likert scales consist of a series of statements where the respondent provides answers in the form of degree of agreement or disagreement. This expresses attitude towards the concept under study. The respondent selects a numerical score for each statement to indicate the degree of agreement or otherwise. Each such score is finally added up to measure the respondent's attitude. The various steps involved in developing a Likert scale are given below.

  - Identify the concept that needs to be measured
  - Develop a series of statements (say, 100) that articulate respondents' feelings towards the concept
  - Every test item is categorized by the respondent as generally favourable or unfavourable based on the attitude that needs to be measured
  - A pre-test is conducted to measure the intensity of the favourable or unfavourable attitude of respondents towards each test item. The scale would have intensity descriptors like, highly favourable, favourable, neutral, unfavourable, and highly unfavourable. These responses are given a numerical weight.
  - The total attitude score is represented by the algebraic sum of the weights of the items. To make the measuring process uniform, the weights are consistently assigned. For instance, if 5 were assigned to reflect strong agreement with a favourable situation, then 5 should be assigned to show strong disagreement with an unfavourable situation too.
  - After the results have been obtained, the researcher selects items that reveal a clear discrimination between high and low total scorers by identifying the highest and lowest quartiles based on total scores. Subsequently, mean differences are computed for these high and low groups.
  - Finally, a set of items is chosen that represent the greatest difference between the highest and the lowest mean values.

Exhibit 11

Stapel Scale for Measuring the Attitudes of Flight Passengers

Below given is a stapel scale designed to measure your attitude on three attributes. Please circle one number from the following three columns that best describes your attitude towards them.

+5	+5	+5
+4	+4	+4
+3	+3	+3
+2	+2	+2
+1	+1	+1
Friendly Cabin Crew	Comfortable Interiors	Accurate timings
-1	-1	-1
-2	-2	-2
-3	-3	-3
-4	-4	-4
-5	-5	-5

Likert scales are very popular among researchers for measuring the attitudes of people. But, in practical situations, commercial researchers are more concerned with finding the respondents attitudes towards individual components, rather than overall positive or negative attitudes of respondents. For instance, the manufacturer of a brand of shoes will be more interested in finding out why people are not buying the brand rather than respondents' attitudes towards shoes in general. A typical Likert scale is discussed in the exhibit 12.

Exhibit 12

Likert Scale

A Likert scale for evaluating the attitudes of customers, who have not used a vacuum cleaner, but are aware of its existence is given below.

Here are some statements that describe how customers might feel about vacuum cleaners.

Please indicate your agreement or disagreement. For each statement given below, please circle the appropriate number to indicate whether you: 1- strongly agree, 2- agree, 3- neutral, 4- disagree and 5- strongly disagree.

	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The product is costlier	1	2	3	4	5
I don't find time to use a vacuum cleaner	1	2	3	4	5
Advertising of the product is not convincing enough	1	2	3	4	5
I have never used a vacuum cleaner	1	2	3	4	5
I am satisfied with the way I am cleaning my house right now	1	2	3	4	5
Using a vacuum cleaner has better features	1	2	3	4	5
Competitor's vacuum cleaner has better features	1	2	3	4	5

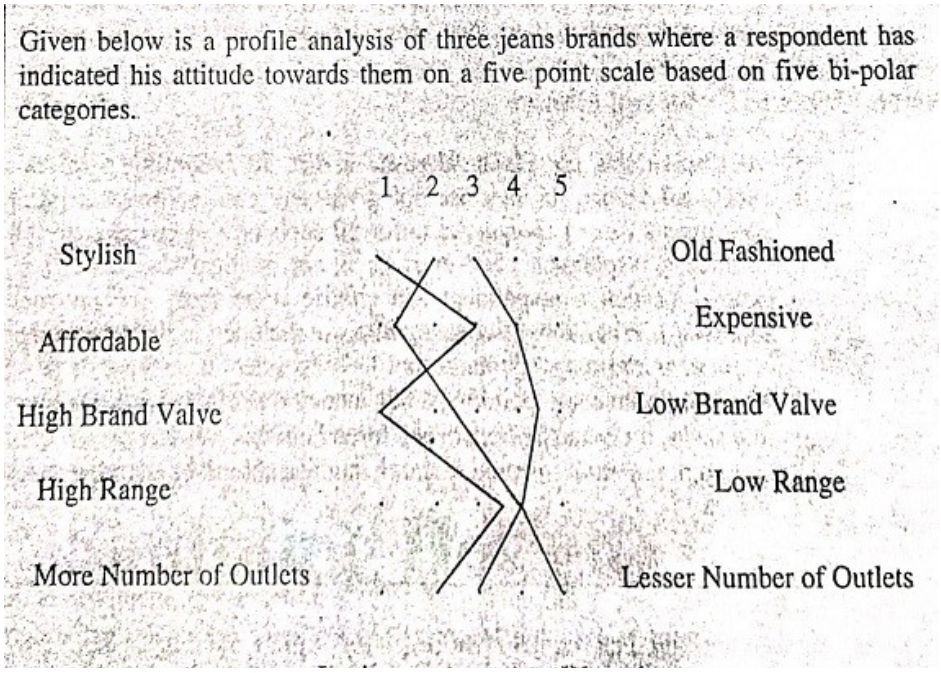
- **Thurstone scales:**
  In Thurstone scales, researchers select a group of 80 to 100 items indicating the different degrees of favourable attitude towards a concept under study. Once items are selected, they are given to a group of judges, who are asked to categorize them according to how much they favour or disfavour them. The judges are asked to treat intervals between categories as equal and analyse each item without expressing their own attitudes. Once the results are obtained, all those items that have a consensus from the judges are selected and items where there was no consensus are eliminated. These results are then distributed uniformly on a scale of favourability. This scale is then administered to a set of respondents for measuring their attitude towards a particular concept. Although the Thurstone method is time

consuming as it involves a two- stage procedure, it is easy to administer. This method comes under criticism because the scale values are developed based on the attitudes of the judges.

- **Profile analysis:**
  Profile analysis is a process where two or more objects are rated by respondents on a scale. Profile analysis can be considered as an application of the semantic differential scale. Comparing different objects visually, based on different attributes, is possible in this approach. The major disadvantage is that it is very difficult to interpret the profiles as the number of objects increases. The profile analysis is used in exhibit 13 to compare the three jeans brands.

Exhibit 13

A Profile Analysis for Measuring the Attitudes of Respondents towards Jeans Brands



9.5 SUMMARY

Developing measurement scales is a crucial aspect of business research, as it helps quantify both tangible and intangible variables. A measurement scale is a structured system of numbers or symbols used to assign values to objects, individuals, or events according to defined rules.

Quantitative parameters such as sales**,** profit**,** and productivity are relatively easy to measure**.** However, psychological or behavioral attributes like motivation**,** attitude**,** or customer satisfaction pose measurement challenges due to their subjective nature. Effective measurement requires creating a trustful environment with respondents and ensuring clarity about the purpose of data collection.

The four main types of measurement scales are:

1. Nominal Scale – Used for classification (e.g., gender, region, brand).
2. Ordinal Scale – Used for ranking or ordering (e.g., satisfaction levels: high, medium, low).
3. Interval Scale – Measures differences between values but lacks a true zero (e.g., temperature in Celsius).
4. Ratio Scale – Has an absolute zero and allows all mathematical operations (e.g., income, weight, age).

The choice of scale depends on the research objectives and the type of analysis intended.

9.6 TECHNICAL TERMS

1. **Measurement Scale:** A system of symbols or numbers used to quantify or categorize data.
2. **Nominal Scale:** A scale that classifies data into categories without any order or magnitude.
3. **Ordinal Scale:** A scale that ranks data but does not specify the difference between ranks.
4. **Interval Scale:** A scale that measures the distance between points, without a true zero.
5. **Ratio Scale:** A scale that has equal intervals and a true zero, allowing full mathematical comparison.
6. **Quantitative Data:** Data that can be measured numerically.
7. **Qualitative Data:** Non-numerical data expressing qualities or characteristics.
8. **Respondent Cooperation:** The willingness of individuals to share truthful and accurate information during research.

9.7 SELF-ASSESSMENT QUESTIONS

1. Define a **measurement scale** and explain its significance in business research.
2. Differentiate between **quantitative** and **qualitative** measurements with examples.
3. What are the **four types of measurement scales**? Provide examples for each.

4. How does a **ratio scale** differ from an **interval scale**?
5. Why is it challenging to measure **attitudes and motivations** of respondents?
6. Explain two strategies a researcher can use to **gain respondent trust** during data collection.
7. Why is it important to align the **measurement scale** with the **research objective**?

9.8 SUGGESTED READINGS

1. Kothari, C. R. (2014). *Research Methodology: Methods and Techniques.* New Age International Publishers.
2. Malhotra, N. K. (2020). *Marketing Research: An Applied Orientation.* Pearson Education.
3. Cooper, D. R., & Schindler, P. S. (2017). *Business Research Methods.* McGraw-Hill Education.
4. Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students.* Pearson.
5. Tashakkori, A., & Teddlie, C. (2010). *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Sage Publications.

Dr. K. NAGA SUNDARI

LESSON-10

DATA COLLECTION

OBJECTIVES OF THE LESSON

By the end of this lesson, learners should be able to:

1. Identify and explain various methods of collecting primary data.
2. Distinguish between primary data and secondary data.
3. Evaluate the advantages and limitations of different data collection methods.
4. Select appropriate data collection techniques based on research objectives and context.

STRUCTURE OF THE LESSON

10.1 INTRODUCTION TO DATA COLLECTION

The task of data collection begins after a research problem has been defined and research design/ plan chalked out. While deciding about the method of data collection to be used for the study, the researcher should keep in mind two types of data viz., primary and secondary.

The primary data are those which are collected afresh and for the first time, and thus happen to be original in character. The secondary data, on the other hand, are those which have already been collected by someone else and which have already been passed through the statistical process. The researcher would have to decide which sort of data he would be using (thus collecting) for his study and accordingly he will have to select one or the other method of data collection.

The methods of collecting primary and secondary data differ since primary data are to be originally collected, while in case of secondary data the nature of data collection work is merely

that of compilation. We describe the different methods of data collection, with the pros and cons of each method.

10.2 METHODS OF COLLECTION OF PRIMARY DATA

We collect primary data during the course of doing experiments in an experimental research but in case we do research of the descriptive type and perform surveys, whether sample surveys or census surveys, then we can obtain primary data either through observation or through direct communication with respondents in one form or another or through personal interviews. This, in other words, means that there are several methods of collecting primary data, particularly in surveys and descriptive researches. Important ones are i) observation method, ii) interview method, iii) through questionnaires, iv) through schedules, and v) other methods which include a) warranty cards; b) distributor audit; c) pantry audits; d) consumer panels; e) using mechanical devices; f) through projective techniques; g) depth interviews, and (h) content analysis. We briefly take up each method separately.

10.2.1 Observation Method

The observation method is the most commonly used method especially in studies relating to behavioural sciences. In a way we all observe things around us, but this sort of observation is not scientific observation. Observation becomes a scientific tool and the method of data collection for the researcher when it serves a formulated research purpose, is systematically planned and recorded and is subjected to checks and controls on validity and reliability. Under the observation method, the information is sought by way of investigator's own direct observation without asking from the respondent. For instance, in a study relating to consumer behaviour, the investigator instead of asking the brand of wrist watch used by the respondent may himself look at the watch. The main advantage of this method is that subjective bias is eliminated, if observation is done accurately. Secondly, the information obtained under this method relates to what is currently happening; it is not complicated by either the past behaviour or future intentions or attitudes. Thirdly, this method is independent of respondents' willingness to respond and as such is relatively less demanding of active cooperation on the part of respondents as happens to be the case in the interview or the questionnaire method. This method is particularly suitable in studies which deal with subjects (i.e., respondents) who are not capable of giving verbal reports of their feelings for one reason or the other.

However, observation method has various limitations. Firstly, it is an expensive method. Secondly, the information provided by this method is very limited. Thirdly, sometimes unforeseen factors may interfere with the observational task. At times, the fact that some people are rarely accessible to direct observation creates obstacle for this method to collect data effectively.

While using this method, the researcher should keep in mind things like: What should be observed? How the observations should be recorded? Or how the accuracy of observation can be ensured? In case the observation is characterised by a careful definition of the units to be observed, the style of recording the observed information, standardised conditions of observation and the selection of pertinent data of observation, then the observation is called as structured observation. But when observation is to take place without these characteristics to be thought of in advance, the same is termed as unstructured observation. Structured observation is considered

appropriate in descriptive studies, whereas in an exploratory study the observational procedure is most likely to be relatively unstructured.

We often talk about participant and non-participant types of observation in the context of studies particularly of social sciences. This distinction depends upon the observer's sharing or not sharing the life of the group he is observing. If the observer observes by making himself, more or less, a member of the group he is observing so that he can experience what the members of the group experience, the observation is called as the participant observation. But when the observer observes as a detached emissary without any attempt on his part to experience through participation what others feel, the observation of this type is often termed as non-participant observation. (When the observer is observing in such a manner that his presence may be unknown to the people he is observing; such an observation is described as disguised observation).

There are several merits of the participant type of observation**:** (i) the researcher is enabled to record the natural behaviour of the group, (ii) the researcher can even gather information which could not easily be obtained if he observes in a disinterested fashion, (iii) the researcher can even verify the truth of statements made by informants in the context of a questionnaire or a schedule. But there are also certain demerits of this type of observation viz., the observer may lose the objectivity to the extent he participates emotionally; the problem of observation-control is not solved; and it may narrow down the researcher's range of experience.

Sometimes we talk of controlled and uncontrolled observation. If the observation takes place in the natural setting, it may be termed as uncontrolled observation, but when observation takes place according to definite pre-arranged plans, involving experimental procedure, the same is then termed controlled observation. In non-controlled observation, no attempt is made to use precision instruments. The major aim of this type of observation is to get a spontaneous picture of life and persons. It has a tendency to supply naturalness and completeness of behaviour, allowing sufficient time for observing it. But in controlled observation, we use mechanical (or precision) instruments as aids to accuracy and standardisation. Such observation has a tendency to supply formalised data upon which generalisations can be built with some degree of assurance. The main pitfall of non-controlled observation is that of subjective interpretation. There is also the danger of having the feeling that we know more about observed phenomena than we actually do. Generally, controlled observation takes place in various experiments that are carried out in a laboratory or under controlled conditions, whereas uncontrolled observation is resorted to in case of exploratory researches.

10.2.2 Interview Method

The interview method of collecting data involves presentation of oral-verbal stimuli and reply in terms of oral-verbal responses. This method can be used through personal interviews and, if possible, through telephone interviews.

a) **Personal interviews:** Personal interview method requires a person known as the interviewer asking questions generally in a face-to-face contact to the other person or persons. (At times the interviewee may also ask certain questions and the interviewer responds to these, but usually the interviewer initiates the interview and collects the information). This sort of interview may be in the form of direct personal investigation or it may be indirect oral investigation. In the

case of direct personal investigation, the interviewer has to collect the information personally from the sources concerned. He has to be on the spot and has to meet people from whom the data have to be collected. This method is particularly suitable for intensive investigations. But in certain cases it may not be possible or worthwhile to contact directly the persons concerned or on account of the extensive scope of enquiry, the direct personal investigation technique may not be used. In such cases an indirect oral examination can be conducted under which the interviewer has to cross-examine other persons who are supposed to have knowledge about the problem under investigation and the information, obtained is recorded. Most of the commissions and committees appointed by government to carry on investigations make use of this method.

The method of collecting information through personal interviews is usually carried out in a structured way. As such we call the interviews as structured interviews. Such interviews involve the use of a set of predetermined questions and of highly standardised techniques of recording. Thus, the interviewer in a structured interview follows a rigid procedure laid down, asking questions in a form and order prescribed. As against it, the unstructured interviews are characterised by a flexibility of approach to questioning. Unstructured interviews do not follow a system of pre-determined questions and standardised techniques of recording information. In a non-structured interview, the interviewer is allowed much greater freedom to ask, in case of need, supplementary questions or at times he may omit certain questions if the situation so requires. He may even change the sequence of questions. He has relatively greater freedom while recording responses to include some aspects and exclude others. But this sort of flexibility results in lack of comparability of one interview with another and the analysis of unstructured responses becomes much more difficult and time-consuming than that of the structured responses obtained in case of structured interviews. Unstructured interviews also demand deep knowledge and greater skill on the part of the interviewer. Unstructured interview, however, happens to be the central technique of collecting information in case of exploratory or formulative research studies. But in case of descriptive studies, we quite often use the technique of structured interview because of its being more economical, providing a safe basis for generalisation and requiring relatively lesser skill on the part of the interviewer.

We may as well talk about focussed interview, clinical interview and the non-directive interview. Focussed interview is meant to focus attention on the given experience of the respondent and its effects. Under it the interviewer has the freedom to decide the manner and sequence in which the questions would be asked and has also the freedom to explore reasons and motives. The main task of the interviewer in case of a focussed interview is to confine the respondent to a discussion of issues with which he seeks conversance. Such interviews are used generally in the development of hypotheses and constitute a major type of unstructured interviews. The clinical interview is concerned with broad underlying feelings or motivations or with the course of individual's life experience. The method of eliciting information under it is generally left to the interviewer's discretion. In case of non-directive interview, the interviewer's function is simply to encourage the respondent to talk about the given topic with a bare minimum of direct questioning. The interviewer often acts as a catalyst to a comprehensive expression of the respondents' feelings and beliefs and of the frame of reference within which such feelings and beliefs take on personal significance.

Despite the variations in interview-techniques, the major advantages and weaknesses of personal interviews can be enumerated in a general way. The chief merits of the interview method are as follows:

a) More information and that too in greater depth can be obtained.
b) Interviewer by his own skill can overcome the resistance, if any, of the respondents; the interview method can be made to yield an almost perfect sample of the general population.
c) There is greater flexibility under this method as the opportunity to restructure questions is always there, especially in case of unstructured interviews.
d) Observation method can as well be applied to recording verbal answers to various questions.
e) Personal information can as well be obtained easily under this method.
f) Samples can be controlled more effectively as there arises no difficulty of the missing returns; non-response generally remains very low.
g) The interviewer can usually control which person(s) will answer the questions. This is not possible in mailed questionnaire approach. If so desired, group discussions may also be held.
h) The interviewer may catch the informant off-guard and thus may secure the most spontaneous reactions than would be the case if mailed questionnaire is used.
i) The language of the interview can be adopted to the ability of educational level of the person interviewed and as such misinterpretations concerning questions can be avoided.
j) The interviewer can collect supplementary information about the respondent's personal characteristics and environment which is often of great value in interpreting results.

But there are also certain weaknesses of the interview method. Among the important weaknesses, mention may be made of the following:

a) It is a very expensive method, especially when large and widely spread geographical sample is taken.
b) There remains the possibility of the bias of interviewer as well as that of the respondent; there also remains the headache of supervision and control of interviewers.
c) Certain types of respondents such as important officials or executives or people in high income groups may not be easily approachable under this method and to that extent the data may prove inadequate.
d) This method is relatively more-time-consuming, especially when the sample is large and recalls upon the respondents are necessary.
e) The presence of the interviewer on the spot may over-stimulate the respondent, sometimes even to the extent that he may give imaginary information just to make the interview interesting.
f) Under the interview method the organisation required for selecting, training and supervising the field-staff is more complex with formidable problems.
g) Interviewing at times may also introduce systematic errors.
h) Effective interview presupposes proper rapport with respondents that would facilitate free and frank responses. This is often a very difficult requirement.

Pre-requisites and basic tenets of interviewing: For successful implementation of the interview method, interviewers should be carefully selected, trained and briefed. They should be honest, sincere, hardworking, impartial and must possess the technical competence and necessary practical experience. Occasional field checks should be made to ensure that interviewers are neither cheating, nor deviating from instructions given to them for performing their job efficiently. In addition, some provision should also be made in advance so that appropriate action may be taken if some of the selected respondents refuse to cooperate or are not available when an interviewer calls upon them.

In fact, interviewing is an art governed by certain scientific principles. Every effort should be made to create friendly atmosphere of trust and confidence, so that respondents may feel at ease while talking to and discussing with the interviewer. The interviewer must ask questions properly and intelligently and must record the responses accurately and completely. At the same time, the interviewer must answer legitimate question(s), if any, asked by the respondent and must clear any doubt that the latter has. The interviewers approach must be friendly, courteous, conversational and unbiased. The interviewer should not show surprise or disapproval of a respondent's answer but he must keep the direction of interview in his own hand, discouraging irrelevant conversation and must make all possible effort to keep the respondent on the track.

b) Telephonic interviews: This method of collecting information consists in contacting respondents on telephone itself. It is not a very widely used method, but plays important part industrial surveys, particularly in developed regions. The chief merits of such system are:

1. It is more flexible in comparison to mailing method.
2. It is faster than other methods i.e., a quick way of obtaining information.
3. It is cheaper than personal interviewing method; here the cost per response is relatively low.
4. Recall is easy; call backs are simple and economical.
5. There is higher rate of response than what we have in mailing method; the non-response is generally very low.
6. Replies can be recorded without causing embarrassment to respondents.
7. Interviewer can explain requirements more easily.
8. At times, access can be gained to respondents who otherwise cannot be contacted for one reason or the other.
9. No field staff is required.
10. Representative and wider distribution of sample is possible.

But this system of collecting information is not free from demerits. Some of these may be highlighted.

1. Little time is given to respondents for considered answers; interview period is not likely to exceed five minutes in most cases.
2. Surveys are restricted to respondents who have telephone facilities.
3. Extensive geographical coverage may get restricted by cost considerations.
4. It is not suitable for intensive surveys where comprehensive answers are required to various questions.
5. Possibility of the basis of the interviewer is relatively more.

6. Questions have to be short and to the point; probes are difficult to handle.

10.2.3 Collection of Data through Questionnaires:

This method of data collection is quite popular, particularly in case of big enquiries. It is being adopted by private individuals, research workers, private and public organisations and even by governments. In this method a questionnaire is sent (usually by post) to the persons concerned with a request to answer the questions and return the questionnaire. A questionnaire consists of a number of questions printed or typed in a definite order on a form or set of forms. The questionnaire is mailed to respondents who are expected to read and understand the questions and write down the reply in the space meant for the purpose in the questionnaire itself. The respondents have to answer the questions on their own.

The method of collecting data by mailing the questionnaires to respondents is most extensively employed in various economic and business surveys. The merits claimed on behalf of this method are as follows:

1. There is low cost even when the universe is large and is widely spread geographically.
2. It is free from the bias of the interviewer; answers are in respondents' own words.
3. Respondents have adequate time to give well thought out answers.
4. Respondents, who are not easily approachable, can also be reached conveniently.
5. Large samples can be made use of and thus the results can be made more dependable and reliable.

The main demerits of this system can also be listed here:

1. Low rate of return of the duly filled in questionnaires; bias due to no-response as often
2. It can be used only when respondents are educated and cooperating.
3. The control over questionnaire may be lost once it is sent.
4. There is inbuilt inflexibility because of the difficulty of amending the approach once questionnaires have been despatched.
5. There is also the possibility of ambiguous replies or omission of replies altogether to certain questions; interpretation of omissions is difficult.
6. It is difficult to know whether willing respondents are truly representative.
7. This method is likely to be the slowest of all.

Before using this method, it is always advisable to conduct 'pilot study' (Pilot Survey) for testing the questionnaires. In a big enquiry the significance of pilot survey is felt very much. Pilot survey is in fact the replica and rehearsal of the main survey. Such a survey, being conducted by experts, brings to the light the weaknesses (if any) of the questionnaires and also of the survey techniques. From the experience gained in this way, improvement can be effected.

Main aspects of a questionnaire: Quite often questionnaire is considered as the heart of a survey operation. Hence it should be very carefully constructed. If it is not properly set up, then the survey is bound to fail. This fact requires us to study the main aspects of a questionnaire viz., the general form, question sequence and question formulation and wording. Researcher should note the following with regard to these three main aspects of a questionnaire:

1. **General form:** So far as the general form of a questionnaire is concerned, it can either be structured or unstructured questionnaire. Structured questionnaires are those questionnaires in which there are definite, concrete and pre-determined questions. The questions are presented with exactly the same wording and in the same order to all respondents. Resort is taken to this sort of standardisation to ensure that all respondents reply to the same set of questions. The form of the question may be either closed (i.e., of the type 'yes' or 'no') or open (i.e., inviting free response) but should be stated in advance and not constructed during questioning. Structured questionnaires may also have fixed alternative questions in which responses of the informants are limited to the stated alternatives. Thus a highly structured questionnaire is one in which all questions and answers are specified and comments in the respondent's own words are held to the minimum. When these characteristics are not present in a questionnaire, it can be termed as unstructured or non-structured questionnaire. More specifically, we can say that in an unstructured questionnaire, the interviewer is provided with a general guide on the type of information to be obtained, but the exact question formulation is largely his own responsibility and the replies are to be taken down in the respondent's own words to the extent possible; in some situations tape recorders may be used to achieve this goal.

Structured questionnaires are simple to administer and relatively inexpensive to analyse. The provision of alternative replies, at times, helps to understand the meaning of the question clearly. But such questionnaires have limitations too. For instance, wide range of data and that too respondent’s own words cannot be obtained with structured questionnaires. They are usually considered inappropriate in investigations where the aim happens to be to probe for attitudes and reasons for certain actions or feelings. They are equally not suitable when a problem is being first explored and working hypotheses sought. In such situations, unstructured questionnaires may be used effectively. Then on the basis of the results obtained in pre-test (testing before final use) operations from the use of unstructured questionnaires, one can construct a structured questionnaire for use in the main study.

2. **Question sequence:** In order to make the questionnaire effective and to ensure quality to the replies received a researcher should pay attention to the question-sequence in preparing the questionnaire. A proper sequence of questions reduces considerably the chances of individual questions being misunderstood. The question-sequence must be clear and smoothly-moving, meaning thereby that the relation of one question to another should be readily apparent to the respondent, with questions that are easiest to answer being put in the beginning. The first few questions are particularly important because they are likely to influence the attitude of the respondent and in seeking his desired cooperation. The opening questions should be such as to arouse human interest. The following type of questions should generally be avoided as opening questions in a questionnaire:

1. Questions that put too great a strain on the memory or intellect of the respondent:

2. Questions of a personal character

3. Questions related to personal wealth, etc.

Following the opening questions, we should have questions that are really vital to the research problem and a connecting thread should run through successive questions. Ideally, the question-sequence should conform to the respondent's way of thinking. Knowing what information

is desired, the researcher can rearrange the order of the questions (this is possible in case of unstructured questionnaire) to fit the discussion in each particular case. But in a structured questionnaire the best that can be done is to determine the question sequence with the help of a Pilot Survey which is likely to produce good rapport with most respondents. Relatively difficult questions must be relegated towards the end so that even if the respondent decides not to answer such questions, considerable information would have already been obtained. Thus, question-sequence should usually go from the general to the more specific and the researcher must always remember that the answer to a given question is a function not only of the question itself, but of all previous questions as well. For instance, if one question deals with the price usually paid for coffee and the next with reason for preferring that particular brand, the answer to this latter question may be couched largely in terms of price differences.

3. Question formulation and wording: With regard to this aspect of questionnaire, the researcher should note that each question the very clear for any sort of misunderstanding can do irreparable harm to a survey. Question should also be impartial in order not to give a biased picture of the true state of affairs. Question should be constructed with a view to their forming a logical part of a well thought out tabulation plan. In general, all questions should meet the following standards (a) should be easily understood; (b) should be simple i.e., should convey only one thought at a time; (c) should be concrete and should conform as much as possible to the respondent's way of thinking. For instance, instead of asking. “How many razor blades do you use annually?" The more realistic question would be to ask, “How many razor blades did you use last week?"

Concerning the form of questions, we can talk about two principal forms, viz., multiple choice questions and the open-end question. In the former the respondent selects one of the alternative possible answers put to him, whereas in the latter he has to supply the answer in his own words. The question with only two possible answers (usually 'Yes' or 'No') can be taken as a special case of the multiple choice question, or can be named as a "closed question. There are some advantages and disadvantages of each possible form of question. Multiple choice or closed questions have the advantages of easy handling, simple to answer, quick and relatively inexpensive to analyse. They are most amenable to statistical analysis. Sometimes, the provision of alternative replies helps to make clear the meaning of the question. But the main drawback of fixed alternative questions is that of "putting answers in people's mouths" i.e., they may force a statement of opinion on an issue about which the respondent does not in fact have any opinion. They are not appropriate when the issue under consideration happens to be a complex one and also when the interest of the researcher is in the exploration of a process. In such situations, open-ended questions which are designed to permit a free response from the respondent rather than one limited to certain stated alternatives are considered appropriate. Such questions give the respondent considerable latitude in phrasing a reply. Getting the replies in respondent's own words is, thus, the major advantage of open-ended questions. But one should not forget that, from an analytical point of view, open-ended questions are more difficult to handle, raising problems of interpretation, comparability and interviewer bias.

In practice, one rarely comes across a case when one questionnaire relies on one form of questions alone. The various forms complement each other. As such questions of different forms are included in one single questionnaire. For instance, multiple-choice questions constitute the basis of a structured questionnaire, particularly in a mail survey. But even there, various open-

ended questions are generally inserted to provide a more complete picture of the respondent's feelings and attitudes.

Researcher must pay proper attention to the wordings of questions since reliable and meaningful returns depend on it to a large extent. Since words are likely to affect responses, they should be properly chosen. Simple words, which are familiar to all respondents, should be employed. Words with ambiguous meanings must be avoided. Similarly, danger words, catch-words or words with emotional connotations should be avoided. Caution must also be exercised in the use of phrases which reflect upon the prestige of the respondent. Question wording, in no case, should bias the answer. In fact, question wording and formulation is an art and can only be learnt by practice.

Essentials of a good questionnaire: To be successful, questionnaire should be comparatively short and simple i.e., the size of the questionnaire should be kept to the minimum. Questions should proceed in logical sequence moving from easy to more difficult questions. Personal and intimate questions should be left to the end. Technical terms and vagut expressions capable of different interpretations should be avoided in a questionnaire. Questions may be dichotomous (yes or no answers), multiple choice (alternative answers listed) or open-ended. The latter types of questions are often difficult to analyse and hence should be avoided in a questionnaire to the extent possible. There should be some control questions in the questionnaire which indicate the reliability of the respondent. For instance, a question designed to determine the consumption of particular material may be asked first in terms of financial expenditure and later in terms of weight. The control questions, thus, introduce a cross-check to see whether the information collected is correct or not. Questions affecting the sentiments of respondents should be avoided. Adequate space for answers should be provided in the questionnaire to help editing and tabulation. There should always be provision for indications of uncertainty, e.g., “do not know,” “no preference" and so on. Brief directions with regard to filling up the questionnaire should invariably be given in the questionnaire itself. Finally, the physical appearance of the questionnaire affects the cooperation the researcher receives from the recipients and as such an attractive looking questionnaire, particularly in mail surveys, is a plus point for enlisting cooperation. The quality of the paper, along with its colour, must be good so that it may attract the attention of recipients.

10.2.4 Collection of Data through Schedules:

This method of data collection is very much like the collection of data through questionnaire, with little difference which lies in the fact that schedules (proforma containing a set of questions) are being filled in by the enumerators who are specially appointed for the purpose. These enumerators along with schedule go to respondents, put to them the questions from the proforma in the order the questions are listed and record the replies in the space meant for the same in the proforma. In certain situations, schedules may be handed over to respondents and enumerators may help them in recording their answers to various questions in the said schedules. Enumerators explain the aims and objects of the investigation and also remove the difficulties which any respondent may feel in understanding the implications of a particular question or the definition or concept of difficult terms.

This method requires the selection of enumerators for filling up schedules or assisting respondents to fill up schedules and as such enumerators should be very carefully selected. The enumerators should be trained to perform their job well and the nature and scope of the

investigation should be explained to them thoroughly so that they may well understand the implications of different questions put in the schedule. Enumerators should be intelligent and must possess the capacity of cross examination in order to find out the truth. Above all, they should be honest, sincere, hardworking and should have patience and perseverance.

This method of data collection is very useful in extensive enquiries and can lead to fairly reliable results. It is however, very expensive and is usually adopted in investigations conducted by governmental agencies or by some big organisations. Population census all over the world is conducted through this method.

10.2.5 Difference between Questionnaires and Schedules:

Both questionnaire and schedule are popularly used methods of collecting data in research surveys. There is much resemblance in the nature of these two methods and this fact has made many people to remark that from a practical point of view, the two methods can be taken to be the same. But from the technical point of view there is difference between the two. The important points of difference are as under.

1. The questionnaire is generally sent through mail to informants to be answered as specified in a covering letter, but otherwise without further assistance from the sender. The schedule is generally filled out by the research worker or the enumerator, who can interpret questions when necessary.
2. To collect data through questionnaire is relatively cheap and economical since we have to spend money only in preparing the questionnaire and in mailing the same to respondents. Here no field staff required. To collect data through schedules is relatively more expensive since considerable amount of money has to be spent in appointing enumerators and in importing training to them. Money is also spent in preparing schedules.
3. Non-response is usually high in case of questionnaire as many people do not respond and many return the questionnaire without answering all questions. Bias due to non-response often remains indeterminate. As against this, non-response is generally very low in case of schedules because these are filled by enumerators who are able to get answers to all questions. But there remains the danger of interviewer bias and cheating.
4. In case of questionnaire, it is not always clear as to who replies, but in case of schedule the identity of respondent is known.
5. The questionnaire method is likely to be very slow since many respondents do not return the questionnaire in time despite several reminders, but in case of schedules the information is collected well in time as they are filled in by enumerators.
6. Personal contact is generally not possible in case of the questionnaire method as questionnaires are sent to respondents by post who also in turn returns the same by post. But in case of schedules direct personal contact is established with respondents.
7. Questionnaire method can be used only when respondents are literate and cooperative, but in case of schedules the information can be gathered even when the respondents happen to be illiterate.
8. Wider and more representative distribution of sample is possible under the questionnaire method, but in respect of schedules there usually remains the difficulty in sending enumerators over a relatively wider area.

9. Risk of collecting incomplete and wrong information is relatively more under the questionnaire method, particularly when people are unable to understand questions properly. But in case of schedules, the information collected is generally complete and accurate as enumerators can remove the difficulties, if any, faced by respondents in correctly understanding the questions. As a result, the information collected through schedules is relatively more accurate than that obtained through questionnaires.
10. The success of questionnaire method lies more on the quality of the questionnaire itself, but in the case of schedules much depends upon the honesty and competence of enumerators.
11. In order to attract the attention of respondents, the physical appearance of questionnaire must be quite attractive, but this may not be so in case of schedules as they are to be filled in by enumerators and not by respondents.
12. Along with schedules, observation method can also be used but such a thing is not possible while collecting data through questionnaires.

10.3 COLLECTION OF SECONDARY DATA

Secondary data means data that are already available i.e., they refer to the data which have already been collected and analysed by someone else. When the researcher utilises secondary data, then he has to look into various sources from where he can obtain them. In this case he is certainly not confronted with the problems that are usually associated with the collection of original data. Secondary data may either be published data or unpublished data. Usually published data are available in: (a) various publications of the central, state are local governments; (b) various publications of foreign governments or of international bodies and their subsidiary organisations; (c) technical and trade journals; (d) books, magazines and newspapers; (e) reports and publications of various associations connected with business and industry, banks, stock exchanges, etc.; (f) reports prepared by research scholars, universities, economists, etc, in different fields; and (g) public records and statistics, historical documents, and other sources of published information. The sources of unpublished data are many; they may be found in diaries, letters, unpublished biographies and autobiographies and also may be available with scholars and research workers, trade associations, labour bureaus and other public/private individuals and organisations.

Researcher must be very careful in using secondary data. He must make a minute scrutiny because it is just possible that the secondary data may be unsuitable or may be inadequate in the context of the problem which the researcher wants to study. In this connection Dr. A.L. Bowley very aptly observes that it is never safe to take published statistics at their face value without knowing their meaning and limitations and it is always necessary to criticise arguments that can be based on them.

By way of caution, the researcher, before using secondary data, must see that they possess following characteristics:

1. Reliability of data: The reliability can be tested by finding out such things about the said data: (a) Who collected the data? (b) What were the sources of data? (c) Were they collected by using proper methods (d) At what time were they collected? (e) Was there any bias of the compiler? (f) What level of accuracy was desired? Was it achieved?

2. Suitability of data: The data that are suitable for one enquiry may not necessarily be found suitable in another enquiry. Hence, if the available data are found to be unsuitable, they should not

be used by the researcher. In this context, the researcher must very carefully scrutinise the definition of various terms and units of collection used at the time of collecting the data from the primary source originally. Similarly, the object, scope and nature of the original enquiry must also be studied. If the researcher finds differences in these, the data will remain unsuitable for the present enquiry and should not be used.

3. Adequacy of data: If the level of accuracy achieved in data is found inadequate for the purpose of the present enquiry, they will be considered as inadequate and should not be used by the researcher. The data will also be considered inadequate, if they are related to an area which may be either narrower or wider than the area of the present enquiry.

From all this we can say that it is very risky to use the already available data. The already available data should be used by the researcher only when he finds them reliable, suitable and adequate. But he should not blindly discard the use of such data if they are readily available from authentic sources and are also suitable and adequate for in that case it will not be economical to spend time and energy in field surveys for collecting information. At times, there may be wealth of usable information in the already available data which must be used by an intelligent researcher but with due precaution.

10.4 PROCESSING AND ANALYSIS OF DATA

The data, after collection, has to be processed and analysed in accordance with the outline laid down for the purpose at the time of developing the research plan. This is essential for a scientific study and for ensuring that we have all relevant data for making contemplated comparisons and analysis. Technically speaking, processing implies editing, coding, classification and tabulation of collected data so that they are amenable to analysis. The term analysis refers to the computation of certain measures along with searching for patterns of relationship that exist among data-groups. Thus, in the process of analysis, relationships or differences supporting or conflicting with original or new hypotheses should be subjected to statistical tests of significance to determine with what validity data can be said to indicate any conclusions". But there are persons (Selltiz, Jahoda and others) who do not like to make difference between processing and analysis. They opine that analysis of data in a general way involves a number of closely related operations which are performed with the purpose of summarising the collected data and organising these in such a manner that they answer the research question(s). We, however, shall prefer to observe the difference between the two terms as stated here in order to understand their implications more clearly.

Processing Operations:

With this brief introduction concerning the concepts of processing and analysis, we can now proceed with the explanation of all the processing operations.

1. **Editing:** Editing of data is a process of examining the collected raw data (especially in surveys) to detect errors and omissions and to correct these when possible. As a matter of fact, editing involves a careful scrutiny of the completed questionnaires and/or schedules. Editing is done to assure that the data are accurate, consistent with other facts gathered, uniformly entered, as completed as possible and have been well arranged to facilitate coding and tabulation.

With regard to points or stages at which editing should be done, one can talk of field editing and central editing. Field editing consists in the review of the reporting forms by the investigator for completing (translating or rewriting) what the latter has written in abbreviated and/or in illegible form at the time of recording the respondents' responses. This type of editing is necessary in view of the fact that individual writing styles often can be difficult for others to decipher. This sort of editing should be done as soon as possible after the interview, preferably on the very day or on the next day. While doing field editing the investigator must restrain himself and must not correct errors of omission by simply guessing what the informant would have said if the question had been asked.

Central editing should take place when all forms of schedules have been completed and returned to the office. This type of editing implies that all forms should get a thorough editing by a single editor in a small study and by a team of editors in case of a large inquiry. Editor(s) may correct the obvious errors such as an entry in the wrong place, entry recorded in months when it should have been recorded in weeks, and the like. In case of inappropriate on missing replies, the editor can sometimes determine the proper answer by reviewing the other information in the schedule. At times, the respondent can be contacted for clarification. The editor must strike out the answer if the same is inappropriate and he has no basis for determining the correct answer or the response. In such a case an editing entry of 'no answer' is called for. All the wrong replies, which are quite obvious, must be dropped from the final results, especially in the context of mail surveys.

Editors must keep in view several points while performing their work: (a) They should be familiar with instructions given to the interviewers and coders as well as with the editing instructions supplied to them for the purpose. (b) While crossing out an original entry for one reason or another, they should just draw a single line on it so that the same may remain legible. (c) They must make entries (if any) on the form in some distinctive colour and that too in a standardised form. (d) They should initial all answers which they change or supply. (e) Editor's initials and the date of editing should be placed on each completed form or schedule.

2. **Coding:** Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness (i.e., there must be a class for every data item) and also that of mutual exclusively which means that a specific answer can be placed in one and only one cell in a given category set. Another rule to be observed is that of unidimensionality by which is meant that every class is defined in terms of only one concept.

   Coding is necessary for efficient analysis and through it the several replies may be reduced to a small number of classes which contain the critical information required for analysis. Coding decisions should usually be taken at the designing stage of the questionnaire. This makes it possible to pre-code the questionnaire choices and which in turn is helpful for computer tabulation as one can straight forward key punch from the original questionnaires. But in case of hand coding some standard method may be used. One such standard method is to code in the margin with a coloured pencil. The other method can be to transcribe the data from the

questionnaire to a coding sheet. Whatever method is adopted; one should see that coding errors are altogether eliminated or reduced to the minimum level.

3. **Classification:** Most research studies result in a large volume of raw data which must be reduced into homogeneous groups if we are to get meaningful relationships. This fact necessitates classification of data which happens to be the process of arranging data in groups or classes on the basis of common characteristics. Data having a common characteristic are placed in one class and in this way the entire data get divided into a number of groups or classes. Classification can be one of the following two types, depending upon the nature of the phenomenon involved**:**

(a) Classification according to attributes: As stated above, data are classified on the basis of common characteristics which can either be descriptive (such as literacy, sex, honesty, etc.) or numerical (such as weight, height, income, etc.). Descriptive characteristics refer to qualitative phenomenon which cannot be measured quantitatively, only their presence or absence in an individual item can be noticed. Data obtained this way on the basis of certain attributes are known as statistics of attributes and their classification is said to be classification according to attributes.

Such classification can be simple classification or manifold classification. In simple classification we consider only one attribute and divide the universe into two classes ---- one class consisting of items possessing the given attribute and the other class consisting of items which do not possess the given attribute. But in manifold classification we consider two or more attributes simultaneously, and divide that data into a number of classes (total number of classes of final order is given by 2^n, where n= no. of attributes considered). Whenever data are classified according to attributes, the researcher must see that the attributes are defined in such a manner that there is least possibility of any doubt/ambiguity concerning the said attributes.

(b) Classification according to class-intervals: Unlike descriptive characteristics, the numerical characteristics refer to quantitative phenomenon which can be measured through some statistical units. Data relating to income, production, age, weight, etc come under this category. Such data are known as statistics of variables and are classified on the basis of class intervals. For instance, persons whose incomes, say, are within Rs 201 to Rs 400 can form one group; those whose incomes are within Rs 401 to Rs 600 can form another group and so on. In this way the entire data may be divided into a number of groups or classes or what are usually called, 'class-intervals.' Each group of class-interval, thus, has an upper limit as well as a lower limit which are known as class limits. The difference between the two class limits is known as class magnitude. We may have classes with equal class magnitudes or with unequal class magnitudes. The number of items which fall in a given class is known as the frequency of the given class. All the classes or groups, with their respective frequencies taken together and put in the form of a table, are described as group frequency distribution or simply frequency distribution. Classification according to class intervals usually involves the following three main problems:

(i) How may classes should be there? What should be their magnitudes?

There can be no specific answer with regard to the number of classes. The decision about this calls for skill and experience of the researcher. However, the objective should be to display the data in such a way as to make it meaningful for the analyst. Typically, we may have 5 to 15 classes. With regard to the second part of the question, we can say that, to the extent possible, class-intervals should be of equal magnitudes, but in some cases unequal magnitudes may result in better classification. Hence the researcher's objective judgement plays an important part in this connection. Multiples of 2, 5 and 10 are generally preferred while determining class magnitudes. Some statisticians adopt the following formula, suggested by H.A. Sturges, determining the size of class interval**:**

$$i = R/(1 + 3.3 \log N)$$

Where,

i = size of class interval;

R= Range (i.e., difference between the values of the largest item and smallest item among the given items):

N = Number of items to be grouped.

It should also be kept in mind that in case one or two or very few items have very high or very low values, one may use what are known as open-ended intervals in the overall frequency distribution. Such intervals may be expressed like under R 500 or Rs 10001 and over. Such intervals are generally not desirable, but often cannot be avoided. The researcher must always remain conscious of this fact while deciding the issue of the total number of class intervals in which the data are to be classified

(ii) How to choose class limits?

While choosing class limits, the researcher must take into consideration the criterion that the mid-point (generally worked out first by taking the sum of the upper limit and lower limit of a class and then divide this sum by 2) of a class-interval and the actual average of items of that class interval should remain as close to each other as possible. Consistent with this, the class limits should be located at multiples of 2, 5, 10, 20, 100 and such other figures. Class limits may generally be stated in any of the following forms:

Exclusive type class intervals: They are usually stated as follows:

10-20
20-30
30-40
40-50

The above intervals should be read as under:

10 and under 20
20 and under 30
30 and under 40
40 and under 50

Thus, under the exclusive type class intervals, the items whose values are equal to the upper limit of a class are grouped in the next higher class. For example, an item whose value is exactly 30 would be put in 30-40 class interval and not in 20-30 class interval.

In simple words, we can say that under exclusive type class intervals, the upper limit of a class interval is excluded and items with values less than the upper limit (but not less than the lower limit) are put in the given class interval.

Inclusive type class intervals: They are usually stated as follows:

11-20
21-30
31-10
41-50

In inclusive type class intervals, the upper limit of a class interval is also included in the concerning class interval. Thus, an item whose value is 20 will be put in 11-20 class interval. The stated upper limit of the class interval 11-20 is 20 but the real limit is 20.99999 and as such 11-20 class interval really means 11 and under 21.

When the phenomenon under consideration happens to be a discrete one (i.e., can be measured and stated only in integers), then we should adopt inclusive type classification. But when the phenomenon happens to be a continuous one capable of being measured in fractions as well, we can use exclusive type class intervals.

(iii) How to determine the frequency of each class?

This can be done either by tally sheets or by mechanical aids. Under the technique or tally sheet, the class-groups are written on a sheet of paper (commonly known as the tally sheet) and for each item a stroke (usually a small vertical line) is marked against the class group in which it falls. The general practice is that after every four small vertical lines in a class group, the fifth line for the item falling in the same group is indicated as horizontal line through the said four lines and the resulting flower (IIII) represents five items. All this facilitates the counting of items in each one of the class groups. An illustrative tally sheet can be shown as under:

An Illustrative Tally Sheet for Determining the Number of 70 Families in Different Income Groups

Income groups (Rupees)	Tally mark	No. of families or (Class Frequency)
Below 400	𝍸 𝍸 III	13
401-800	𝍸 𝍸 𝍸 𝍸	20
801-1200	𝍸 𝍸 II	12
1201-1600	𝍸 𝍸 𝍸 III	18
1601 and above	𝍸 II	7
Total		70

Alternatively, class frequencies can be determined, especially in case of large inquires and surveys, by mechanical aids i.e., with the help of machines viz., sorting machines that are available for the purpose. Some machines are hand operated, whereas other work with electricity. There are machines which can sort out cards at a speed of something like 25000 cards per hour. This method is fast but expensive.

4. **Tabulation:** When a mass of data has been assembled, it becomes necessary for the researcher to arrange the same in some kind of concise and logical order. This procedure is referred to as tabulation. Thus, tabulation is the process of summarising raw data and displaying the same in compact form (i.e., in the form of statistical tables) for further analysis. In a broader sense, tabulation is an orderly arrangement of data in columns and rows.

   Tabulation is essential because of the following reasons.

   1. It conserves space and reduces explanatory and descriptive statement to a minimum.
   2. It facilitates the process of comparison.
   3. It facilitates the summation of items and the detection of errors and omissions.
   4. It provides a basis for various statistical computations.

   Tabulation can be done by hand or by mechanical or electronic devices. The choice depends on the size and type of study, cost considerations, time pressures and the availability of tabulating machines or computers. In relatively large inquiries, we may use mechanical or computer tabulation if other factors are favourable and necessary facilities are available. Hand tabulation is usually preferred in case of small inquiries where the number of questionnaires is small and they are of relatively short length. Hand tabulation may be done using the direct tally, the list and tally or the cards sort and count methods. When there are simple codes, it is feasible to tally directly from the questionnaire. Under this method, the codes are written on a sheet of paper, called tally sheet, and for each response a stroke is marked against the code in which it falls. Usually after every four strokes against a particular code, the fifth response is indicated by drawing a diagonal or horizontal line through the strokes. These groups of five are easy to count and the data are sorted against each code conveniently. In the listing method, the code responses may be transcribed onto a large work sheet, allowing a line for each questionnaire. This way a large number of questionnaires can be listed on one worksheet. Tallies are then made for each question. The card sorting method is the most flexible hand tabulation. In this method the data are recorded on special cards of convenient size and shape with a series of holes. Each hole stands for a code and when cards are stacked, a needle passes through particular hole representing a particular code. These cards are then separated and counted. In this way frequencies of various codes can be found out by the repetition of this technique. We can as well use the mechanical devices or the computer facility for tabulation purpose in case we want quick results, our budget permits their use and we have a large volume of straightforward tabulation involving a number of cross breaks.

   Tabulation may also be classified as simple and complex tabulation. The former type of tabulation gives information about one or more groups of independent questions, whereas the latter type of tabulation shows the division of data in two or more categories and as such is designed to give information concerning one or more sets of inter-related questions. Simple tabulation generally results in one-way tables which supply answers to questions about one characteristic of data only. As against this, complex tabulation usually results in two-way tables (which give information about two inter-related characteristics of data), three-way tables giving information about three interrelated characteristics of data) or still higher order tables, also known as manifold tables, which supply information about

several interrelated characteristics of data. Two-way tables, three-way tables or manifold tables are all examples of what is sometimes described as cross tabulation.

Generally accepted principles of tabulation: Such principles of tabulation, particularly of constructing statistical tables, can briefly states as follows:

1. Every table should have a clear, concise and adequate title so as to make the table intelligible without reference to the text and this title should always be placed just above the body of the table.
2. Every table should be given a distinct number to facilitate easy reference.
3. The column headings (captions) and the row headings (stubs) of the table should be clear and brief.
4. The units of measurement under each heading or sub-heading must always be indicated.
5. Explanatory footnotes, if any, concerning the table should be placed directly beneath the table, along with the reference symbols used in the table.
6. Source of sources from where the data in the table have been obtained must be indicated just below the table.
7. Usually the columns are separated from one another by lines which make the table more readable and attractive. Lines are always drawn at the top and bottom of the table and below the captions.
8. There should be thick lines to separate the data under one class from the data under another class and the lines separating the sub-divisions of the classes should be comparatively thin lines.
9. The columns may be numbered to facilitate reference.
10. Those columns whose data are to be compared should be kept side by side. Similarly, percentages and/or averages must also be kept close to the data.
11. It is generally considered better to approximate figures before tabulation as the same would reduce unnecessary details in the table itself.
12. In order to emphasise the relative significance of certain categories, different kinds of type, spacing and indentations may be used.
13. It is important that all column figures be properly aligned. Decimal points and (+) or ( ) signs should be in perfect alignment.
14. Abbreviations should be avoided to the extent possible and ditto marks should not be used in the table.
15. Miscellaneous and exceptional items, if any, should be usually placed in the last row of the table.
16. Table should be made as logical, clear, accurate and simple as possible. If the data happen to be very large, they should not be crowded in a single table for that would make the table unwieldy and inconvenient.
17. Total of rows should normally be placed in the extreme right column and that of columns should be placed at the bottom.
18. The arrangement of the categories in a table may be chronological, geographical, alphabetical or according to magnitude to facilitate comparison. Above all, the table must suit the needs and requirements of an investigation.

10.5 SUMMARY

Data collection is a fundamental stage of the research process that begins after defining the research problem and designing the research plan. Data can be classified into primary data, which are collected firsthand for a specific purpose, and secondary data, which are pre-existing data originally collected by others. Primary data can be collected through several methods such as observation, interviews, questionnaires, and schedules, as well as through specialized techniques like consumer panels, distributor audits, and projective methods. Secondary data are gathered from existing sources, including government publications, research reports, trade journals, books, and unpublished documents such as letters, diaries, or organizational records. Choosing the right method depends on factors such as the nature of the research, availability of data, resources, and time constraints. Effective data collection ensures reliability, accuracy, and validity of research findings.

10.6 TECHNICAL TERMS

1. **Primary Data:** Data collected firsthand for a specific research purpose.
2. **Secondary Data:** Data that already exist and have been collected by others for different purposes.
3. **Observation Method:** Collecting data by watching and recording behaviors or events as they occur.
4. **Interview Method:** Gathering information directly from respondents through structured or unstructured interaction.
5. **Questionnaire:** A written set of questions used to collect information from respondents.
6. **Schedule:** A structured form filled by an interviewer during personal interaction with respondents.
7. **Projective Techniques:** Indirect methods used to uncover underlying attitudes, motives, or feelings.
8. **Consumer Panel:** A group of consumers who provide continuous information over time about their behavior or opinions.
9. **Published Data:** Data made available in reports, journals, or official documents.
10. **Unpublished Data:** Data not formally published, often found in private or organizational records.

10.7 SELF-ASSESSMENT QUESTIONS

1. What is the difference between **primary** and **secondary** data?
2. List any five **methods of collecting primary data**.
3. Explain the **observation method** of data collection with an example.
4. What are the **advantages and limitations** of the interview method?
5. Mention at least four **sources of secondary data**.
6. How do **published** and **unpublished** data differ?
7. Why is it important to choose an appropriate **method of data collection** in research?
8. Give examples of situations where **secondary data** may be more useful than primary data.

10.8 SUGGESTED READINGS

1. Kothari, C. R. (2014). *Research Methodology: Methods and Techniques.* New Age International Publishers.
2. Malhotra, N. K. (2020). *Marketing Research: An Applied Orientation.* Pearson Education.
3. Cooper, D. R., & Schindler, P. S. (2017). *Business Research Methods.* McGraw-Hill Education.
4. Saunders, M., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students.* Pearson.
5. Tashakkori, A., & Teddlie, C. (2010). *Mixed Methodology: Combining Qualitative and Quantitative Approaches.* Sage Publications.

Dr. K. NAGA SUNDARI

LESSON-11

ANALYSIS OF VARIANCE (ANOVA), MULTIPLE CORRELATION AND REGRESSION

OBJECTIVES OF THE LESSON

After studying this lesson, students will be able to:

1. Understand the concept and purpose of Analysis of Variance (ANOVA).
2. Distinguish between one-way and two-way ANOVA and their applications.
3. Apply ANOVA to practical business, agricultural, or experimental research problems.
4. To understand how Multiple Correlation measures the combined strength of the relationship between one dependent variable and two or more independent variables.
5. To learn how Multiple Regression develops a predictive model that explains the effect of several independent variables on a single dependent variable.

STRUCTURE OF THE LESSON

11.1 INTRODUCTION TO ANOVA

Analysis of Variance (ANOVA) is a statistical method commonly used in research areas such as economics, biology, education, psychology, sociology, business, and industry. It is particularly helpful when comparing more than two groups or samples. While the t-test or z-test can compare the means of two samples, ANOVA determines whether significant differences exist among the means of three or more samples at the same time.

ANOVA allows a researcher to determine whether different samples originate from populations with the same mean, for example, if a researcher aims to compare the yield of

different seed varieties, the fuel efficiency of various car brands, etc. ANOVA offers a systematic way to test whether the differences observed among group means are statistically significant. It eliminates the need for multiple pairwise comparisons, which can be time-consuming, costly, and prone to interpretation errors.

Meaning of ANOVA

The terms variance and the technique of ANOVA were introduced by Professor R.A. Fisher, who laid its theoretical foundation and demonstrated its practical applications. Later, Professor Snedecor and others refined and expanded the method.

ANOVA primarily aims to test the homogeneity of various groups. The main concept is to split the total variation observed in the data into:

1. **Variation due to chance** (random variation within samples), and
2. **Variation due to specific causes** (differences between sample means due to a factor or treatment).

Thus, ANOVA analyses the total variance into different components attributable to various sources. It helps determine whether these sources significantly influence the dependent variable.

This technique is helpful in a variety of practical contexts, such as:

- Determining whether different types of fertilisers or seeds significantly affect crop yield.
- Comparing the effectiveness of different drugs on patients.
- Evaluating the performance of sales staff in an organisation.
- Comparing different training methods or teaching strategies in education research.

If only one factor with multiple categories is studied, the procedure is called One-Way ANOVA. When two factors are studied together, the method becomes Two-Way ANOVA, which also permits the examination of interaction effects between factors.

Basic Principle of ANOVA

The core principle of ANOVA is to examine the ratio of variation between groups to the variation within groups:

- Variation within samples is assumed to occur due to random or unexplained factors.
- Variation between sample means is assumed to arise due to the influence of a specific factor or treatment effect.

ANOVA works by making two independent estimates of population variance:

1. Variance Based on Between-Group Differences, and
2. Variance Based on Within-Group Differences.

These two estimates are compared using the **F-test**, where:

$$F = \frac{\text{Variance Between Groups}}{\text{Variance Within Groups}}$$

If the calculated F value is equal to or greater than the critical F value from the table at a specific significance level and degrees of freedom, we conclude that there are significant differences among the sample means. Otherwise, we conclude that all sample means are statistically the same.

Assumptions of ANOVA

For the ANOVA test to be valid, the following assumptions must be satisfied:

1. The populations from which samples are drawn are normally distributed.
2. The populations have equal variances (homogeneity of variance).
3. Samples are randomly and independently drawn.
4. Factors other than those tested are controlled or held constant.

11.2 ONE-WAY (OR SINGLE FACTOR) ANOVA:

Procedure of One-Way ANOVA

In a One-Way ANOVA, we study the effect of a single factor on a dependent variable. The factor may have several categories or groups (samples), and our objective is to determine whether the group means differ significantly.

The steps involved in conducting One-Way ANOVA are as follows:

1. Calculate the mean of each of the *k* samples:

$$\bar{x}_1, \bar{x}_2, \bar{x}_3, \dots, \bar{x}_k$$

2. Compute the Grand Mean: The grand mean is the mean of all sample means, given by:

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \cdots + \bar{x}_k}{k}$$

3. Calculate the Sum of Squares Between Samples (SS Between) For each sample mean, find its deviation from the grand mean, square it, multiply by the number of observations in that sample, and then sum these values:

$$SS_{\text{Between}} = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(\bar{x}_2 - \bar{\bar{x}})^2 + \cdots + n_k(\bar{x}_k - \bar{\bar{x}})^2$$

4. Compute Mean Square Between Samples (MS Between) Divide SS Between by its degrees of freedom (k – 1):

$$MS_{\text{Between}} = \frac{SS_{\text{Between}}}{k-1}$$

5. Calculate the Sum of Squares Within Samples (SS Within) For each sample, take the deviation of each observation from its sample mean, square these deviations, and sum them for all samples:

$$SS_{\text{Within}} = \sum(x_{1i} - \bar{x}_1)^2 + \sum(x_{2i} - \bar{x}_2)^2 + \cdots + \sum(x_{ki} - \bar{x}_k)^2$$

6. Compute Mean Square Within Samples (MS Within). Divide SS Within by its degrees of freedom (n – k), where *n* is the total number of observations in all samples:

$$MS_{\text{Within}} = \frac{SS_{\text{Within}}}{n-k}$$

7. Check the Total Sum of Squares (Optional Verification): The total variation should satisfy:

$$SS_{\text{Total}} = SS_{\text{Between}} + SS_{\text{Within}}$$

8. Degrees of freedom also follow the additive rule:

$$(n-1) = (k-1) + (n-k)$$

9. Calculate the F-Ratio

$$F = \frac{MS_{\text{Between}}}{MS_{\text{Within}}}$$

10. Decision Rule: Compare the calculated F value with the critical F value from the F-table at a chosen significance level:
    a. If F calculated < F table → Differences among sample means are not significant. The groups are assumed to come from the same population.
    b. If F calculated ≥ F table → Differences are significant, meaning the sample means are not equal, and the factor has a real effect.

Interpretation

- A higher F-value (greater than the table value) provides more substantial evidence that there are significant differences between the sample group means.
- If the F-value is small, any observed differences are likely due to random chance.

Illustration 1:

A researcher wants to study the effect of wheat variety on per-acre production. The data for three varieties of wheat (A, B, C), each grown on four plots, are as follows:

Plot of land	*Per-acre production data*		
	Variety of wheat		
	A	***B***	***C***
1	6	5	5
2	7	5	4
3	3	3	3
4	8	7	4

Solution:

Before performing ANOVA, we define the null and alternative hypotheses:

Null Hypothesis (H_0): There is no significant difference in per-acre production among the three wheat varieties.

$$\mu_A = \mu_B = \mu_C$$

Alternative Hypothesis (H_1): At least one variety differs significantly in per-acre production.

Step 1: Compute the Mean of Each Variety

$$\bar{x}_A = \frac{6+7+3+8}{4} = 6$$
$$\bar{x}_B = \frac{5+5+3+7}{4} = 5$$
$$\bar{x}_C = \frac{5+4+3+4}{4} = 4$$

Step 2: Compute the Grand Mean

$$\bar{\bar{x}} = \frac{\bar{x}_A + \bar{x}_B + \bar{x}_C}{3} = \frac{6+5+4}{3} = 5$$

Step 3: Calculate Sum of Squares Between Samples (SS Between)

$$SS_{Between} = n_A(\bar{x}_A - \bar{\bar{x}})^2 + n_B(\bar{x}_B - \bar{\bar{x}})^2 + n_C(\bar{x}_C - \bar{\bar{x}})^2$$
$$SS_{Between} = 4(6-5)^2 + 4(5-5)^2 + 4(4-5)^2 = 4 + 0 + 4 = 8$$

Step 4: Calculate Sum of Squares Within Samples (SS Within)

$$SS_{Within} = \sum(x_{Ai} - \bar{x}_A)^2 + \sum(x_{Bi} - \bar{x}_B)^2 + \sum(x_{Ci} - \bar{x}_C)^2$$
$$SS_{Within} = (0 + 1 + 9 + 4) + (0 + 0 + 4 + 4) + (1 + 0 + 1 + 0) = 14 + 8 + 2 = 24$$

Step 5: Verify Total Sum of Squares

$$SS_{Total} = SS_{Between} + SS_{Within} = 8 + 24 = 32$$

This confirms that the total variation is correctly partitioned.

Step 6: Compute Mean Squares

$$MS_{Between} = \frac{SS_{Between}}{df_{Between}} = \frac{8}{3-1} = 4.00$$
$$MS_{Within} = \frac{SS_{Within}}{df_{Within}} = \frac{24}{12-3} = 2.67$$

Step 7: Calculate F-Ratio

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{4.00}{2.67} \approx 1.5$$

Step 8: Compare with F-Critical

From the F-table at 5% significance level with $df_1 = 2$and $df_2 = 9$:

$$F_{critical} = 4.26$$

Since $F_{calculated} = 1.5 < F_{critical} = 4.26$; The differences among the variety means are not statistically significant.

Step 9: Conclusion

The observed differences in per-acre wheat production among the three varieties may have arisen by chance. Therefore, we fail to reject the null hypothesis and conclude that there is no significant difference in wheat yield between the varieties.

ANOVA Table:

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (from the F-table)*
Between sample	8	(3 – 1) = 2	8/2 = 4.00	4.00/2.67 = 1.5	*F* (2, 9) = 4.26
Within sample	24	(12 – 3) = 9	24/9 = 2.67		
Total	32	(12 – 1) = 11			

11.3 TWO-WAY ANOVA

Introduction

In research and data analysis, the outcome, or dependent variable, is often influenced by multiple factors. While one-way ANOVA allows us to examine the effect of a single factor on a dependent variable, two-way ANOVA is used when the data are classified according to two independent factors simultaneously.

Two-way ANOVA not only helps identify whether each factor individually influences the outcome but also enables us to examine the interaction between the two factors—that is, whether the effect of one factor depends on the level of the other. This makes it a powerful tool for analysing complex experimental designs.

Applications of Two-Way ANOVA

1. Agriculture:
    i. Studying crop yield classified by seed variety and fertiliser type.
    ii. Helps determine which combination of seed and fertiliser maximises output.
2. Business and Marketing:
    i. Analysing sales performance classified by salesperson and region.
    ii. Helps identify whether differences in sales are due to individual salespersons, regional markets, or their interaction.
3. Manufacturing and Industry:
    i. Evaluating product quality classified by machine type and worker skill level.
    ii. Helps determine if machine type, worker skill, or both factors together influence production quality.
4. Healthcare and Medicine:
    i. Testing the effect of dosage and treatment method on patient recovery.
    ii. Helps identify whether the outcome is affected by dosage, treatment method, or the combination of both.

Advantages of Two-Way ANOVA

- Allows simultaneous examination of two factors, saving time and effort compared to conducting multiple one-way ANOVAs.
- Provides insight into interaction effects that one-way ANOVA cannot detect.
- Helps in making more informed decisions in experimental and observational studies.

When to Use Two-Way ANOVA

Two-way ANOVA is appropriate when:

i. There are two independent variables (factors).
ii. Each factor has two or more levels (categories).
iii. The dependent variable is continuous.
iv. Observations are independent and approximately normally distributed.
v. Variances across groups are homogeneous (similar).

Steps in Two-Way Anova:

Step 1: Total Sum of Squares (SS Total): Total variation among all observations.

$$SS_{\text{Total}} = \sum x_{ij}^2 - CF$$

where $CF = \frac{T^2}{n}$is the correction factor, T= sum of all observations, n= total number of observations.

Step 2: Sum of Squares Between Columns (SS Columns): Variation due to factor 1.

$$SS_{\text{Columns}} = \sum \frac{T_j^2}{n_j} - CF$$

T_j= total of observations in column j, n_j= number of items in column.

Step 3: Sum of Squares Between Rows (SS Rows): Variation due to factor 2.

$$SS_{\text{Rows}} = \sum \frac{T_i^2}{n_i} - CF$$

T_i= total of observations in row i, n_i= number of items in row.

Step 4: Residual Error Sum of Squares (SS Residual): Variation not explained by the factors.

$$SS_{\text{Residual}} = SS_{\text{Total}} - (SS_{\text{Columns}} + SS_{\text{Rows}})$$

Degrees of freedom:

Source of Variation	Degrees of Freedom
Total	(c. r - 1)
Between Columns	(c - 1)
Between Rows	(r - 1)
Residual/Error	((c-1) (r-1))

Where c= number of columns, r= number of rows.

Step 5: Compute Mean Squares (MS): Mean squares are computed by dividing the sum of squares by their respective degrees of freedom:

$$MS_{\text{Columns}} = \frac{SS_{\text{Columns}}}{c-1}, MS_{\text{Rows}} = \frac{SS_{\text{Rows}}}{r-1}, MS_{\text{Residual}} = \frac{SS_{\text{Residual}}}{(c-1)(r-1)}$$

Step 6: Compute F-Ratios

F-ratios test whether the variance explained by each factor is significantly larger than the residual variance:

$$F_{\text{Columns}} = \frac{MS_{\text{Columns}}}{MS_{\text{Residual}}}, F_{\text{Rows}} = \frac{MS_{\text{Rows}}}{MS_{\text{Residual}}}$$

Compare the calculated F-values with F-critical values from the F-table at the chosen significance level (e.g., 5%).

- If $F_{\text{calculated}} \geq F_{\text{critical}}$, the factor has a significant effect**.**
- If $F_{\text{calculated}} < F_{\text{critical}}$, the factor effect is not significant**.**

Interaction Effect

- If the design includes repeated measurements, the interaction effect between the two factors can also be analyzed.
- Interaction indicates that the effect of one factor depends on the level of the other factor**.**
- Interaction SS can be calculated as:

$$SS_{\text{Interaction}} = SS_{\text{Total}} - (SS_{\text{Columns}} + SS_{\text{Rows}} + SS_{\text{Residual}})$$

- F-ratio for interaction is computed by dividing MS Interaction by MS Residual**.**

ANOVA table can be set up as shown below:

Source of variation	*Sum of squares (SS)*	*Degrees of freedom (d.f.)*	*Mean square (MS)*	*F-ratio*
Between columns treatment	$\sum \frac{(T_j)^2}{n_j} - \frac{(T)^2}{n}$	$(c-1)$	$\frac{SS \text{ between columns}}{(c-1)}$	$\frac{MS \text{ between columns}}{MS \text{ residual}}$
Between rows treatment	$\sum \frac{(T_i)^2}{n_i} - \frac{(T)^2}{n}$	$(r-1)$	$\frac{SS \text{ between rows}}{(r-1)}$	$\frac{MS \text{ between rows}}{MS \text{ residual}}$

Residual or error	Total SS – (SS between columns + SS between rows)	$(c-1)(r-1)$	$\frac{SS \text{ residual}}{(c-1)(r-1)}$	
Total	$\sum X_{ij}^2 - \frac{(T)^2}{n}$	$(c.r-1)$		

Key Points for Interpretation

1. MS Residual captures natural variability (sampling fluctuations).
2. F-Ratios test whether factor differences are statistically significant.
3. Higher F-values indicate a more substantial effect of the factor.
4. Two-way ANOVA allows simultaneous testing of two factors and residual error, providing a complete view of variability in the data.

Illustration 2

A researcher aims to examine the impact of various seed varieties and fertiliser types on crop yield. The experiment involves three seed varieties (A, B, C) and four fertiliser types (W, X, Y, Z). The recorded yield (in kg) for each combination is provided as follows.

Varieties of seeds	*A*	*B*	*C*
Varieties of fertilisers			
W	6	5	5
X	7	5	4
Y	3	3	3
Z	8	7	4

Apply Two-Way ANOVA to examine:

1. Whether the type of fertiliser significantly affects crop yield.
2. Whether the variety of seed significantly affects crop yield.
3. Whether there is a significant interaction effect between fertiliser type and seed variety on crop yield.

Solution:

Factor 1: Fertiliser

- Null hypothesis H_{0F} There is no significant difference in crop yield among the different fertilisers.
- Alternative hypothesis H_{1F} There is a significant difference in crop yield among the different fertilisers.

Factor 2: Seed Variety

- Null hypothesis H_{0S}: There is no significant difference in crop yield among the different seed varieties.
- Alternative hypothesis H_{1S}: There is a significant difference in crop yield among the different seed varieties.

Step (i)	$T = 60$, $n = 12$, □ Correction factor $= \frac{(T)^2}{n} = \frac{60\ X\ 60}{12} = 300$
Step (ii)	Total $SS = (36 + 25 + 25 + 49 + 25 + 16 + 9 + 9 + 9 + 64 + 49 + 16) - \left(\frac{60\ X\ 60}{12}\right)$ $= 332\text{-}300$ $= 32$
Step (iii)	SS between columns treatment □ $[\ \frac{24\ X\ 24}{4} + \frac{20\ X\ 20}{4} + \frac{16\ X\ 16}{4}\ ] - [\frac{60\ X\ 60}{12}\ ]$ $= 144 + 100 + 64 - 300$ $= 8$
Step (iv)	SS between rows treatment □ $[\ \frac{16\ X\ 16}{3} + \frac{16\ X\ 16}{3} + \frac{9\ X\ 9}{3} + \frac{19\ X\ 19}{3}] - [\frac{60\ X\ 60}{12}\ ]$ $= 85.33+85.33+27.00+120.33-300$ $= 18$
Step (v)	SS residual or error = Total SS – (SS between columns + SS between rows) $= 32 - (8 + 18)$ $= 6$

ANOVA Table

Source of variation	*SS*	*d.f.*	*MS*	*F-ratio*	*5% F-limit (or the table values)*
Between columns (i.e., between varieties of seeds)	8	(3 – 1) = 2	8/2 = 4	4/1 = 4	$F(2,6) = 5.14$
Between rows (i.e., between varieties of fertilizers)	18	(4 – 1) = 3	18/3 = 6	6/1 = 6	$F(3,6) = 4.76$
Residual or error	6	(3 – 1) × (4 – 1) = 6	6/6=1		
Total	32	(3 × 4) – 1 = 11			

From the said ANOVA table, we find that differences concerning varieties of seeds are insignificant at 5% level as the calculated F-ratio of 4 is less than the table value of 5.14, but the variety differences concerning fertilisers are significant as the calculated F-ratio of 6 is more than its table value of 4.76. Therefore, it can be concluded that crop yield is significantly influenced by the choice of fertiliser, whereas the seed variety does not have a significant impact.

11.4 MULTIPLE CORRELATION

When one dependent variable is influenced by two or more independent variables, the strength of their combined linear relationship is measured using Multiple Correlation.

- If X_1 depends on X_2 and X_3, the multiple correlation coefficient is $R_{1.23}$.
- If X_2 depends on X_1 and X_3, it is $R_{2.13}$.
- If X_3 depends on X_1 and X_2, it is $R_{3.12}$.

Multiple correlation represents how well the dependent variable is predicted by the combined effect of the other two variables.

Formulas for Multiple Correlation

1. Multiple Correlation of X_1 on X_2 and X_3

$$R_{1.23}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}$$

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}$$

2. Multiple Correlation of X_2 on X_1 and X_3

$$R^2_{2.13} = \frac{r^2_{12} + r^2_{23} - 2r_{12}r_{23}r_{13}}{1 - r^2_{13}}$$

$$R_{2.13} = \sqrt{\frac{r^2_{12} + r^2_{23} - 2r_{12}r_{23}r_{13}}{1 - r^2_{13}}}$$

3. Multiple Correlation of X_3 on X_1 and X_2

$$R^2_{3.12} = \frac{r^2_{13} + r^2_{23} - 2r_{13}r_{23}r_{12}}{1 - r^2_{12}}$$

$$R_{3.12} = \sqrt{\frac{r^2_{13} + r^2_{23} - 2r_{13}r_{23}r_{12}}{1 - r^2_{12}}}$$

Illustration 3

A simple correlation coefficient between yield(x_1), temperature(x_2) and rainfall(x_3) are given by r_{12}=0.6, r_{13}=0.5 and r_{23}=0.8. Find the multiple correlation $R_{1.23}$, $R_{2.13}$ and $R_{3.12}$

Solution:

Multiple Correlation $R_{1.23}$

Given r_{12}=0.6, r_{13}=0.5, r_{23}=0.8

$$R_{1.23} = \sqrt{\frac{r^2_{12} + r^2_{13} - 2r_{12}r_{13}r_{23}}{1 - r^2_{23}}}$$

$$R_{1.23} = \sqrt{\frac{(0.6)^2 + (0.5)^2 - 2(0.6)(0.5)(0.8)}{1 - (0.8)^2}}$$

$$R_{1.23} = \sqrt{\frac{0.36 + 0.25 - 0.48}{1 - 0.64}}$$

$$R_{1.23} = \sqrt{\frac{0.13}{0.36}}$$

$$R_{1.23} = \sqrt{0.3611}$$

$$\boxed{R_{1.23} \approx 0.60}$$

Multiple Correlation $R_{2.13}$

$$R_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{23}r_{13}}{1 - r_{13}^2}}$$

$$R_{2.13} = \sqrt{\frac{(0.6)^2 + (0.8)^2 - 2(0.6)(0.8)(0.5)}{1 - (0.5)^2}}$$

$$R_{2.13} = \sqrt{\frac{0.36 + 0.64 - 0.48}{1 - 0.25}}$$

$$R_{2.13} = \sqrt{\frac{0.52}{0.75}}$$

$$R_{2.13} = \sqrt{0.6933} = 0.833$$

$$\boxed{R_{2.13} \approx 0.833}$$

Multiple Correlation $R_{3.12}$

$$R_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{13}r_{23}r_{12}}{1 - r_{12}^2}}$$

$$R_{3.12} = \sqrt{\frac{(0.5)^2 + (0.8)^2 - 2(0.5)(0.8)(0.6)}{1 - (0.6)^2}}$$

$$R_{3.12} = \sqrt{\frac{0.25 + 0.64 - 0.48}{1 - 0.36}}$$

$$R_{3.12} = \sqrt{\frac{0.41}{0.64}}$$

$$R_{3.12} = \sqrt{0.6406} = 0.800$$

$$\boxed{R_{3.12} \approx 0.800}$$

Interpretation

Proportion of variance explained

- $R_{1.23}^2 = 0.361111 \rightarrow$ 36.11% of variance in yield (X_1) is explained jointly by temperature (X_2) and rainfall (X_3).
- $R_{2.13}^2 = 0.693333 \rightarrow$ 69.33% of variance in temperature (X_2) is explained jointly by yield (X_1) and rainfall (X_3).

- $R^2_{3.12} = 0.640625 \rightarrow$ 64.06% of variance in rainfall (X_3) is explained jointly by yield (X_1) and temperature (X_2).
- $R_{1.23} \approx 0.601$ (moderate positive): Yield has a moderate positive combined linear relationship with temperature and rainfall — about 36% of yield variability is explained by those two predictors together. Other factors still explain the remaining ~64%.
- $R_{2.13} \approx 0.833$ (strong positive): Temperature is strongly predictable from yield and rainfall together — nearly 69% of its variance is explained by X_1 and X_3.
- $R_{3.12} \approx 0.800$ (strong positive): Rainfall is also strongly predictable from yield and temperature together — about 64% of its variance is explained by X_1 and X_2.

11.5 MULTIPLE REGRESSION

Multiple regression is a statistical technique used to examine the relationship between one dependent variable and two or more independent variables. It helps in understanding how each predictor contributes to changes in the outcome variable. By considering multiple factors simultaneously, it provides a more accurate and reliable prediction than simple regression. Multiple regression also identifies the relative importance of each variable in explaining the variation in the dependent variable. It is widely used in business, economics, social sciences, and research for forecasting and decision-making.

Where:

- X_1= Dependent variable
- $a_{1.23}$= Intercept when $X_1$is regressed on $X_2$and X_3
- $b_{12.3}$= Partial regression coefficient of X_2on X_1keeping X_3constant
- $b_{13.2}$= Partial regression coefficient of X_3on X_1keeping X_2constant

This is the standard form of a multiple regression equation with X_1as the dependent variable.

Formula for $b_{12.3}$

Partial regression coefficient of X_2 on X_1 keeping X_3 constant:

$$b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \cdot \frac{\sigma_1}{\sigma_2}$$

Formula for $b_{13.2}$

Partial regression coefficient of X_3 on X_1 keeping X_2 constant:

$$b_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{\sigma_1}{\sigma_3}$$

Where:

- r_{12}, r_{13}, r_{23}= simple correlation coefficients

- $\sigma_1, \sigma_2, \sigma_3$= standard deviations of X_1, X_2, X_3

Formula for the Intercept $a_{1.23}$

$$a_{1.23} = \bar{X}_1 - b_{12.3}\bar{X}_2 - b_{13.2}\bar{X}_3$$

Where:

- $\bar{X}_1, \bar{X}_2, \bar{X}_3$= means of the variables

Illustration 4

r_{12}=0.8, r_{13}=0.7, r_{23}=0.6, $\sigma_1 = 10\ \sigma_2 = 8, \sigma_3 = 5$ .Determine the regression equation of X_1 on X_2 and X_3

Solution:

Regression equation of X_1 on X_2 and X_3 is

X1= $b_{12.3}$ X2 + $b_{13.2}$ X3

$$b_{12.3} = \frac{r_{12} - r_{13}r_{23}}{1 - r_{23}^2} \cdot \frac{\sigma_1}{\sigma_2}, \qquad b_{13.2} = \frac{r_{13} - r_{12}r_{23}}{1 - r_{23}^2} \cdot \frac{\sigma_1}{\sigma_3}$$

$$b_{12.3} = \frac{0.8 - (0.7 \cdot 0.6)}{1 - 0.6^2} \cdot \frac{10}{8}, \qquad b_{13.2} = \frac{0.7 - (0.8 \cdot 0.6)}{1 - 0.6^2} \cdot \frac{10}{5}$$

$$b_{12.3} = 0.7375, \qquad b_{13.2} = 0.68$$

Regression equation is

$$X_1 = b_{12.3}X_2 + b_{13.2}X_3$$

Substitute $b_{12.3}$and $b_{13.2}$from above:

$$X_1 = 0.7375\,\bar{X}_2 + 0.68\,\bar{X}_3$$

Interpretation:

The regression equation $X_1 = 0.7375X_2 + 0.68X_3$shows that both $X_2$and X_3have a positive effect on X_1. A unit increase in X_2increases $X_1$by 0.7375 units, while a unit increase in X_3increases $X_1$by 0.68 units. This indicates that X_2has a slightly stronger influence on X_1than X_3.

11.6 SUMMARY:

ANOVA is a statistical technique used to compare the means of two or more groups to determine if they differ significantly. It partitions total variation into between-group and within-group variation and uses the F-test to assess significance. It is widely used in experiments and research to test hypotheses about group differences.

Multiple correlation measures the strength of the relationship between one dependent variable and two or more independent variables. The multiple correlation coefficient (R) ranges from 0 to 1, indicating how well the independent variables together predict the dependent variable. It helps in understanding combined effects of predictors.

Multiple regression is a technique to model the relationship between one dependent variable and two or more independent variables. It provides a regression equation:

$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_kX_k$$

This equation predicts the dependent variable and shows the relative contribution of each predictor. It is widely used in forecasting, decision-making, and research analysis.

11.7 TECHNICAL TERMS

- **ANOVA (Analysis of Variance):** A statistical technique to compare means of multiple groups.
- **Factor:** An independent variable in ANOVA.
- **Level:** The different categories or groups within a factor.
- **Sum of Squares (SS):** Measure of variation in the data.
- **Mean Square (MS):** SS divided by its degrees of freedom.
- **F-Ratio:** Ratio of variance between groups to variance within groups; used to test hypotheses.
- **Multiple Correlation Coefficient (R):** Measures the strength of relationship between one dependent variable and multiple independent variables.
- **Coefficient of Determination (R^2):** Proportion of variance in the dependent variable explained by independent variables.
- **Partial Regression Coefficient:** Shows the effect of an independent variable on the dependent variable keeping other variables constant.
- **Intercept (a):** The expected value of the dependent variable when all independent variables are zero.

11.8 SELF-ASSESSMENT QUESTIONS

1. Explain the meaning of analysis of variance. Describe briefly the technique of analysis of variance for one-way and two-way classifications.
2. Below are given the yields per acre of wheat for six plots entering a crop competition, three of the plots being sown with wheat of variety *A* and three with variety *B*.

Variety	*Yields in pe acre fields*		
	1	*2*	*3*
A	30	32	22
B	20	18	16

Set up a table of analysis of variance and calculate *F*. State whether the difference between the yields of two varieties is significant, taking 7.71 as the table value of *F* at 5% level for $v_1 = 1$ and $v_2 = 4$.

3. Differentiate between one-way and two-way ANOVA.
4. List the main assumptions of ANOVA.
5. Explain the difference between main effects and interaction effects in two-way ANOVA.
6. How do you interpret an F-ratio in ANOVA?
7. The simple correlation coefficients between yield (X_1), temperature (X_2), and rainfall (X_3) are $r_{12} = 0.6$, $r_{13} = 0.5$, and $r_{23} = 0.8$. The standard deviations are $\sigma_1 = 12$, $\sigma_2 = 8$, and $\sigma_3 = 6$.
   i. Determine the multiple correlation coefficient ($R_{1.23}$) between yield and the two independent variables.
   ii. Construct the regression equation of X_1 on X_2 and X_3.

11.9 SUGGESTED READINGS

1. Kothari, C. R. (2022). *Research Methodology: Methods and Techniques* (4th ed.). New Delhi: New Age International Publishers.
2. Kumar, R. (2019). *Research Methodology: A Step-by-Step Guide for Beginners* (5th ed.). London: SAGE Publications.
3. Monette, D. R., Sullivan, T. J., & DeJong, C. R. (2014). *Applied Social Research: A Tool for the Human Services* (9th ed.). Boston: Cengage Learning.
4. Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* (3rd ed.). New York: Routledge.
5. Fletcher, C. (2012). *Analysis of Variance, Design, and Regression*. New Delhi: SAGE Publications.

Dr. G. MALATHI

LESSON-12

DISCRIMINANT ANALYSIS

OBJECTIVES OF THE LESSON

After studying this lesson, students will be able to:

1. Understand the concept and purpose of Discriminant Analysis in research.
2. Differentiate between types of Discriminant Analysis and their applications.
3. Learn the process and assumptions involved in conducting Discriminant Analysis.
4. Interpret the results of Discriminant Analysis and assess their research implications.
5. Apply Discriminant Analysis in real-world managerial and social science research contexts.

STRUCTURE OF THE LESSON

12.1 INTRODUCTION TO DISCRIMINANT ANALYSIS

Discriminant Analysis is a robust multivariate statistical technique used to determine which variables separate two or more naturally occurring groups. In research methodology, it serves both descriptive and predictive purposes — helping researchers understand the reasons and mechanisms behind group differences, as well as classifying new observations into those predefined groups with a measurable level of accuracy.

The main idea behind Discriminant Analysis is to develop a discriminant function, which is a linear combination of independent variables that provides the most significant separation between the groups. The function has the general mathematical form:

$$D = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Where D represents the discriminant score, a is a constant, $b_1, b_2, \ldots, b_n$ are the discriminant coefficients, and $X_1, X_2, \ldots, X_n$ are the independent (predictor) variables. The coefficients are estimated so that the ratio of between-group to within-group variance is maximised, ensuring optimal discrimination.

Discriminant Analysis is useful when the dependent variable is categorical (nominal) — for example, customer satisfaction level (high/low), student performance (pass/fail), or creditworthiness (good/insufficient) — and the independent variables are continuous (metric), such as age, income, test scores, or financial ratios.

Unlike regression analysis, which predicts a continuous outcome, discriminant analysis predicts group membership. Conceptually, it is similar to logistic regression but relies on certain statistical assumptions such as multivariate normality and homogeneity of covariance matrices. When these assumptions are satisfied, Discriminant Analysis tends to deliver more efficient and interpretable results.

Historical Context and Development

Discriminant Analysis was first introduced by the statistician R. A. Fisher in 1936 in his influential paper on classifying iris flowers based on their morphological measurements—a study that remains a classic example in statistical literature. Fisher's Linear Discriminant Function laid the foundation for modern classification techniques and has since evolved to handle multiple groups, non-linear boundaries, and extensive data structures.

Discriminant Analysis addresses questions such as:

- Which variables best distinguish one group from another?
- How accurately can cases be classified into their true groups?
- How distinct are the groups based on the observed characteristics?

It operates on the principle that, given a set of predictor variables, individuals within the same group share similar characteristics, while members of different groups differ significantly. The discriminant function thus defines an axis in multidimensional space along which the separation between group centroids (mean vectors) is maximised.

Practical Relevance in Research

In applied research, Discriminant Analysis is extensively used for classification and profiling purposes, for instance:

- In marketing research, it helps classify consumers into market segments or predict brand loyalty.
- In finance, it is used in credit scoring and bankruptcy prediction models.
- In human resource management, it assists in distinguishing successful from unsuccessful job candidates.
- In social and behavioural sciences, it helps identify factors that differentiate groups such as urban versus rural populations or high versus low achievers.

Because of its interpretive strength, Discriminant Analysis not only predicts category membership but also offers insights into which predictor variables are most influential in

differentiating groups—making it a valuable tool for both decision-making and theoretical understanding.

Importance of Discriminant Analysis in Research Methodology

From a methodological perspective, Discriminant Analysis is a vital technique in a researcher's toolkit for addressing problems involving group differentiation and classification. It complements other multivariate methods, such as factor analysis, cluster analysis, and logistic regression, acting as a link between data reduction and predictive modelling. It also aids in hypothesis testing for group differences and serves as the basis for modern machine learning algorithms used in pattern recognition and classification.

Discriminant Analysis allows researchers to go beyond basic group comparisons by creating statistically sound functions that measure the separation between groups, classify new cases, and elucidate the underlying dimensions of group differentiation.

12.2 TYPES OF DISCRIMINANT ANALYSIS

Discriminant Analysis can be categorised into various types based on the number of groups involved and the nature of covariance structures across those groups. The three most common types are Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Multiple Discriminant Analysis (MDA). Each one has a specific analytical purpose and is used under particular statistical conditions.

a. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is the most widely used form of Discriminant Analysis. It assumes that:

1. The data for each group follow a multivariate normal distribution, and
2. The variance–covariance matrices of all groups are equal.

Under these conditions, the best discriminating function between groups is linear, meaning that the decision boundary separating the groups can be expressed as a straight line (in two dimensions) or a hyperplane (in multiple dimensions).

The general form of the linear discriminant function is:

$$D = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

Here, the coefficients (b_1, b_2, …, b_n) are estimated to maximise the ratio of between-group variance to within-group variance—thus ensuring the most significant separation between group centroids.

Purpose: LDA aims to identify a linear combination of predictor variables that best distinguishes between two or more naturally occurring groups. It not only classifies cases but also helps researchers understand which variables most contribute to group differentiation.

Applications

1. Human Resource Management – Distinguishing between "high performers" and "low performers" using predictors such as test scores, experience, and job satisfaction levels.
2. Marketing Research – Predicting whether a customer is a "brand loyalist" or a "brand switcher" based on product usage, income, and satisfaction scores.
3. Medical Research – Differentiating between patients with and without a particular disease based on clinical indicators such as blood pressure, cholesterol, and BMI.
4. Finance – Classifying firms as solvent or insolvent using financial ratios and cash flow metrics.

b. Quadratic Discriminant Analysis (QDA)

Quadratic Discriminant Analysis (QDA) is a classification technique that models each class with a separate, unique covariance matrix. Unlike Linear Discriminant Analysis (LDA), which assumes all classes share the same covariance matrix, QDA permits more flexible, curved decision boundaries by not making this assumption. This flexibility is advantageous when classes exhibit different variances, but it also increases the number of parameters to estimate, making it less suitable for high-dimensional data. The discriminant function for a given class

The discriminant function for a given class k is

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + \log(\pi_k)$$

Where:

- $x =$ The input vector of features.
- μ_k = The mean vector for class k
- $\Sigma_k =$ The covariance matrix for the class k
- $|\Sigma_k| =$ The determinant of the covariance matrix Σ_k
- $\Sigma_k^{-1} =$ The inverse of the covariance matrix
- $log\ \pi_k =$ The log-prior probability of class k

How it works

- **Calculate for each class:** For each class k, compute the mean vector $\mu_k$and the covariance matrix Σ_k
- **Apply discriminant function:** For a new data point x, calculate the discriminant function $\delta_k(x)$ for every class k.

- **Classify:** Assign the data point x, to the class k for which the discriminant function $\delta_k(x)$ is the largest.
- **Decision boundary:** The decision boundary between two classes is quadratic because the terms do not cancel out when the covariance matrices differ. This leads to a curved boundary rather than the linear one observed in LDA. This results in a curved boundary instead of the linear one seen in LDA.

Purpose: QDA is particularly useful when the data exhibit heterogeneity in variance-covariance structures, making it more appropriate than LDA for complex datasets. It allows non-linear separation among groups, improving predictive performance at the cost of increased model complexity.

Applications:

1. Marketing Segmentation – Classifying customers into different market segments (e.g., value buyers, quality seekers, brand aspirants) when spending patterns and variances differ widely among segments.
2. Credit Risk Analysis – Differentiating between low-, medium-, and high-risk borrowers when financial variability differs significantly across risk groups.
3. Biological Sciences – Classifying species or genotypes when the variability in physiological characteristics is not uniform across categories.
4. Industrial Quality Control – Distinguishing among production batches with differing variability in product measurements.

c. Multiple Discriminant Analysis (MDA)

Multiple Discriminant Analysis extends Linear Discriminant Analysis to cases with more than two groups. It generates a series of discriminant functions (up to g – 1, where g is the number of groups), each representing an independent dimension along which the groups vary.

The first discriminant function accounts for the maximum possible variance between groups; the second function explains the next highest amount of remaining variance, and so forth. Each subsequent function is orthogonal (independent) to the previous one.

Mathematical Representation

For *g* groups and *p* independent variables, the discriminant functions are given as:

$$D_1 = a_1 + b_{11}X_1 + b_{12}X_2 + \ldots + b_{1p}X_p$$
$$D_2 = a_2 + b_{21}X_1 + b_{22}X_2 + \ldots + b_{2p}X_p$$
$$\ldots D_{g-1} = a_{g-1} + b_{(g-1)1}X_1 + \ldots + b_{(g-1)p}X_p$$

Each function maximises separation among group centroids while maintaining orthogonality with previously extracted functions.

Purpose: MDA helps researchers visualise and interpret group differences in multidimensional space and is useful for problems involving more than two categories. It not only classifies cases but also highlights which dimensions (functions) best explain group separations.

Applications

1. Education Research – Classifying students into *high*, *moderate*, and *low* academic achievers using predictors such as attendance, internal marks, and study hours.
2. Marketing and Consumer Behaviour – Segmenting customers into *premium*, *value-conscious*, and *discount-seeking* groups based on demographic and psychographic variables.
3. Finance – Categorising firms as *high-growth*, *stable*, or *declining* based on profitability ratios and market indicators.
4. Healthcare and Psychology – Grouping patients into categories of *mild*, *moderate*, or *severe* conditions based on diagnostic scores.

Comparison

Feature	***Linear Discriminant Analysis (LDA)***	***Quadratic Discriminant Analysis (QDA)***	***Multiple Discriminant Analysis (MDA)***
Number of Groups	Two or more	Two or more	More than two
Covariance Matrices	Equal across groups	Unequal across groups	Equal (typically assumed)
Type of Boundary	Linear	Quadratic (curved)	Multiple linear boundaries
Primary Purpose	Classify and explain group differences	Classify with flexible group variability	Classify and interpret multidimensional group differences
Complexity	Moderate	High	Moderate to high
Example Application	Employee performance classification	Market segmentation with unequal variances	Student achievement categorisation

12.3 APPLICATIONS OF DISCRIMINANT ANALYSIS

Discriminant Analysis has wide applications across disciplines:

Field	**Application**
Marketing Research	Classifying consumers into loyal vs. switcher groups predicting brand choice.
Finance and Banking	Credit scoring; distinguishing between solvent and insolven firms.
Human Resource Managemen	Classifying job applicants as suitable or unsuitable based o test scores.

Education Research	Predicting student success or dropout likelihood.
Healthcare	Differentiating between patients with and without a medica condition.

It serves both predictive and descriptive purposes—helping to understand which variables discriminate among groups and how accurately new cases can be classified.

12.4 PROCESS OF CONDUCTING DISCRIMINANT ANALYSIS

The process of conducting Discriminant Analysis is systematic and involves a sequence of well-defined steps that ensure statistical validity and meaningful interpretation. Each stage plays a crucial role in deriving an accurate discriminant function that distinguishes among predefined groups.

The significant steps are described below:

Step 1: Define the Research Problem

The first and most critical step is to clearly define the research objective and identify the grouping variable (dependent variable) and the predictor variables (independent variables).

- The grouping variable must be categorical (nominal) in nature, representing distinct groups such as "high vs. low performers," "pass vs. fail," or "loyal vs. non-loyal customers."
- The predictor variables must be metric (interval or ratio), such as age, income, test score, satisfaction index, or sales turnover.

At this stage, researchers must also decide whether the analysis is descriptive (to understand group differences) or predictive (to classify new cases).

Example:
A researcher may wish to classify bank customers as *creditworthy* or *non-creditworthy* based on predictors such as income, debt-to-income ratio, and credit score.

Step 2: Test Assumptions

Before performing Discriminant Analysis, it is essential to verify that the data meet certain statistical assumptions that ensure the validity of the results.

1. Multivariate Normality: The predictor variables should follow a multivariate normal distribution within each group. This can be tested using statistical tests such as the Shapiro–Wilk or visualised using Q–Q plots.
2. Homogeneity of Variance–Covariance Matrices: The assumption of equal covariance matrices across groups must hold true. This is tested using Box's M Test. A non-significant result ($p > 0.05$) suggests that the assumption is not violated.

3. Absence of Multicollinearity: Predictor variables should not be highly correlated with one another. High multicollinearity can distort the estimation of discriminant coefficients. Variance Inflation Factor (VIF) values less than 10 generally indicate acceptable levels.
4. Independence of Observations: The observations (cases) should be independent of each other; that is, each respondent or case should belong to one group only.

If the data do not satisfy these assumptions, alternative methods such as logistic regression or nonparametric classification techniques may be considered.

Step 3: Estimate the Discriminant Function

Once assumptions are verified, the next step is to estimate the discriminant function(s). This involves determining the coefficients (weights) of each independent variable that best separate the groups.

The general form of the discriminant function is:

$$D = a + b_1X_1 + b_2X_2 + \ldots + b_nX_n$$

where

- D = Discriminant score,
- a = Constant term,
- $b_1, b_2, \ldots, b_n$ = Discriminant coefficients,
- $X_1, X_2, \ldots, X_n$ = Independent variables.

Statistical software such as SPSS, R, SAS, or Python (scikit-learn) is typically used to compute these coefficients by maximising the ratio of between-group to within-group variance.

If there are k groups, up to $k - 1$ discriminant functions can be generated, each representing an independent dimension of discrimination.

Step 4: Evaluate the Significance of the Function

After estimating the discriminant function(s), the next step is to test whether these functions significantly differentiate among groups.

1. **Wilks' Lambda (Λ):** This statistic measures the proportion of total variance in discriminant scores not explained by group differences. Smaller values indicate better discrimination. A significant Chi-square value for Wilks' Lambda indicates that the discriminant function effectively separates the groups.
2. **Canonical Correlation:** This represents the correlation between the discriminant scores and the grouping variable. A high canonical correlation indicates strong discriminating power.
3. **Eigenvalues:** These reflect the amount of variance explained by each discriminant function. The higher the eigenvalue, the greater the function's discriminating ability.

Example:
In a study that classifies customers into loyal, switchers, and price-sensitive groups, a significant Wilks' Lambda ($p < 0.001$) and a canonical correlation of 0.78 indicate that the discriminant function effectively separates these customer segments.

Step 5: Interpret the Function

Once the discriminant functions are found to be statistically significant, they must be **interpreted** to understand which variables contribute most to the differentiation among groups.

1. **Standardised Discriminant Coefficients:** These coefficients indicate the relative importance of each predictor variable in the discriminant function (similar to beta weights in regression analysis).
2. **Structure Matrix (Pooled Within-Groups Correlation):** This shows the correlations between each predictor and the discriminant function, helping identify which variables are most closely associated with group separation.
3. **Group Centroids:** The mean discriminant score for each group (centroid) represents the group's position along the discriminant function. The greater the distance between centroids, the better the discrimination.

Example:
If income and debt ratio have high loadings on the first discriminant function, it suggests these are the primary factors distinguishing *creditworthy* from *non-creditworthy* customers.

Step 6: Validate the Model

Validation ensures that the discriminant function performs consistently when applied to new or unseen data.

1. **Classification Matrix (Confusion Matrix):** This table compares predicted group memberships with actual group memberships. The **hit ratio** (percentage of correctly classified cases) measures predictive accuracy.
2. **Cross-Validation (Leave-One-Out Method):** Each case is classified using a function derived from all other cases except that one. This helps assess the stability of the discriminant function.
3. **Holdout Sample Validation:** The sample is divided into two subsets: one for estimating the function (training sample) and the other for testing its predictive accuracy (validation sample).

A classification accuracy significantly higher than the maximum chance criterion (often 25% or 33%, depending on the number of groups) indicates a strong model.

Step 7: Report the Results

Finally, researchers must present the results in a clear, concise, and interpretable format that supports decision-making or theory building. The report should include:

- The discriminant function(s) with coefficients.
- Wilks' Lambda, Chi-square, eigenvalues, and canonical correlations.
- Group centroids and classification results (hit ratio).
- Interpretation of key variables contributing to group separation.
- Validation results showing the predictive reliability of the model.

Example of Reporting: "The discriminant function was statistically significant (Wilks' Λ = 0.24, χ^2 = 42.15, $p < 0.001$), explaining 72% of between-group variability. The income and expenditure ratios were the strongest predictors. The model correctly classified 88% of cases in the training sample and 84% in the validation sample."

Illustrative Example (Applied Context)

A marketing researcher aims to classify customers into three loyalty groups: *high*, *moderate*, and *low*. Predictors include purchase frequency, customer satisfaction, and average spending.

After testing assumptions and estimating the discriminant functions, the analysis yields two significant functions that explain 78% of the between-group variance. The first function primarily differentiates *high-loyalty customers from others, while the second distinguishes moderate from low-loyalty* customers. The model achieves an overall classification accuracy of 85%, indicating a robust discriminant solution.

12.5 INTERPRETATION OF RESULTS

Key statistics used in interpreting discriminant results include:

- **Canonical Correlation (R^2):** Indicates how well the function discriminates between groups.
- **Wilks' Lambda (Λ):** Measures unexplained variance; lower values indicate better discrimination.
- **Standardised Coefficients:** Show the relative contribution of each variable.
- **Group Centroids:** Mean discriminant scores for each group; used for classification.
- **Classification Matrix (Hit Ratio):** Compares predicted vs. actual group memberships.

12.6 SUMMARY

Discriminant Analysis is a valuable statistical tool for **classifying cases** into predefined categories and **understanding group differences**. It combines elements of regression and multivariate analysis to create discriminant functions that maximise separation among groups.

By carefully following its assumptions and process, researchers can employ Discriminant Analysis to support decision-making across marketing, finance, education, and the social sciences.

12.7 TECHNICAL TERMS

Discriminant unction	: A linear combination of predictors used to distinguish between groups.
Canonical correlation	: Correlation between discriminant scores and groups.
Wilks' Lambda (Λ):	A statistic indicating the significance of the discriminant function.
Group Centroid:	The mean value of discriminant scores for each group.
Hit Ratio:	The percentage of correctly classified cases.
Box's M Test:	A test for equality of covariance matrices.

12.8 SELF-ASSESSMENT QUESTIONS

1. Define Discriminant Analysis and its purpose in research.
2. Differentiate between Linear and Quadratic Discriminant Analysis.
3. State three major assumptions of Discriminant Analysis.
4. What is the role of Wilks' Lambda in Discriminant Analysis?
5. Explain the concept of canonical correlation.
6. Describe the steps involved in conducting a Discriminant Analysis.
7. How can Discriminant Analysis help in predicting consumer behaviour?
8. Discuss the importance of validation in Discriminant Analysis.
9. Compare Discriminant Analysis and Logistic Regression in terms of assumptions and usage.
10. Interpret the following result: Wilks' Lambda = 0.25, $p < 0.001$, Hit Ratio = 89%.

12.9 SUGGESTED READINGS

1. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate Data Analysis* (8th ed.). Pearson Education.
2. Tabachnick, B. G., & Fidell, L. S. (2019). *Using Multivariate Statistics* (7th ed.). Pearson.
3. Malhotra, N. K., & Dash, S. (2020). *Marketing Research: An Applied Orientation* (8th ed.). Pearson India.
4. Field, A. (2020). *Discovering Statistics Using SPSS* (5th ed.). Sage Publications.
5. Kothari, C. R., & Garg, G. (2019). *Research Methodology: Methods and Techniques* (4th ed.). New Age International.

Dr. G. MALATHI

LESSON-13

FACTOR ANALYSIS & CONJOINT ANALYSIS

OBJECTIVES OF THE LESSON

After studying this lesson, the students will be able to:

1. Understand the meaning and purpose of Factor Analysis in research.
2. Explain the concepts of factors, factor loadings, communalities, and eigenvalues.
3. Describe the process and steps involved in conducting Factor Analysis.
4. Interpret the results of Factor Analysis for decision-making in research and business contexts.
5. Understand the concept, purpose, and theoretical foundation of Conjoint Analysis in marketing research.
6. Explain the steps involved in conducting Conjoint Analysis and interpret its applications for managerial decision-making.

STRUCTURE OF THE LESSON

13.1 INTRODUCTION TO FACTOR ANALYSIS

In many research contexts, especially in social sciences, management, psychology, marketing, and behavioural studies, researchers work with large sets of variables. These variables may not always operate independently; instead, some are closely linked. When multiple variables exhibit strong correlations, interpreting the data becomes difficult. An effective method is required to simplify the data without losing crucial information. Factor Analysis accomplishes this by condensing a large number of correlated variables into a smaller set of meaningful and interpretable underlying dimensions called factors.

Factor Analysis is a multivariate statistical technique used to identify groups of variables with strong interconnections. Each group is represented by a single factor that summarises the

common traits of the variables within it. For example, when analysing customer perceptions of a restaurant, variables such as taste, flavour, freshness, and aroma may be closely linked and collectively indicate an underlying factor, like "Food Quality." In this way, Factor Analysis reduces complexity and improves clarity within a dataset.

The heart of Factor Analysis lies in recognising these latent (hidden) factors that cannot be directly measured but influence multiple observable variables. The method assumes that observable variables are partly affected by common underlying factors and partly by unique factors specific to each variable. By analysing the correlations among variables, Factor Analysis determines how strongly each variable relates to each factor. These relationships are represented through factor loadings, which assist the researcher in naming and interpreting the factors.

Factor Analysis is particularly useful when researchers aim to identify patterns and structure within data. It helps them pinpoint the key dimensions that affect consumer attitudes, employee satisfaction, personality traits, product preferences, or perceptions of service quality. The results of Factor Analysis aid in data reduction, scale development, market segmentation, and strategic planning. For instance, rather than analysing ten individual product attributes, a marketing manager might focus on two main dimensions, such as "Value for Money" and "Brand Experience," simplifying decision-making.

Therefore, Factor Analysis plays a crucial role in both academic research and practical business applications. It simplifies complex datasets, uncovers significant structures, and enhances the clarity of research outcomes. Identifying hidden patterns among variables assists researchers and managers in developing informed and scientifically supported decisions conclusions.

13.2 NEED AND PURPOSE OF FACTOR ANALYSIS

In practical research settings, it is common to encounter datasets with many variables that collectively influence outcomes. When these variables are strongly related, understanding their combined behaviour becomes difficult. Factor Analysis provides a systematic way to simplify such data by highlighting common patterns linking variables. It organises related variables into fewer, more meaningful, and easier-to-interpret factors, thereby reducing complexity while preserving essential information.

Factor Analysis is instrumental when a researcher aims to:

- Group related variables into fewer, logical, and interpretable components.
- Simplify complex datasets without losing meaningful insights.
- Identify underlying dimensions or hidden structures that explain observed response patterns.
- Develop or validate measurement scales and questionnaires, ensuring that items collectively measure the intended construct.

The primary purposes of Factor Analysis:

1. **Data Reduction:** Reduces a large number of variables into a smaller set of key factors, making data easier to handle and interpret.
2. **Structure Detection:** Helps in identifying the underlying structure within the data by revealing how variables are grouped and the dimensions they form.
3. **Scale Development and Validation**: Supports the identification of item clusters measuring the same construct (e.g., satisfaction, service quality, attitudes), thereby contributing to the development of reliable and valid measurement instruments.

13.3 BASIC CONCEPTS IN FACTOR ANALYSIS

Factor Analysis relies on several essential statistical concepts that help in identifying and understanding the underlying structure of a set of variables. The key concepts are explained below:

a) Factor: A factor is a latent (hidden) dimension that influences a set of observed variables. It represents the common pattern or theme shared by those variables. Variables that load strongly on the same factor are considered to measure the same underlying construct. Therefore, a factor can be seen as a summary measure that captures the combined effect of multiple variables.

b) Factor Loading: Factor loading refers to the degree of association between a variable and a factor. It indicates how strongly a particular variable contributes to or is explained by a factor.

- The value of a factor loading functions similarly to a correlation coefficient.
- A higher loading suggests that the variable is strongly related to the factor.
- As a rule of thumb, loadings of **±0.40 or above** are usually considered meaningful and worthy of interpretation.

c) Communality: Communality reflects the proportion of variance in a variable that is accounted for by all the retained factors together. It measures how well a variable is represented in the factor solution.

- A communality value close to **1.00** indicates that the factors well explain the variable.
- A low communality suggests that the variable does not fit well within the identified factor structure and may need to be reconsidered.

d) Eigen Value: An **eigenvalue** is a measure of the **importance or explanatory power** of a factor.

- It represents the amount of total variance in the data that is explained by a particular factor.
- In practice, factors with an eigenvalue **greater than 1.00** are generally retained, as they explain more variance than a single original variable would.

e) Total Sum of Squares (TSS): The Total Sum of Squares (TSS) is the sum of the eigenvalues of all the factors under consideration. It represents the **total variance** in the dataset that can be explained. By comparing the variance explained by individual factors with the total variance, researchers evaluate how effectively the factor solution summarises the data.

13.4 PROCESS OF CONDUCTING FACTOR ANALYSIS

The process of Factor Analysis involves a sequence of structured steps to ensure that the factors extracted are meaningful and statistically sound. The significant steps are outlined below:

1. **Selection of Variables:** The first step is to identify and select the variables to include in the analysis. These variables should be conceptually relevant to the study and are expected to show some degree of interrelation. Factor analysis works best when the variables have logical and theoretical connections.
2. **Testing the Suitability of Data:** Before conducting Factor Analysis, it is crucial to evaluate whether the dataset is suitable for the method. Two common tests are used at this stage. The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy assesses the proportion of shared variance among variables; values of 0.6 or higher indicate that the data is appropriate. Bartlett's Test of Sphericity tests whether the correlation matrix significantly differs from an identity matrix; a significant result confirms that there are meaningful correlations among the variables.
3. **Extraction of Initial Factors:** After confirming eligibility, the next step is to extract the initial set of factors. Techniques such as Principal Component Analysis (PCA) or Common Factor Analysis are commonly employed. These methods detect patterns in the data and group variables based on shared variance.
4. **Determination of the Number of Factors:** Once the initial factors are extracted, the researcher must decide how many to retain for interpretation. This is usually done using the eigenvalue greater than one rule and the Scree Plot, which visually shows the point at which the contribution of additional factors begins to decrease.
5. **Rotation of Factors:** To clarify and simplify the factor structure, rotation is applied. Varimax rotation (orthogonal rotation) is used when factors are assumed to be independent, while Promax rotation (oblique rotation) is employed when factors are expected to be correlated. Rotation redistributes the variance across factors, making the pattern of factor loadings easier to interpret.
6. **Interpretation and Naming of Factors:** After rotation, the variables that load strongly on each factor are examined to identify the shared theme they represent. Based on this core meaning, each factor is assigned an appropriate and meaningful name.
7. **Validation and Application of Factors:** The final step is to validate the derived factor structure and use the factors in further statistical analysis or decision-making. These factors might be applied in building predictive models, segmenting consumers, designing marketing strategies, or developing measurement scales.

13.5 APPLICATIONS OF FACTOR ANALYSIS

Factor Analysis has a wide range of applications in social sciences, management, marketing, psychology, and behavioural research. It provides insights into the underlying data structures and helps researchers and managers make well-informed decisions. Some primary areas of application are outlined below.

1. New Product Development: In product development and innovation studies, organisations often analyse several product attributes to understand consumer preferences. Factor Analysis helps group these attributes into broader, meaningful factors such as product quality, price sensitivity, convenience, or brand appeal. By identifying the most influential factors, firms can design or modify products that better match consumer expectations. This ensures that product improvements focus on features that matter most to customers.

2. Consumer Segmentation: Factor analysis can also be utilised to group consumers rather than variables. When individuals exhibit similar response patterns on surveys or rating scales, they can be segmented into homogeneous groups. This segmentation helps organisations understand different consumer profiles, such as value-conscious buyers, quality-focused consumers, or brand-loyal customers. The insights gained from such segmentation assist managers in crafting targeted marketing strategies and personalised promotional campaigns.

3. Psychometric Scale Development: In fields such as psychology, management studies, and educational research, Factor Analysis is essential for developing and validating measurement scales. It assists in determining whether individual questionnaire items assess the same underlying construct, such as job satisfaction, organisational commitment, leadership style, or service quality. By identifying groups of closely related items, researchers can ensure that the scale is both reliable and valid.

4. Market Research and Strategic Decision-Making: Factor Analysis is widely used in market research to identify the key dimensions that influence consumer perceptions, attitudes, and purchasing decisions. By recognising these underlying factors, companies can gain a better understanding of market behaviour and develop more effective marketing strategies. For example, a company's brand positioning, promotional campaigns, and pricing strategies can be improved when managers understand the core factors affecting customer choices.

13.6 INTRODUCTION TO CONJOINT ANALYSIS

In contemporary markets characterised by intense competition, diverse product selections, and rapidly evolving consumer expectations, purchase decisions are rarely based on a single feature. Instead, consumers perceive products and services as comprehensive packages, evaluating how different attributes—such as brand recognition, functional performance, aesthetic appeal, price, durability, warranty, and after-sales support—contribute to overall value. This all-encompassing assessment reflects the reality that consumer preferences are complex and multi-dimensional.

Conjoint Analysis is a specialised quantitative research method used to understand how consumers value different features of a product or service. The term conjoint comes from the phrase "considered jointly", emphasising that consumers form preferences by evaluating combinations of features, not individual features in isolation. Instead of directly asking consumers which feature is most important, Conjoint Analysis infers the relative significance of features by observing how consumers rank, rate, or choose between various product profiles.

The theoretical foundation of Conjoint Analysis is based on utility theory, which states that consumers choose the option that offers the highest satisfaction (or utility). By estimating these utilities for different attribute levels, the method allows researchers to model consumer decision-making and predict responses to new product designs or market changes.

Therefore, Conjoint Analysis serves as a powerful tool in product design, pricing strategies, market segmentation, and competitive positioning, enabling organisations to align their offerings with consumer preferences and enhance market success. It provides managers with practical insights into which features consumers value most, how trade-offs are made among different attributes, and which configurations are likely to be chosen in real-world purchases situations.

13.7 PROCESS OF CONDUCTING CONJOINT ANALYSIS

The application of Conjoint Analysis involves a series of structured activities aimed at identifying consumer preferences for products made up of multiple attributes. The process can be summarised as follows:

Step 1: Identification of Relevant Attributes and Their Levels

The first step is to identify the set of attributes that significantly influence consumers' assessments of the product or service. Each attribute must represent a trait that the consumer considers when making a purchasing decision. For each attribute, the researcher defines a range of levels (variations or options).

Example:
For a **smartphone**, important attributes may include:

- **Brand** (Brand A, Brand B, Brand C)
- **Battery Life** (10 hours, 18 hours, 26 hours)
- **Camera Quality** (12 MP, 48 MP, 64 MP)
- **Price** (₹15,000; ₹25,000; ₹35,000)

Attributes must be **mutually exclusive**, **clearly defined**, and **realistic** so that respondents can evaluate them meaningfully.

Step 2: Construction of Product Profiles

Once the attributes and levels are identified, combinations of attribute levels are created to develop product profiles. Each profile represents a potential product configuration. However, when many attributes and levels are involved, the total number of combinations can become very large. Therefore, fractional factorial designs or orthogonal arrays are used to select a manageable number of statistically representative profiles.

Example:

A full combination (3×3×3×3) would generate 81 profiles. Using an orthogonal design, this could be reduced to about 12–16 profiles for respondents without sacrificing statistical accuracy.

Step 3: Presentation of Profiles to Respondents

The product profiles are then presented to respondents for evaluation. The mode of presentation may include written descriptions, visual mock-ups, or digital product cards. Respondents express their preferences using one of the following approaches:

- **Ranking** the product profiles in order of preference
- **Rating** each profile on a numerical scale (e.g., 1 to 10)
- **Pairwise Choice** (choosing the preferred profile from pairs)

The goal is to simulate realistic consumer judgment situations.

Step 4: Collection of Preference Data

The responses collected from ranking, rating, or choices show the perceived usefulness or desirability of each product profile. These data serve as the basis for estimating how each attribute level contributes to the overall preference.

Step 5: Estimation of Utility (Part-Worth) Values

Using statistical analysis—often regression analysis, logistic modelling, or specialised conjoint estimation procedures—the researcher determines utility scores (also called part-worths) for each attribute level. These scores show how much each level contributes to the consumer's overall satisfaction.

Interpretation of Example: If the utility for *Battery Life (26 hours)* is significantly higher than that for *Battery Life (10 hours)*, it indicates that consumers strongly value longer battery performance.

Step 6: Determination of Attribute Importance

The relative importance of each attribute is assessed by examining the range of utility scores across its levels. A broader range indicates a greater influence on consumer preference.

Example:

Attribute	**Utility Range**	**Relative Importance (%)**
Price	2.4	35%
Camera Quality	1.9	28%
Battery Life	1.5	22%
Brand	1.1	15%
Attribute	**Utility Range**	**Relative Importance (%)**

This suggests that **Price** has the most significant influence on consumer choice in this smartphone market.

Step 7: Interpretation and Managerial Decision-Making

The final step involves turning results into practical strategic choices. Conjoint outcomes assist managers in identifying:

- The **most preferred product configuration**
- **Trade-offs** consumers are willing to make (e.g., accepting a lower-resolution camera for lower price)
- The **market share potential** of alternative product designs
- Opportunities for **market segmentation** based on differences in preference patterns

Example Interpretation:
If the analysis shows the highest utility profile as:
Brand B + 26-hour Battery + 48 MP Camera + ₹25,000,
The company may prioritise developing and promoting this configuration as its flagship offering.

Illustrative Summary Example

Attribute	**Selected Level (Most Preferred)**
Brand	Brand B
Battery Life	26 hours
Camera Quality	48 MP
Price	₹25,000

This combination provides the **highest total utility**, indicating the **optimal smartphone design** from the customer's perspective.

13.8 APPLICATIONS OF CONJOINT ANALYSIS

Conjoint Analysis is a vital analytical tool for understanding consumer preference structures and assisting strategic marketing decisions. Estimating the relative significance of product or service attributes allows organisations to develop offerings that meet customer expectations and suit competitive market conditions. Some of the key areas where Conjoint Analysis is utilised include:

1. **New Product Development and Product Improvement:** Conjoint Analysis assists firms in identifying the optimal combination of attributes to incorporate into new products. By simulating how consumers evaluate alternative product designs, organizations can determine which features are essential, desirable, or unnecessary. This helps in reducing development risk and improving market acceptance of new products.
2. **Pricing Strategy and Value-Based Pricing:** Since consumers' preferences are influenced by both product characteristics and price, Conjoint Analysis provides insight into how sensitive customers are to price changes in relation to other attributes. Firms can determine acceptable price ranges, evaluate trade-offs consumers are willing to make, and adopt **value-based pricing strategies** that reflect perceived benefit rather than cost alone.
3. **Market Segmentation and Targeting:** Conjoint Analysis can identify segments of consumers who value attributes differently. By grouping consumers based on their preference patterns, firms can tailor product designs, communication messages, and promotional efforts to specific target segments, enhancing competitive advantage.
4. **Competitive Strategy and Market Share Simulation:** Organisations can apply conjoint results to predict how changes in product attributes (either by themselves or by competitors) will influence consumer switching behaviour. This enables firms to estimate potential shifts in market share, assess competitive threats, and make informed strategic decisions on positioning and product differentiation.
5. **Brand Positioning and Repositioning:** Conjoint Analysis reveals the attributes most strongly associated with brand perception and preference. This helps marketers position their brand effectively in the consumer's mind or reposition existing offerings to strengthen brand identity in a competitive environment.
6. **Service Design and Service Quality Optimisation:** The method is not limited to tangible goods; it is equally helpful in service industries such as banking, telecom, healthcare, tourism, aviation, and hospitality. By analysing preferences for service components—such as response time, employee behaviour, accessibility, digital support, or customisation—organisations can redesign service delivery to enhance customer satisfaction and loyalty.
7. **Forecasting Customer Response to Strategic Changes:** Conjoint models enable firms to conduct "what-if" analyses, predicting how consumers may respond to changes in product configuration, pricing, or promotional strategy. This helps plan product launches, modify existing offerings, and evaluate alternative marketing strategies before implementation.

13.9 SUMMARY

Factor Analysis is a technique that reduces complex data sets by grouping correlated variables into fewer meaningful factors. Key statistical concepts include factor loadings, communalities, and eigenvalues. It is highly useful in areas such as product development, market segmentation, consumer behaviour analysis, and scale development. Simplifying data structures provides clearer insights for managerial decision-making.

Conjoint Analysis is an advanced research technique designed to measure consumer preferences for products consisting of multiple attributes. By analysing how consumers evaluate different combinations of attributes, organisations can determine the relative importance of features and identify the most desirable product configurations. Despite the complexity of its design and data interpretation, Conjoint Analysis remains a valuable tool for strategic decision-making in areas such as product development, pricing, and competitive positioning.

13.10 TECHNICAL TERMS

1. **Factor:** A factor refers to an underlying or latent dimension that influences a group of observed variables. Highly correlated variables are understood to be governed by the same factor.
2. **Factor Loading:** Factor loading represents the degree of relationship between a variable and a factor. It indicates how strongly a variable contributes to the meaning of that factor. Higher loading values suggest a stronger association.
3. **Communality:** Communality is the proportion of variance in an observed variable that is explained jointly by all the retained factors. A higher communality value indicates that the variable is well represented in the factor structure.
4. **Eigen Value:** An eigen value reflects the relative importance of a factor. It shows the amount of total variance accounted for by a particular factor. Generally, factors with eigenvalues greater than one are considered significant.
5. **Varimax Rotation:** Varimax rotation is a commonly used technique in factor analysis to simplify the interpretation of factors. It redistributes variance across factors in such a way that each factor becomes easier to interpret by producing clearer and more distinct factor loadings.
6. **Attribute:** A measurable characteristic or feature of a product or service that influences consumer evaluation.
7. **Attribute Level:** The specific variations or alternatives within each attribute offered for comparison.
8. **Utility:** The degree of satisfaction or perceived value a consumer derives from a product or attribute combination.
9. **Part-Worth:** The estimated utility contribution of each attribute level in shaping overall consumer preference.
10. **Profile:** A product or service description formed by combining one level from each attribute for evaluation.

13.11 SELF-ASSESSMENT QUESTIONS

1. What is Factor Analysis?
2. Define Factor Loading.
3. What is the purpose of factor rotation?
4. Why is Conjoint Analysis based on the concept of evaluating attribute combinations rather than individual attributes?
5. What is the significance of utility in Conjoint Analysis?
6. Explain the process of Factor Analysis with suitable examples.
7. Discuss the applications of Factor Analysis in marketing research.
8. Describe the meaning of communality and eigenvalue with interpretation.
9. Explain the steps involved in conducting Conjoint Analysis with a suitable example.
10. Discuss the applications of Conjoint Analysis in new product development and pricing decisions.
11. Evaluate the advantages and limitations of Conjoint Analysis for managerial decision-making.

13.12 SUGGESTED READINGS

1. Malhotra, N.K. (2019). *Marketing Research: An Applied Orientation*. Pearson.
2. Hair, J.F., Anderson, R.E., Tatham, R.L., & Black, W.C. (2020). *Multivariate Data Analysis*. Prentice Hall.
3. Kothari, C.R. (2011). *Research Methodology: Methods and Techniques*. New Age International Publishers.
4. Green, P. E. & Srinivasan, V. (1978). "Conjoint Analysis in Consumer Research." *Journal of Consumer Research*.

Dr. G. MALATHI

LESSON-14

MULTIDIMENSIONAL SCALING & CLUSTER ANALYSIS

OBJECTIVES OF THE LESSON

After completing this chapter, students will be able to:

1. Understand the concept, purpose, and importance of Multidimensional Scaling (MDS) in research.
2. Apply MDS to understand consumer perceptions and competitive positioning.
3. Recognize the relationship between MDS and related multivariate techniques such as Factor Analysis.
4. Understand the concept and purpose of Cluster Analysis in research.
5. Identify situations where Cluster Analysis is applicable.
6. Explain the steps involved in conducting Cluster Analysis.

STRUCTURE OF THE LESSON

14. 1 INTRODUCTION TO MULTIDIMENSIONAL SCALING

In the fields of marketing, psychology, social sciences, and managerial decision-making, researchers often encounter constructs such as perceptions, attitudes, preferences, and image. These constructs are intangible and subjective, existing within individuals' cognition rather than in directly observable or measurable form. For example, consumers might perceive one smartphone brand as more stylish or view a particular airline as more comfortable compared to others. Such judgments are inherently personal and are shaped by experience, expectations, cultural context, and individual value systems. Because these perceptions cannot be measured using simple numerical scales, specialised analytical tools are required to represent them meaningfully.

Multidimensional Scaling (MDS) is an advanced statistical method. It allows researchers to measure and visually display perceived similarities or differences among objects—such as

brands, products, or service providers—by converting perceptual data into a geometric spatial arrangement, usually in two or three dimensions. The visual display created through MDS is often called a Perceptual Map.

A perceptual map enables researchers and decision-makers to see how consumers mentally position different options in relation to one another. This visual tool helps identify competitive clusters, differentiation patterns, market gaps, and strategic opportunities. By understanding how products or brands are perceived, organisations can enhance their positioning strategies, communication messages, product development choices, and competitive responses more effectively and with greater insight.

Concept and Working Mechanism of Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) is essentially based on the idea that perceptions of similarity or dissimilarity among a set of objects can be represented spatially. The objects analysed may include brands, products, services, attributes, or consumer preference profiles. MDS helps convert qualitative perceptual judgments into a measurable visual representation, where the relative positions of objects indicate how closely or distinctly they are perceived by respondents.

At its core, MDS addresses one essential question:

How can we represent consumers' subjective perceptions in an objective, measurable way?

The technique achieves this by mapping objects within a geometric space, where distance signifies perceived similarity — the smaller the distance, the greater the similarity; the larger the distance, the greater the perceived difference.

14.2 TYPES OF MULTIDIMENSIONAL SCALING

Multidimensional Scaling (MDS) can be broadly divided into two main types based on the nature of the input data and the underlying measurement assumptions. These are Metric MDS and Non-Metric MDS. Both methods aim to depict perceived similarities or dissimilarities among objects within a geometric space, but they differ in how they handle the measurement scale of the input data.

(a) Metric Multidimensional Scaling

Metric MDS is used when the similarity or dissimilarity data are measured on **interval or ratio scales**, where the numerical values represent meaningful and measurable differences between objects. In this approach, the distances in the perceptual map are required to closely correspond to the actual dissimilarity values. Hence, Metric MDS preserves both the **order and magnitude** of differences across objects. This method produces a **more precise and mathematically rigorous spatial representation**, making it appropriate in situations where respondents can accurately quantify how different or similar objects are (e.g., physical measurements, performance ratings, or sensory tests in product development).

(b) Non-Metric Multidimensional Scaling

Non-Metric MDS is suitable when the input data are measured on an **ordinal scale**, where respondents provide **rankings** rather than exact numerical ratings. Instead of trying to match the exact magnitude of differences, Non-Metric MDS focuses on preserving the **relative order** of similarities or dissimilarities among objects. This makes Non-Metric MDS particularly valuable in **marketing and consumer behavior studies**, where perceptions, preferences, and judgments are often subjective and difficult to quantify precisely. Here, what matters is not the exact distance between brands but whether a consumer perceives one brand as *more similar* to another relative to a third.

Steps in Conducting MDS

1. **Collection of Perception Data:** Respondents are asked to evaluate how similar or different each pair of objects is. This evaluation may be done using:
    - Rating scales (e.g., 1 = Very Similar, 7 = Very Different)
    - Rank ordering
    - Paired comparison judgments
2. **Construction of a Dissimilarity Matrix:** The collected data are organized into a square matrix, where each cell represents the degree of perceived difference between two objects.
   This matrix serves as the input for MDS.
3. **Application of the MDS Algorithm:** Statistical software applies an optimization algorithm that locates each object in a multidimensional space, such that the distances between points reflect perceived dissimilarities as closely as possible.
4. **Generation of a Perceptual Map:** The resulting map visually displays the objects:
    - Objects close together → Perceived as similar
    - Objects far apart → Perceived as dissimilar
5. **Interpretation and Labelling of Dimensions:** Researchers examine the configuration and infer the underlying attributes that best describe the axes. Common perceptual dimensions include:
    - Price (High to Low)
    - Quality (Premium to Basic)
    - Style (Modern to Traditional)
    - Reliability (Consistent to Uncertain)

This interpretation is not provided automatically by the software; it requires researcher judgment and contextual understanding.

Illustrative Example: Perceptual Mapping of Smartphone Brands

Suppose a researcher aims to understand how consumers perceive four smartphone brands—Brand A, Brand B, Brand C, and Brand D—in terms of overall style and sophistication. Respondents are asked to rate the similarity between each pair of brands. These similarity ratings are then converted into a dissimilarity matrix, which serves as input for Multidimensional Scaling (MDS).

After applying MDS, the researcher obtains a perceptual map, where the brands are positioned based on perceived similarities.

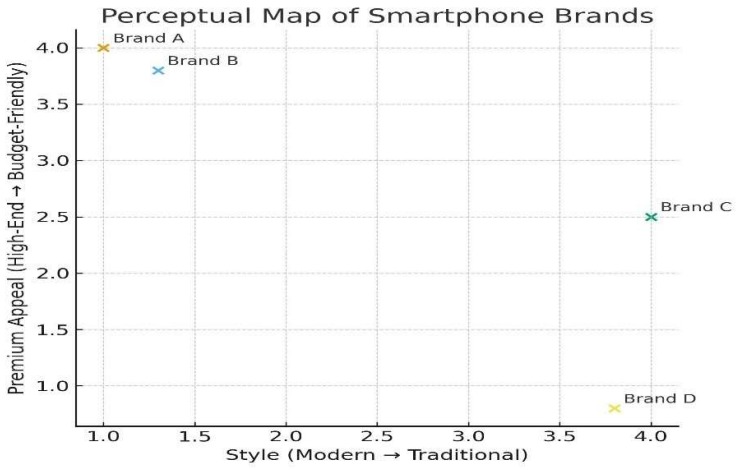


Figure 1: Perceptual Map of Smartphone Brands Based on Style and Premium Appeal

Interpreting the Dimensions

The perceptual map reveals two meaningful dimensions:

- Horizontal Axis: Style (Modern ↔ Traditional)
- Vertical Axis: Premium Appeal (High-End ↔ Budget-Friendly)

These dimensions reflect how consumers mentally categorise the smartphone brands.

Interpretation of the Perceptual Map

- Brand A and Brand B appear close to each other, indicating that consumers perceive them as similar, likely representing modern, sleek, premium-style smartphones.
- Brand C is placed farther away from these two brands, suggesting that it is perceived as more traditional or less stylish.
- Brand D is located at a more distant position, indicating a distinct market identity, possibly associated with rugged or utility-focused design.

Managerial Implications

Insights from the perceptual map can guide strategic decisions such as:

- **Brand Repositioning:** Adjusting brand image and message to influence perception.
- **Target Segment Selection:** Identifying which consumer groups associate with each brand.

- **Advertising Communication:** Designing promotional messages aligned with perceived brand characteristics.
- **Competitive Strategy:** Understanding market proximity and differentiation opportunities.

14.3 APPLICATIONS OF MULTIDIMENSIONAL SCALING (MDS)

Multidimensional Scaling has become a vital analytical tool in business, management, psychology, and social sciences because of its ability to transform subjective perceptions into visual representations. By mapping objects or concepts in a perceptual space, MDS allows researchers and decision-makers to recognise underlying patterns in human judgment. The following are some key applications of MDS:

1. **Brand Positioning:** MDS is widely used in marketing to assess how different brands are perceived in consumers' minds. By plotting brands on a perceptual map, organisations can see which brands are viewed as similar or distinct. This helps managers evaluate their brand's competitive position and spot opportunities for repositioning or differentiation.
2. **Product Development and Innovation:** In product development, MDS assists in identifying the attributes that consumers value most. By studying how consumers perceive similarities among products, firms can detect unmet needs and develop features that enhance product appeal. This technique is particularly useful in industries such as automobiles, electronics, and fast-moving consumer goods.
3. **Market Segmentation:** MDS helps categorise consumers into groups based on similarity of attitudes, perceptions, or preferences. Unlike demographic segmentation, perceptual segmentation focuses on how consumers think about products and brands. This leads to more meaningful and targeted marketing strategies.
4. **Competitor Analysis:** By displaying competing products or brands in a perceptual space, MDS reveals the degree of competitive proximity among them. Brands that appear close together are perceived similarly and therefore compete directly, while those located further apart serve different market niches. Such insights assist firms in shaping strategic responses to competitors.
5. **Service Quality Assessment:** In service-based industries such as hospitality, education, healthcare, and banking, MDS is used to compare perceptions of service quality across providers. It enables organisations to understand how customers perceive attributes like reliability, responsiveness, ambience, and customer care. This helps firms enhance service standards and strengthen customer loyalty.
6. **Psychological and Behavioural Research:** In psychology, MDS is used to explore how individuals mentally represent abstract constructs such as emotions, attitudes, stereotypes, or personality traits. By examining similarity judgments, researchers can uncover hidden cognitive structures and understand how people organise complex information in memory.

14.4 MDS VS. FACTOR ANALYSIS

Both Multidimensional Scaling (MDS) and Factor Analysis (FA) are data reduction and representation techniques used in multivariate analysis. Though they share the objective of simplifying complex data structures, they differ fundamentally in what they analyze, the type of input data, and the way results are interpreted.

Aspect	***Multidimensional Scaling***	***Factor Analysis***
Input Data	Similarity/Dissimilarity judgments	Inter-correlations among variables
Output	Spatial map of objects	Underlying latent factors
Purpose	Understand perceptions and relationships	Identify hidden dimensions influencing variables
Best Used For	Brand positioning, competitive mapping	Dimensionality reduction and scale development

14.5 INTRODUCTION TO CLUSTER ANALYSIS

Cluster Analysis is a multivariate statistical method used to group objects—such as individuals, products, organisations, or geographic locations—into relatively similar groups called clusters. The core of this method is to find patterns of similarity among objects based on chosen features, so that objects within the same cluster are more alike than those in different clusters. Therefore, cluster analysis aims to maximise internal uniformity within clusters and maximise external differences between clusters.

Cluster Analysis does not depend on predefined groups; instead, it reveals natural patterns within the data. This makes it especially useful in exploratory research, where the aim is to identify meaningful classifications that might not be immediately obvious.

In fields such as marketing, psychology, sociology, healthcare, retailing, and management research, cluster analysis acts as a powerful tool for segmentation and strategic decision-making. For example, in marketing, it allows organisations to identify customer segments that differ in needs, preferences, or buying behaviour, thereby supporting the development of targeted marketing strategies. Similarly, in organisational research, employees can be grouped based on skills, work values, or performance orientations to inform training, recruitment, and human resource planning. Cluster analysis helps researchers and decision-makers simplify complex datasets, uncover latent structures, and derive insights that contribute to more informed and effective strategic planning.

Purpose of Cluster Analysis

The key purposes of employing cluster analysis include:

1. **Simplification of Complex Data:** Cluster analysis reduces a large and complex dataset into a smaller number of meaningful groups, thereby facilitating easier interpretation and understanding of relationships among observations.
2. **Identification of Homogeneous Segments:** It helps uncover distinct segments that share similar characteristics, attitudes, demographic attributes, preferences, or behaviors. This allows researchers to understand how subgroups differ within a broader population.
3. **Strategic Decision Support:** The identification of clusters provides a basis for strategic and managerial decisions. Insights from clustering support decisions related to product differentiation, market segmentation, competitive positioning, retail store placement, inventory management, and customer relationship strategies.

For example, a marketing manager might want to group consumers based on lifestyle, demographic factors, purchasing motivations, or brand loyalty. Using cluster analysis, customer segments such as price-sensitive buyers, quality-focused consumers, brand loyalists, and trend-seeking youth can be identified. Once these segments are recognised, companies can create targeted marketing campaigns, tailor product offerings, and adjust pricing strategies to better meet the specific needs of each group.

14.6 TYPES OF CLUSTERING METHODS

Cluster Analysis can be carried out using several methodological approaches, the choice of which depends on the nature of the data, sample size, and the researcher's objectives. Broadly, clustering techniques are grouped into two major categories: **Hierarchical Clustering** and **Non-Hierarchical Clustering**.

A. Hierarchical Clustering: Hierarchical clustering is a procedure that builds a sequence of nested clusters either by successively **merging** smaller clusters or **dividing** larger ones. The process operates in one of two ways:

1. **Agglomerative Approach (Bottom-Up Method)**
   - Begins with each object considered as an individual cluster.
   - Clusters that are most similar are progressively merged at each step.
   - This continues until all objects are combined into a single cluster.
2. **Divisive Approach (Top-Down Method)**
   - Starts with all objects in one single cluster.
   - The cluster is then split into smaller clusters in stages until each object stands alone.

The results of hierarchical clustering are typically represented using a **dendrogram**, a tree-like diagram that illustrates the step-by-step process of cluster formation. By visually examining the dendrogram, the researcher can determine the appropriate number of clusters that best represent the underlying structure in the data.

B. Non-Hierarchical Clustering (K-Means Clustering)

Non-hierarchical methods, particularly **K-Means Clustering**, are widely used when the researcher wishes to directly partition the dataset into a pre-specified number of clusters (denoted as *K*). The K-means algorithm operates as follows:

- The researcher selects the number of clusters in advance.
- Initial cluster centers (or centroids) are assigned.
- Each object is assigned to the cluster with the closest centroid.
- Cluster centroids are recalculated based on newly formed clusters.
- The process iterates until the cluster assignments stabilise and no further changes occur.

K-means is computationally efficient and well-suited for handling large datasets.

Comparison of Hierarchical and K-Means Clustering Methods

Feature	***Hierarchical Clustering***	***K-Means (Non-Hierarchical) Clustering***
Number of Clusters	Determined by examining the dendrogram	Must be specified before analysis
Suitability for Sample Size	Best for small to moderate datasets	Suitable for large datasets
Output/Representation	Produces a dendrogram showing cluster structure	Provides final cluster centroids and group membership
Computational Demand	Higher, especially for large datasets	Lower and more efficient for large datasets
Flexibility in Reassignment	Once cluster assignments are made, they cannot be undone	Allows objects to move between clusters during iteration

14.7 PROCESS OF CONDUCTING CLUSTER ANALYSIS

The application of Cluster Analysis involves a systematic procedure to ensure that the resulting clusters are meaningful, interpretable, and useful for decision-making. The following steps outline the sequential process followed in conducting cluster analysis:

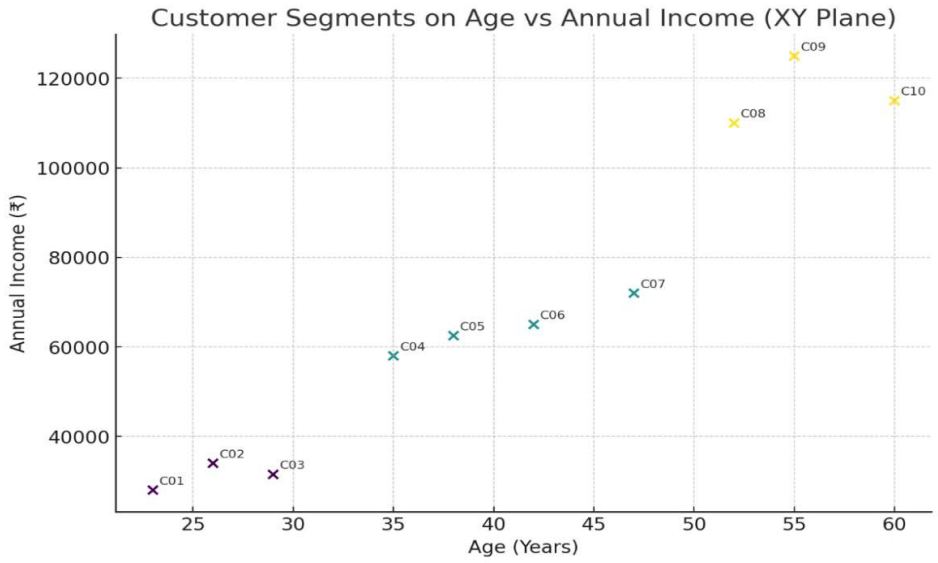
1. **Define the Research Problem and Objectives:** The first step is to clearly articulate the purpose of clustering and identify the entities to be grouped—for example, customers, brands, retail outlets, or geographic regions. A precise statement of the objective ensures that the clustering results are relevant and actionable.
2. **Selection of Variables for Clustering:** Variables that appropriately represent the characteristics of the objects must be identified. These may include demographic factors (age, income), psychographic attributes (lifestyle, values), behavioral data (usage frequency, purchase patterns), or product features. The quality of the clusters depends heavily on the relevance and accuracy of the variables selected.

3. **Standardisation of Data:** Because variables may be measured in different units (e.g., income in rupees and age in years), it is essential to standardise or normalise the data. Standardization ensures that no single variable dominates the clustering process due to scale differences and that all variables contribute proportionately.
4. **Selection of a Similarity or Distance Measure**: Cluster formation is based on the concept of similarity or dissimilarity among objects. Standard distance measures include Euclidean distance, Manhattan distance, correlation-based measures, and simple matching coefficients. The choice of measure depends on the type of data and research context.
5. **Selection of Clustering Method:** The researcher must choose between **Hierarchical** and **Non-Hierarchical** (e.g., K-Means) clustering techniques.
    - **Hierarchical methods** build clusters progressively and are suitable for small to medium samples.
    - **Non-Hierarchical methods** require specifying the number of clusters in advance and are efficient for large datasets.
6. **Derivation and Validation of Clusters:** Once clusters are generated, their stability, consistency, and distinctiveness must be assessed. This involves examining the separation between clusters and the internal similarity within each cluster. Techniques such as statistical tests, comparison with external criteria, or replication with different samples may be used for validation.
7. **Labeling and Profiling the Clusters:** The final step is to interpret and describe each cluster meaningfully. Descriptive statistics are used to characterize the clusters based on the variables included. Clear and concise labels (e.g., "Value Conscious Consumers," "Brand Loyal Professionals," or "Youth Trendsetters") enhance understanding and support decision-making.

Illustrative Example of Cluster Analysis: Customer Segmentation Based on Age and Annual Income

To understand how cluster analysis works, consider a small sample of customers:

Customer	Age (Years)	Annual Income (₹)
C01	23	28,000
C02	26	34,000
C03	29	31,500
C04	35	58,000
C05	38	62,500
C06	42	65,000
C07	47	72,000
C08	52	1,10,000
C09	55	1,25,000
C10	60	1,15,000



The diagram visually represents each customer's position on the Age–Income plane and confirms the three natural groupings:

- Budget Conscious Youth (lower-left region): Young customers with relatively low annual incomes. They cluster tightly near the lower values on the Y-axis and lower values on the X-axis. These customers are likely entry-level earners — appropriate targets for budget products and introductory offers.
- Middle-Class Family Buyers (central band): Customers in their mid-30s to late 40s with moderate incomes. They occupy the middle region on both axes, indicating stable earning capacity and family-oriented purchasing needs. Marketing strategies for this cluster may emphasise value for money, durability, and family benefits.
- Affluent Premium Customers (upper-right region): Older customers with high annual incomes. They are clearly separated from the other clusters along the Y-axis (high income) and are positioned toward higher ages on the X-axis. This segment is suitable for premium products, personalised services, and higher-end promotions.

14.8 SUMMARY

Multidimensional Scaling is a valuable technique in research for examining how individuals perceive products, brands, or concepts. It converts subjective judgments of similarity or difference into a visual spatial representation called a perceptual map. MDS supports strategic marketing activities, including branding, product development, market segmentation, and competitive analysis. While related to Factor Analysis, MDS is distinct in its focus on mapping objects rather than uncovering underlying variable structures.

Cluster Analysis is a vital research tool used for grouping individuals or objects based on similarity. It helps organisations identify meaningful segments for decision-making. Both hierarchical and non-hierarchical clustering methods aid in discovering patterns and support

strategic business planning. Accurate interpretation and validation ensure that the clusters formed are useful and actionable.

14.9 TECHNICAL TERMS

1. **Multidimensional Scaling (MDS):** A statistical technique that maps perceived similarities among objects.
2. **Perceptual Map:** A visual representation of the positioning of objects based on perceptions.
3. **Similarity/Dissimilarity Data:** Information showing how alike or different objects are perceived to be.
4. **Metric MDS:** Uses numerical scales to measure perceptions.
5. **Non-Metric MDS:** Uses ranked or ordered perceptions.
6. **Stress Value:** A measure of how well the perceptual map represents the observed data (lower stress indicates better fit).
7. **Cluster**: A group of objects or individuals that are similar to one another based on selected characteristics.
8. **Cluster Analysis**: A statistical technique used to classify objects into homogeneous groups (clusters) based on similarity.
9. **Segmentation**: The process of dividing a larger population into smaller, meaningful, and homogeneous sub-groups.
10. **Similarity Measure**: A numerical measure that indicates how alike two objects are. Higher similarity means the objects are more alike.
11. **Distance Measure**: A numerical measure of how different two objects are. A larger distance indicates greater dissimilarity.
12. **Hierarchical Clustering**: A clustering method that builds a hierarchy of clusters either by merging smaller clusters (agglomerative) or splitting large clusters (divisive).
13. **Agglomerative Method**: A bottom-up approach in hierarchical clustering where each object starts as a separate cluster and clusters are merged step-by-step.
14. **Divisive Method**: A top-down approach where all objects start in one cluster and are divided into smaller clusters step-by-step.
15. **Dendrogram**: A tree-like diagram used in hierarchical clustering that shows how clusters are formed at each stage.
16. **Non-Hierarchical Clustering (K-Means Clustering)**: A clustering method where the number of clusters is specified in advance, and objects are grouped by minimizing the distance to cluster centers.
17. **Cluster Centroid**: The average value of all variables for the members of a cluster; used to describe the cluster in K-means clustering.
18. **Standardization (Normalization)**: A data preparation process in which variables are scaled to ensure equal weight in cluster formation.
19. **Profile Analysis**: The process of describing the characteristics of each cluster by examining the mean or frequency of variables.
20. **Cluster Validity**: The process of evaluating whether the clusters are meaningful, stable, and useful for decision-making.
21. **Homogeneity Within Clusters**: The degree to which objects inside a cluster are similar.
22. **Heterogeneity Between Clusters**: The extent to which clusters differ from one another.

23. **Segment Size Estimation**: Determining how many objects or individuals belong to each cluster/segment.
24. **Market Segmentation**: The application of cluster analysis in marketing to classify customers into distinct groups with similar needs or characteristics.

14.10 SELF-ASSESSMENT QUESTIONS

1. Define Multidimensional Scaling and state its purpose.
2. What is a perceptual map? Explain with an example.
3. Differentiate between Metric and Non-Metric MDS.
4. How is MDS functional in brand positioning?
5. What is the Stress Value in MDS interpretation?
6. Define Cluster Analysis and explain its objectives.
7. Differentiate between hierarchical and non-hierarchical clustering.
8. Explain the steps involved in conducting Cluster Analysis.
9. Explain the process of constructing a perceptual map using MDS.
10. Discuss the role of MDS in product and market strategy decisions.
11. Compare and contrast Multidimensional Scaling and Factor Analysis.
12. Discuss the role of Cluster Analysis in customer segmentation.
13. What is the importance of similarity and distance measures in clustering

14.11 SUGGESTED READINGS

1. Malhotra, N.K. (2019). Marketing Research: An Applied Orientation. Pearson.
2. Hair, J.F., Black, W., Babin, B., & Anderson, R. (2018). Multivariate Data Analysis. Cengage.
3. Green, P.E., & Srinivasan, V. (1978). Conjoint Analysis in Consumer Research. Journal of Consumer Research.
4. Sharma, S. (1996). Applied Multivariate Techniques. Wiley.

Dr. G. MALATHI

LESSON-15

INTRODUCTION TO DATA ANALYSIS USING SPSS

OBJECTIVES OF THE LESSON

After studying this lesson, students will be able to:

1. Explain the concept and significance of automated data analysis in modern research.
2. Describe the features and uses of SPSS software in managing and analysing research data.
3. Create a new data file in SPSS, including assigning variable names, labels, and coding responses.
4. Enter data accurately in SPSS using Variable View and Data View.
5. Identify and assign missing value codes to ensure data analysis remains meaningful and reliable.
6. Generate value labels for categorical variables to facilitate clear interpretation during statistical analysis.

STRUCTURE OF THE LESSON

15.1 INTRODUCTION TO DATA ANALYSIS

In contemporary research practices, the volume and complexity of data have risen substantially due to advances in digital technology, online surveys, organisational databases, social media analytics, and other forms of electronic data collection. Manual methods of analysing data, though essential, can be time-consuming, susceptible to errors, and limited in their ability to manage large datasets. To address these challenges, researchers increasingly depend on automated data analysis systems.

Automated data analysis involves the systematic use of computer-based software tools to process, organise, evaluate, and interpret data with minimal manual effort. These tools aim to streamline the analytical process by offering standardised procedures for data management, statistical testing, and visualisation. This makes automated analysis not only efficient but also reliable and replicable, which are essential qualities in scientific research.

Automation in data analysis assists researchers in numerous ways. It improves accuracy by minimising human calculation mistakes, accelerates analysis through rapid processing of complex statistical tests, and enhances interpretability by producing clear tables, charts, and output reports. As research increasingly becomes evidence-based across fields such as social sciences, business and management, health sciences, education, and economics, employing automated analytical tools has become essential rather than optional.

Among the various software options available, SPSS (Statistical Package for the Social Sciences) is one of the most widely used and trusted platforms. Its graphical user interface enables users with limited statistical training to perform sophisticated analyses. SPSS offers a broad range of analytical functions, from basic descriptive statistics to advanced inferential and multivariate techniques. Consequently, learning to use SPSS empowers students and researchers to conduct rigorous, data-driven investigations and present their findings convincingly.

15.2 OVERVIEW OF SPSS SOFTWARE

SPSS, originally developed as the Statistical Package for the Social Sciences, is a comprehensive software system designed to facilitate systematic data handling, statistical analysis, and interpretation of results in research. Over time, its application has extended far beyond the social sciences, becoming a widely adopted analytical tool in fields such as business administration, healthcare, education, economics, behavioural sciences, and public policy. Its popularity arises from its intuitive design, powerful analytical capabilities, and ability to manage both small-scale and large-scale datasets efficiently.

SPSS functions through a highly interactive interface supported by menus, dialog boxes, and command syntax. This dual mode of operation enables both beginners and experienced researchers to work comfortably. Beginners can perform analyses using point-and-click menus, while advanced users may rely on syntax commands for greater precision, automation, and reproducibility of results. The software also allows users to import data from various sources, including Excel, CSV, text files, and databases, ensuring flexibility in data collection.

SPSS supports the entire analytical process—beginning with data entry and preparation, then progressing to statistical analysis, and concluding with the interpretation and presentation of results. The software automatically generates structured output tables and visual displays, enhancing clarity and supporting evidence-based decision-making.

15.3 KEY FEATURES OF SPSS

- **User-Friendly Interface:** The menu-driven structure reduces the need for advanced programming knowledge, making it accessible to learners and professionals at all stages.
- **Comprehensive Statistical Tools:** SPSS supports descriptive statistics, correlation, regression, inferential tests, non-parametric tests, and advanced multivariate analysis, catering to diverse research requirements.
- **Data Visualization Capabilities:** SPSS includes tools for creating high-quality charts, histograms, scatter plots, and other graphical representations that aid in interpreting data patterns.

- **Efficient Handling of Large Datasets:** The software is designed to process and analyze datasets with a large number of variables and observations without compromising speed and accuracy.
- **Flexible Export and Reporting Options:** Results can be exported into Word, Excel, PDF, or image formats, enabling seamless incorporation of tables and charts into research reports, dissertations, and presentations.

15.4 CREATING A DATA FILE

This lesson explains how to set up a file with new data. After finishing this lesson, you should be able to create an SPSS data file that will include the data and some labeling that gives more detail about the data. To illustrate this process, we will use a shortened version of the questionnaire used by the General Social Survey conducted by the National Opinion Research Center. For this example, students wanted to see if their opinions on social issues were similar to those of the national sample.

The students knew they were not a representative sample, even of college students, but this questionnaire is an interesting way to learn how to create a new data file. They decided to use the following questions[1]:

- What is your age?
- Are you male or female?
- What is your religious preference?
- Generally speaking, in politics, do you consider yourself as conservative, liberal, or middle of the road?
- What kind of marriage do you think is the more satisfying way of life: one where the husband provides for the family and the wife takes care of the house and children, or one where both the husband and wife have jobs, and both take care of the house and children?
- Do you think it should be possible for a pregnant woman to obtain a legal abortion?
    - If there is a strong chance of a serious defect in the baby?
    - If she is married and does not want any more children?
    - If the woman's own health is seriously endangered by pregnancy?
    - If the family has a very low income and cannot afford any more children?
    - If she became pregnant as a result of rape?
    - If she is not married and does not want to marry the man?
    - If the woman wants it for any reason?

Basic Steps in Creating a Data File

It is best to start a data file with some careful planning.

[1] A copy of this questionnaire is included as Appendix at the end of this Lesson .

1. First, we will assign each respondent an identification number. This is not so we can identify individuals, but so we can keep track of each case when we go back to check the accuracy of the data entry. Each question is a variable in our data set. It needs a variable name that is simple but expresses something important about the data. (SPSS limits variable names to 64 characters or fewer. They may be numbers or letters but not spaces and very few special characters, so don't use any odd symbols.) *Age* and *sex* would be good variable names for the first two questions.[2] For the questions on abortion, we decided to use the first three characters of the variable names used by the General Social Survey. We used *mg* for the preferred type of marriage and called political orientation *conlib*. Each variable name can be given an extended variable label that gives more detail. (Extended variable labels can use spaces or special characters.) For example, *conlib* could have a variable label that said Conservative-Liberal.

2. After we have given each variable a name and label, we give each possible response to the question a code that is often the number corresponding to the order of the answers. (We could use another system, but this is the easiest because SPSS works best with numeric codes to represent the data.) For example, *sex* could use 1 for male and 2 for female; *conlib* could use 1 for conservative, 2 for liberal, and 3 for middle of the road. Values would then be given value labels such as Male, Female, Conservative, Liberal, and Middle of the Road.

3. Sometimes respondents do not answer a question, give more than one answer, or do something else that makes their answers unusable. In our example, respondent #2 marked both yes and no on the last question, respondent #3 wrote in none on question 4, and respondent #13 didn't answer the marriage question. We assign these missing value codes so they don't distort the analysis. Often 9 is used to indicate missing data or 99 if it is a two-digit value.

Everything must be planned carefully before entering the data into SPSS. It is useful to put the data in a matrix like Table 15-1 before entering it into the SPSS Data Editor. For this exercise, we will use only the first four questions and five respondents. (The complete matrix is in Appendix 2.B at the end of this lesson)

Table 15-1 Matrix for Data-entry Exercise

Id	*Age*	*Sex*	*Rel*	*conlib*
01	**20**	**1**	**4**	**2**
02	**24**	**2**	**5**	**2**
03	**21**	**2**	**2**	**9**
04	**24**	**2**	**5**	**3**
05	**26**	**2**	**4**	**2**

[2] For this exercise, we used lower-case italics for the variable names.

15.5 GETTING STARTED IN SPSS

To create the data file in SPSS, open SPSS (probably by clicking on the SPSS icon on the desktop). (See Figure 15-1.) Click on Close to close this window.

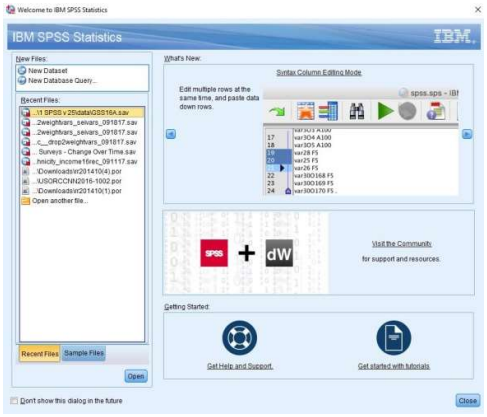


Figure 15-1

This opens a matrix similar to a spreadsheet such as Excel or the matrix we just worked on. The rows will be the cases (the respondents) and the columns will be the variables (answers to the questions). So, the upper-left cell will contain an identification number for the first case and the cells to the right will be data about that case. The SPSS Data Editor has tabs in the lower-left corner that let you work with your data in two ways.

Variable View is used to set up the data—names, variable labels, value labels, etc. The other tab, Data View, is used to actually enter the data. SPSS probably opened in the Data View mode, if not, click the Data View tab at the bottom left of the SPSS screen now. (See Figure 15-2.)

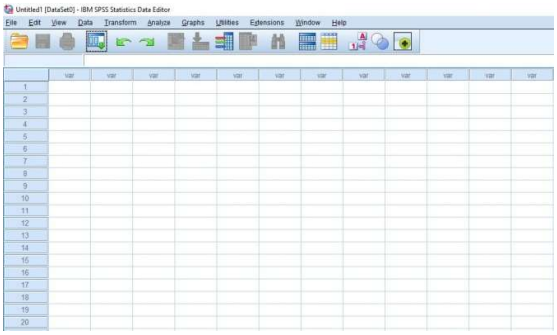


Figure 15-2

Entering Variable and Value Names and Labels:

In Data View, we will use the first column for the respondents' ID numbers, so type **001** into the first cell and press Enter and 1.00 will appear. (See Figure 15-3.) [3]

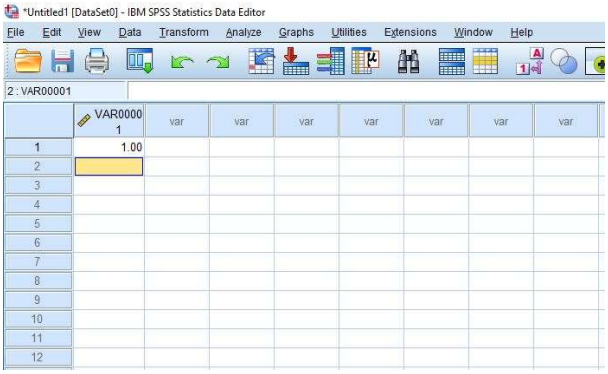


Figure 15-3

We will use the Variable View tab to assign variable names and longer variable labels as well as value labels that will make it easier to use the data for tables and charts. Click Variable View now and click the VAR0001 in the top left column. Type in id. (Press Enter and VAR00001 changes to our variable name, id.) Go back to Data View and notice that the first column is now titled id. (See Figure 15-4.)

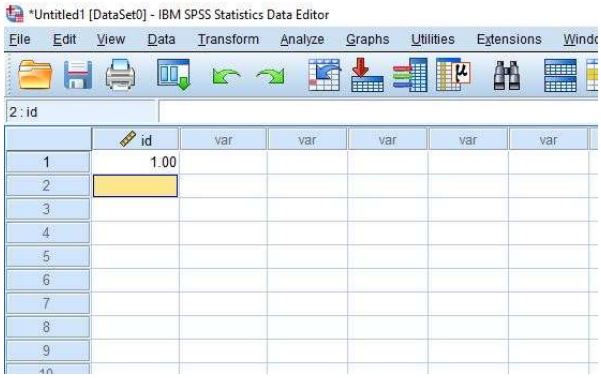


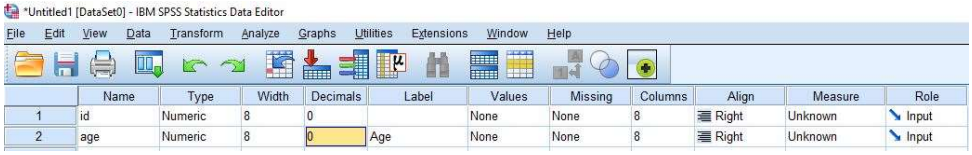
Figure 15-4

The second variable will be the student's age, so change back to Variable View and type **age** under name in the second row. Remember to press Enter after you type in each entry.

[3] It is wise to save your computer work early and often. You might want to save this file and call it something like **Data Entry Exercise 1**. Notice that SPSS saves it as a .sav file. This means it contains the data in the format for SPSS analysis.

SPSS makes some assumptions about data that might not be appropriate. In the fourth column, notice that it plans to use two decimal points even when the values for *age* are integers. To avoid these inappropriate decimal points, in the decimals column, click the cell for age and then click the blue box and click on the down arrow to change the value to 0. (Remember to do this whenever a numeral doesn't really refer to a numerical value.)

Since the short variable name usually doesn't give enough information about the variable, we want a longer or clearer variable label for our analysis. This one would be simple. To add a variable label to *age*, just tab over to the label column and type in **Age**. (See Figure 15-5.) Although it may not seem necessary to have a variable label for age, for most variables, a longer variable label is very useful.



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	8	0		None	None	8	Right	Unknown	Input
2	age	Numeric	8	0	Age	None	None	8	Right	Unknown	Input

Figure 15-5

Sometimes respondents don't answer a question, give two answers, or do something else so the data can't be used in the analysis. To have accurate results, missing or invalid data need to be indicated. Still in Variable View, tab over to Missing and click the blue box. This dialog box lets you specify up to three missing values. For our data, click Discrete and type **99** in the first text box. Leave the other boxes empty. Then Click OK. Now if someone doesn't answer a question, it will be marked as missing. (See Figure 15-6.)

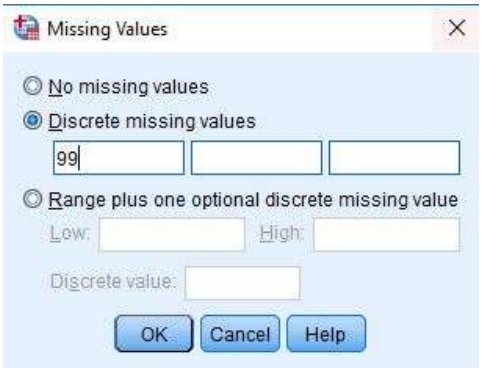
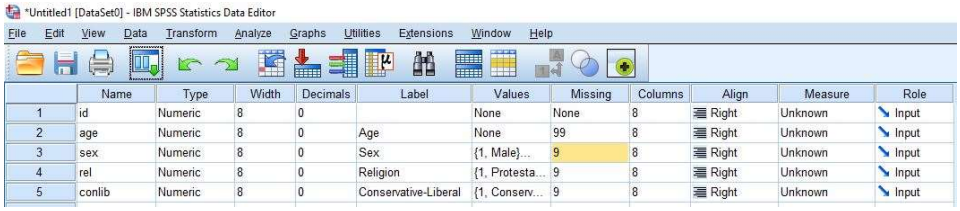


Figure 15-6

The third variable will be the sex of the respondent, so type **sex** in the third row under Name and Sex as the variable label. Since we're going to use the code 1 for males and 2 for females, we're going to need value labels in words for each category. Tab over to the cell under Values and click the little blue box to get the value labels menu. Type a **1** in the Value box and then **Male** in the Value Label box. Click Add and it shows that Value 1 will be Male. Type a **2** in the Value box, and type **Female** in the Value label space. Click Add and then click OK to save these. Now, SPSS knows that 1 and 2 in Sex are really male and female respectively. (See Figure 15-7.)

Figure 15-7

For this exercise, we are also using religion and conservative-liberal as variables. Add those variables in rows 4 and 5. Give each variable a name and label—*rel* gets Religion and *conlib* gets something like Conservative-Liberal as variable name and label. Then add value names and labels. Notice that *rel* has five possibilities—Protestant, Catholic, Jewish, other, and no religion. Go ahead and work out the value names and value labels. Make arrangements for missing values just as you did before. (You can refer to Appendix , Codebook for Student Questionnaire at the end of this lesson.) Remember to type variable labels and value labels exactly the way you would want them in a table when you do the analysis—often this is with the first letter of each important word capitalized. (Your Variable View might look like Figure 15-8.)



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	id	Numeric	8	0		None	None	8	Right	Unknown	Input
2	age	Numeric	8	0	Age	None	99	8	Right	Unknown	Input
3	sex	Numeric	8	0	Sex	{1, Male}...	9	8	Right	Unknown	Input
4	rel	Numeric	8	0	Religion	{1, Protesta...	9	8	Right	Unknown	Input
5	conlib	Numeric	8	0	Conservative-Liberal	{1, Conserv...	9	8	Right	Unknown	Input

Figure 15-8

Entering the Data:

Enter the codes for each variable using Data View[4]. Then check the accuracy of your data entry by scanning down each column looking for codes that would be impossible. For example, sex can have only three possibilities since male is 1, female is 2, and missing information is 9, so a 5 or 6 would be a mistake. Then check everything carefully. The best check is to have one person read the codes while another checks the entries on Data View.

[4] Some people, especially those who are used to working with spreadsheets, like to enter all the data in Data View before they set up the variable names, etc. In this example, we set up the variable names, etc., before we enter any data. (You'll have to figure out what works best for you.) You can also enter data from a spreadsheet like Excel.

Student Survey Questionnaire

(1) What is your age? _____

(2) Are you ____ male or ___ female?

(3) What is your religious preference?

___ Protestant ___Catholic ___ Jewish ___ Some other religion ___No religion

(4) Generally speaking, in politics, do you consider yourself as

___conservative ___ liberal __ or middle of the road

(5) What kind of marriage do you think is the more satisfying way of life?

___ One where the husband provides for the family and the wife takes care of the house and children

___ One where both the husband and wife have jobs and both take care of the house and children

Do you think it should be possible for a pregnant woman to obtain a legal abortion?

(6) If there is a strong chance of serious defect in the baby?

__Yes __ No ___Don't Know

(7) If she is married and does not want any more children?

__Yes __ No ___Don't Know

(8) If the woman's own health is seriously endangered by pregnancy?

__Yes __ No ___Don't Know

(9) If the family has a very low income and cannot afford any more children?

__Yes __ No ___Don't Know

(10) If she became pregnant as a result of rape?

__Yes __ No ___Don't Know

(11) If she is not married and does not want to marry the man?

__Yes __No __ Don't Know

(12) If the woman wants it for any reason

__Yes __ No ___Don't Know

Appendix: Codebook for Student Questionnaire

Missing Values	9 or 99
Age	Age at last birthday
Sex	1 = male, 2 = female
Religious Preference	1 = Protestant, 2 = Catholic, 3 = Jewish, 4 = Other, 5 = None
Political Orientation	1 = Conservative, 2 = Liberal, 3 = Middle of the road
Preferred Marriage	1 = Traditional, 2 = Shared
Abortion if Birth Defect	1= Yes, 2 = No, 3 = Don't Know
Abortion if No More Children	1= Yes, 2 = No, 3 = Don't Know
Abortion if Health Risk	1= Yes, 2 = No, 3 = Don't Know
Abortion if Poor	1= Yes, 2 = No, 3 = Don't Know
Abortion if Rape	1= Yes, 2 = No, 3 = Don't Know
Abortion if Not Married	1= Yes, 2 = No, 3 = Don't Know
Abortion For Any Reason	1= Yes, 2 = No, 3 = Don't Know

Appendix: Planning Matrix for Data-entry Exercise

	age	*Sex*	*rel*	*Conlib*	*mg*	*Abd*	*abn*	*Abh*	*abp*	*abr*	*abs*	*aba*
01	20	1	4	2	2	2	2	1	3	1	2	2
02	24	2	5	2	2	1	1	1	1	1	1	9
03	21	2	2	9	2	2	2	2	2	2	2	2
04	24	2	5	3	2	1	1	1	1	1	1	1
05	26	2	4	2	2	1	1	1	1	1	1	1
06	28	2	2	2	2	2	2	1	2	1	2	2
07	23	1	1	2	2	1	2	1	1	1	2	2
08	22	2	4	3	1	1	1	1	1	1	1	1
09	22	1	5	2	2	1	1	1	1	1	1	1
10	22	2	4	4	2	1	1	1	1	1	1	1
11	23	1	2	2	1	2	2	1	2	1	2	3
12	24	2	2	3	2	1	1	1	1	1	1	2
13	51	2	1	2	9	1	1	1	1	1	1	1
14	22	2	2	3	2	1	1	1	1	1	1	1
15	21	2	4	3	2	1	1	1	1	1	1	1
16	37	1	1	3	2	1	2	1	2	1	2	2
17	22	2	4	2	2	1	1	1	1	1	2	2
18	22	2	3	3	2	1	2	1	2	1	2	2
19	22	2	4	3	2	3	2	1	2	1	1	1
20	30	2	5	2	2	1	1	1	1	1	1	1
21	25	2	5	2	2	1	1	1	1	1	1	1
22	23	1	2	2	2	1	1	1	1	1	1	1
23	21	1	1	2	1	1	1	2	1	2	1	1

15.6 SUMMARY

In modern research, automated data analysis has become essential due to the increasing volume and complexity of data. Software tools such as SPSS simplify the process of organising, analysing, and interpreting data by minimising manual work and reducing errors. SPSS is widely used across academic and professional disciplines because of its user-friendly interface and powerful statistical capabilities.

This lesson introduces the basic steps in creating a new data file in SPSS. It covers assigning variable names, setting variable and value labels, coding responses, and defining missing values. Learners also become familiar with the two main working views in SPSS: the Variable View for data setup and the *Data View* for data entry. Careful planning and accuracy checks are emphasised to ensure reliable data analysis results.

15.7 TECHNICAL TERMS:

1. **Automated Data Analysis:** The use of computer-based software to analyse data with minimal human intervention, improving efficiency and reducing the chances of manual error.
2. **SPSS (Statistical Package for the Social Sciences):** A widely used statistical software application designed to manage, analyse, and present research data across various academic and professional fields.
3. **Variable:** A characteristic or attribute that can take different values among individuals in a dataset. Each survey question typically becomes a variable in SPSS.
4. **Variable Name**: A short, computer-friendly identifier assigned to each variable in SPSS (e.g., *age*, *sex*, *rel*). It must not contain spaces or special symbols.
5. **Variable Label:** A longer descriptive text explaining the meaning of a variable (e.g., "Age of Respondent"). Used for clarity when generating tables and reports.
6. **Value Labels**: Text descriptions assigned to numerical codes representing categories of responses (e.g., 1 = Male, 2 = Female), making outputs easier to interpret.
7. **Coding:** The process of converting qualitative or categorical responses into numerical values so they can be entered and analysed in SPSS.
8. **Missing Value Code:** An exceptional numeric value used to indicate that data is absent, unclear, or not usable (commonly coded as *9* or *99*) to avoid distortion in analysis.
9. **Data View:** The worksheet-like SPSS screen where actual case-by-case responses are entered. Rows represent respondents, and columns represent variables.
10. **Variable View:** The SPSS screen used to define the properties of each variable, such as variable name, label, coding, type, decimals, and missing values.
11. **Data File:** A structured electronic file in SPSS that contains all the variables and coded responses for each respondent in the study.

15.8 SELF-ASSESSMENT QUESTIONS:

1. Explain the need for automated data analysis in modern research.
2. Distinguish between *Variable View* and *Data View* in SPSS.
3. What is the purpose of assigning value labels in SPSS?
4. Describe the role of missing value codes in data analysis.
5. Discuss the importance and advantages of using SPSS for data analysis in research. Provide examples of how SPSS supports data management and decision-making.
6. Explain the step-by-step procedure for creating a new data file in SPSS, including assigning variable names, coding responses, and entering data.
7. A researcher collects data on gender, age, religious preference, and political orientation. Describe how these variables should be coded and entered into SPSS to ensure clarity and accuracy in analysis.

15.9 SUGGESTED READINGS

1. IBM Corp. (2023). *SPSS Statistics User Guide*. Armonk, NY: IBM Corporation.
2. IBM Corp. (2023). *SPSS Statistics: Base System Documentation*. Armonk, NY: IBM Corporation.
3. IBM Corp. (2023). *SPSS Statistics for Windows, Version 29.0* [Computer software]. Armonk, NY: IBM Corporation.

Dr. G. MALATHI

LESSON -16

UNIVARIATE ANALYSIS

OBJECTIVES OF THE LESSON

After studying this lesson, students will be able to:

1. Understand the concept and purpose of univariate analysis.
2. Identify appropriate SPSS procedures for analysing single variables (Frequencies, Descriptives, Explore).
3. Generate, read, and interpret frequency tables and graphical outputs.
4. Select suitable descriptive statistics based on the level of measurement (nominal, ordinal, interval, ratio).
5. Recognise the importance of univariate analysis as the first step in data analysis.
6. Detect and correct data entry errors and missing data issues through univariate inspection.

STRUCTURE OF THE LESSON

16.1 INTRODUCTION TO UNIVARIATE ANALYSIS

Univariate analysis, which looks at individual variables, is usually the initial step when analysing data for the first time. There are several reasons why it is the starting point, and most of these will be discussed at the end of this lesson, but for now, let us focus on the “basic” results. If we are analysing a survey, we are interested in how many people said “Yes” or “No,” or how many 'Agreed' or 'Disagreed' with a statement. We aren't conducting a traditional hypothesis test with an independent and dependent variable; we're simply observing the distribution of responses.

The SPSS tools for analysing single variables include the following procedures: Frequencies, Descriptives, and Explore, all found under the Analyse menu.

This lesson will use the GSS16A file (This is a subset of the 2016 General Social Survey, you can download these data files from the web by going to SPSS Statistics for Windows 25: A Basic Tutorial), so start SPSS and bring the file into the Data Editor. To begin the process, start SPSS and open the GSS16A data file. Under the Analyse menu, select Descriptive Statistics and the procedure you want: Frequencies, Descriptives, Explore, or Crosstabs.

Frequencies:

Generally, a frequency distribution is used to examine detailed information about nominal and ordinal (categorical) data, describing the results. Categorical data applies to variables such as gender, where males are coded as "1" and females as "2." Frequency options include a table displaying counts and percentages, statistics such as percentile values, measures of central tendency, dispersion, and distribution, and charts including bar charts and histograms. The steps for using the Frequencies procedure are to click the Analyse menu, choose Descriptive Statistics, then from the submenu, select Frequencies and pick your variables for analysis. You can then choose statistics options, chart options, and format options, and have SPSS calculate your request**.**

For this example, we will examine attitudes on the abortion issue. The 2016 General Social Survey, GSS16A(Questionnaire used in Lesson 15), includes the variable ABORTION--FOR ANY REASON. We will consider this variable for our initial analysis investigation.

Choosing Frequencies Procedure:

From the Analyze menu, highlight Descriptive Statistics, Figure 16-1, then move your mouse across to the sub menu and click on Frequencies.

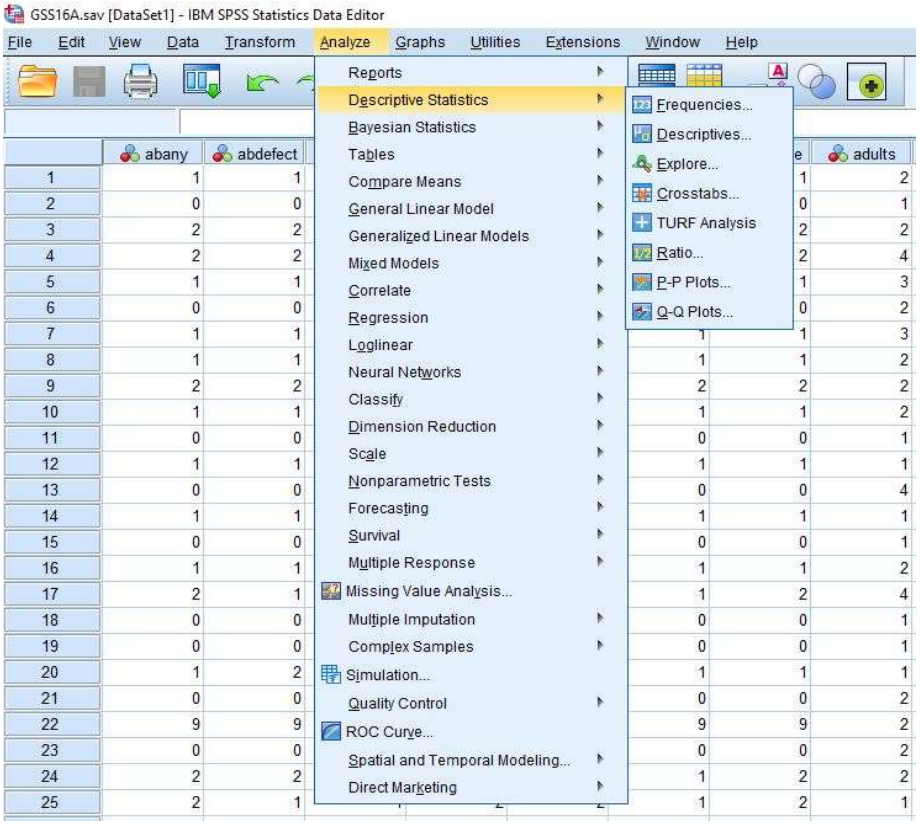


Figure 16-1

A Dialog box, Figure 16-2, will appear providing a scrollable list of the variables on the left, a Variable(s) choice box, and buttons for Statistics, Charts and Format options.[1]

Figure 16-2

Selecting Variables for Analysis:

First select your variable from the main Frequencies Dialog box, Figure 16-2, by clicking the variable name on the left side. (Use the scroll bar if you do not see the variable you want.) In this case *abany* is the first variable and will be selected (i.e., highlighted). Thus, you need not click on it.

Click the arrow on the right of the Variable List box, Figure 16-2, to move *abany* into the Variable(s) box. All variables selected for this box will be included in any procedures you decide to run. We could click OK to obtain a frequency and percentage distribution of the variables. In most cases we would continue and choose one or more statistics.

Choosing Statistics for Variables:

Click the Statistics button, right top of Figure 16-2, and a Dialog box of statistical choices will appear, Figure 16-3.

This variable, *abany*, is a nominal (category) variable so click only the Mode box within the central Tendency choices. See Figure 16-3.

[1] If you want to change the display to labels or know more information about a variable, the label, codes, etc., place the mouse pointer on the variable name in the Variable List, right click the mouse button.

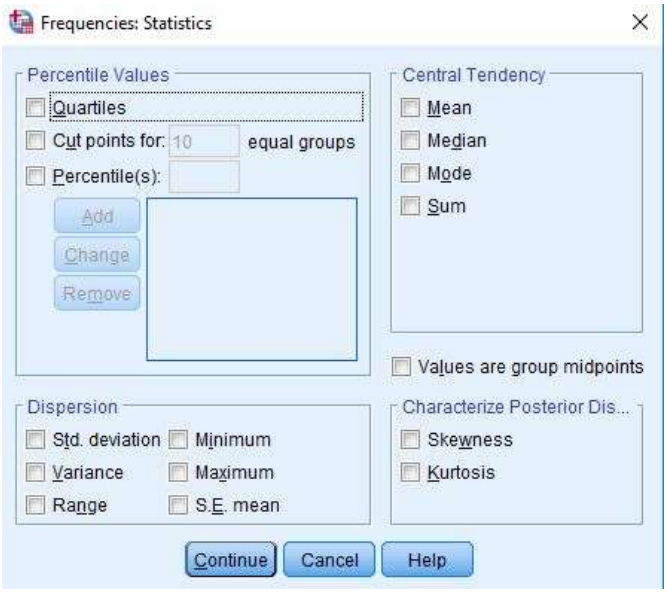


Figure 16-3

After clicking the Mode box, click the Continue button, bottom left, and we return to the main Frequencies dialog box, Figure 16-2.

You can now click OK, and SPSS will calculate and display the frequency and percentage distribution (click OK if you wish), but typically, we will proceed to include options for charts and review the Options settings. If you click OK, simply go to the Analysis menu, select Descriptive Statistics, then Frequencies from the submenu, and you will return to this screen with your variable and statistics already chosen.

Choosing Charts for Variables:

On the main frequencies window, click the Charts button, Figure 16-2, and a dialogue box of chart choices, Figure 16-4, will appear.

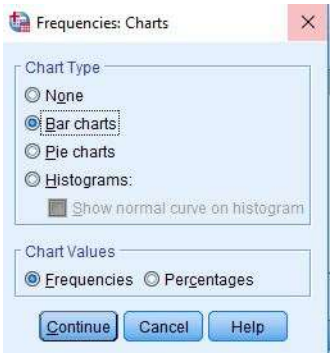


Figure 16-4

Click Bar Chart, as I have done, since this is a categorical variable, then click Continue to return to the main Frequencies window box. If you have a continuous variable, choose Histograms and the With Normal Curve option would be available. Choose the With Normal Curve option to have a normal curve drawn over the distribution so that you can visually see how close the distribution is to normal. Note: Frequencies is automatically chosen for chart values but if desired you could change that to Percentages, bottom Figure 16-4.

Now click OK on the main frequencies dialog box and SPSS will calculate and present a frequency and percent distribution with our chosen format, statistics, and chart. (Note: We could look to see if additional choices should be made by clicking the Format button. In this case we don't need to do this because all the Format defaults are appropriate since we are looking at one variable.)

Looking at Output from Frequencies:

We will now take a brief look at our output from the SPSS frequencies procedure. (Patience, processing time for SPSS to perform the analysis in the steps above will depend on the size of the data set, the amount of work you are asking SPSS to do and the CPU speed of your computer.) The output outline, left side, and the output, right side, will appear when SPSS has completed its computations. Either scroll down to the chart in the right window, or click the Bar Chart icon in the outline pane to the left of the output in Figure 16.5.

abany ABORTION IF WOMAN WANTS FOR ANY REASON

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 YES	819	28.6	44.9	44.9
	2 NO	1006	35.1	55.1	100.0
	Total	1825	63.7	100.0	
Missing	0 IAP	964	33.6		
	8 DK	49	1.7		
	9 NA	29	1.0		
	Total	1042	36.3		
Total		2867	100.0		

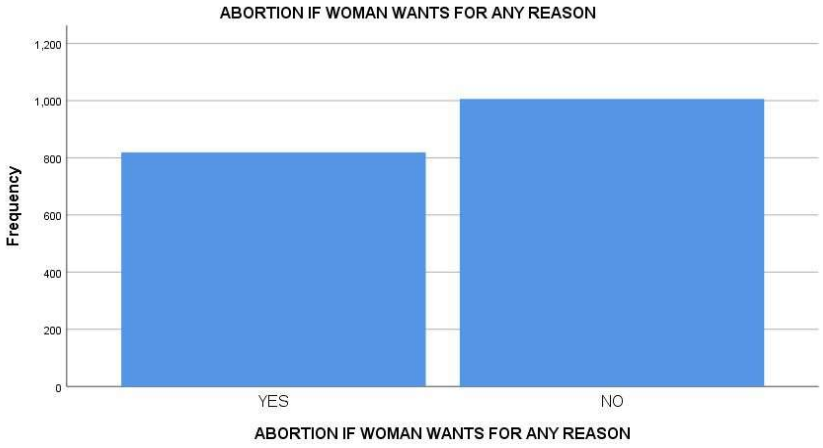


Figure 16.5

Interpreting the Chart:

We now see the chart, Figure 16-6. The graphic is a bar chart with the categories at the bottom, the X-axis, the frequency scale on the left, and the Y-axis. The variable label ABORTION IF WOMAN WANTS FOR ANY REASON is displayed at the top of the chart. We see from the frequency distribution that there are more "no," 35.1%, answers than "yes," 28.6% answers (see Figure 16-7), when respondents were asked if a woman should be able to get an abortion for any reason. A much smaller number, which does not appear on this chart, 1.7% (see Figure 16-7), selected "don't know," "DK." If a chart were the only data presented for this variable in a report, you should look at the frequency output and report the total responses and/or percentages of YES, NO and DK answers. You should also label the chart with frequencies and/or percentages. There are a lot of possibilities for enhancing this chart within SPSS.

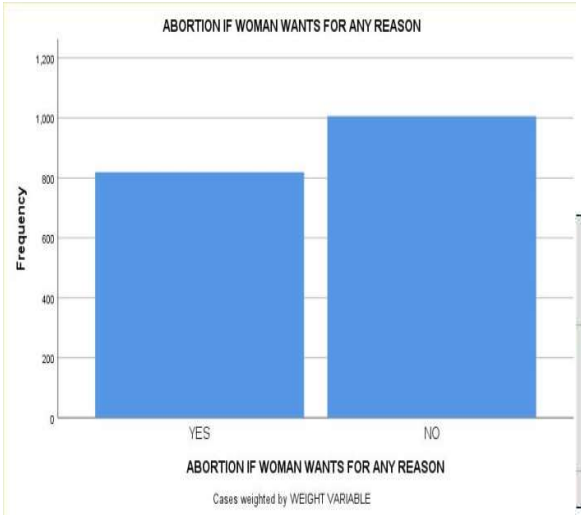


Figure 16-6

abany ABORTION IF WOMAN WANTS FOR ANY REASON

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 YES	819	28.6	44.9	44.9
	2 NO	1006	35.1	55.1	100.0
	Total	1825	63.7	100.0	
Missing	0 IAP	964	33.6		
	8 DK	49	1.7		
	9 NA	29	1.0		
	Total	1042	36.3		
Total		2867	100.0		

Figure 16-7

If we choose to copy our chart to a word processor program for a report, first select the chart by clicking the mouse on the bar chart. A box with handles will appear around the chart. Select Copy Special from the Edit menu and choose the format that you want to use (JPG is a good choice). Start your word processing document, click the mouse where you want the chart to appear then choose Paste Special from the down arrow on Paste. Choose an option in the paste special dialog box that appears and click OK to paste the chart into your document.

Interpreting Frequency Output:

To view the frequency distribution, move the scroll bar on the right of our output window to view the table. Another way is to click the Frequencies icon in the Outline box to the left of the output window. To view a large table, you may want to click on the Maximise Arrow in the upper right corner of the SPSS Output Navigator window to enlarge the output window. Use the scroll bars to display different parts of a large table. The most relevant part of the frequency distribution for *abany* is in Figure 16-7.

We can now see some of the specifics of the SPSS frequencies output for the variable *abany*. At the top is the variable label ABORTION IF WOMEN WANT FOR ANY REASON. The central part of the display shows the value labels (YES, NO, Total), and the missing categories, IAP (Inapplicable), DK (Don't Know), and NA (Not Answered), Total and the Frequency, Percent, Valid Percent, Cumulative Percent (the cumulative % for values as they increase in size), for each classification of the variable. The "Total" frequency and percent is listed at the bottom of the table. When asked if a woman should be able to have an abortion for any reason, 35.1% responded no. DK, don't know, was chosen by 1.7% and 1.0% were NA [Not Answered]. The 33.6% "IAP [Inapplicable], was that portion of the sample that was not asked this question. In a written paper, you should state that the "Valid Percent" excludes the "missing" answers.

It is important to use these concepts correctly, so a review at this point is appropriate. A Variable name is the short name you gave to each variable, or question in a survey. The table below is designed to help you keep these separate.

Variable Name	**Variable Label**	**Value**	**Value Label**
SEX	Respondent's gender.	1 or 2	(1) Male, (2) Female
AGE	Respondent's age at last birthday.	18, 19, 20, 21… 89, 98, 99	None needed
AGED	Should aged live with their children.	1, 2, 3, 0, 8, 9	(1) A good idea, (2) Depends, (3) A bad idea (0) IAP [Inapplicable], (8) DK [Don't Know], (9) NA [Not Answered]

Understanding these allows you to intelligently customize SPSS for Windows so that it is easier for you to use. You can set SPSS so that you can see the variable names when you scroll through a listing of variables, or so that you can see the variable labels as you scroll through the listing. You can set SPSS so that you get only the values, only the labels, or both in the output. Below are two examples of Frequencies Dialog box.

Figure 16-8 shows the listing as variable labels. This is the default setting when SPSS for Windows is installed. This example has the cursor on the variable label ABORTION IF WOMAN WANTS FOR ANY REASON (is displayed). You can change the listing, however, so that you see only variable names, as in Figure 16-9. Changing this is a matter of personal taste. This Lesson uses variable names, Figure 16-9.

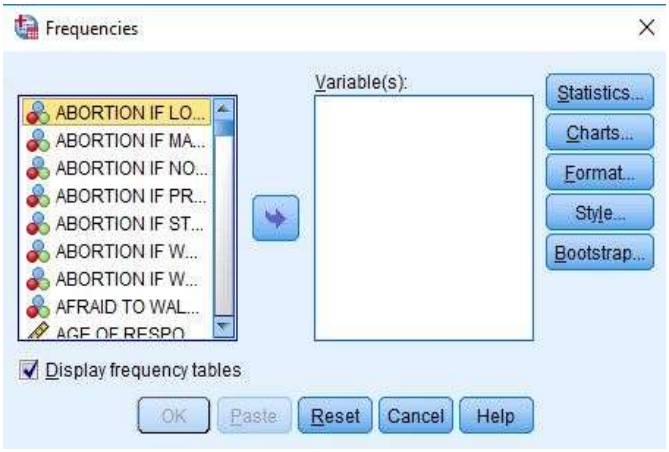


Figure 16-8

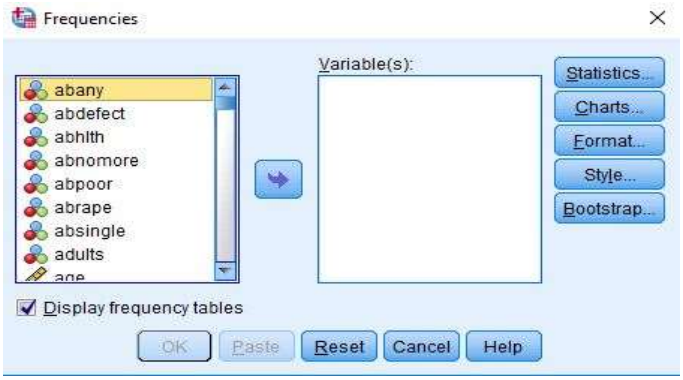


Figure 16-9

You can change the display listing when running a procedure by right clicking on the list in the left box of a procedure and choosing a display format, Figure 16-9. For this Lesson we choose Display Names and Alphabetical so that variable names will be displayed alphabetically as in Figure 16-9.

Changing the display option for the Variable Selection dialog box, as well as other display formats, can be done for all dialog choices *before* running a procedure. After starting SPSS, to set the display option, click Edit then choose Options. The General tab on the Options dialog box will appear, Figure 16-10. Under Variable Lists section, top right quadrant, click your choices, again we choose Display Names and Alphabetical, then click OK.

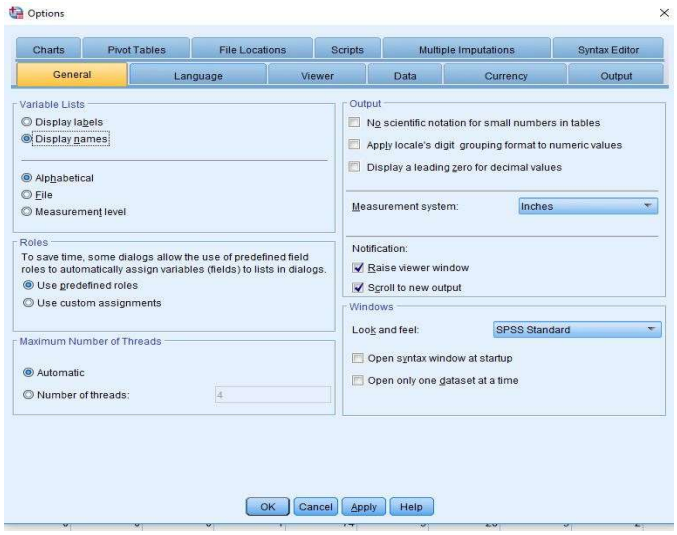


Figure 16-10

Displaying Values, Value Labels or Both in Your Output:

One other option you might want to use is in the table format for your SPSS output. You can choose to have displayed variable labels, values (e.g., 1, 2, 3, etc.), value labels (YES, No, DK, etc.) or both values and labels (1 YES, 2 NO, 3 DK). To make these choices, click the Edit menu and choose Options, then click the Output tab, click your choices on the options dialog box. My choices are seen in Figure 16-11. The output resulting from my choices for a Frequencies procedure is Figure 16-12.

abany ABORTION IF WOMAN WANTS FOR ANY REASON

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1 YES	819	28.6	44.9	44.9
	2 NO	1006	35.1	55.1	100.0
	Total	1825	63.7	100.0	
Missing	0 IAP	964	33.6		
	8 DK	49	1.7		
	9 NA	29	1.0		
	Total	1042	36.3		
Total		2867	100.0		

Figure 16-11 **Figure 16-12**

Descriptives

Descriptives (Analysis, Descriptive Statistics, Descriptives, Figure 16-13) is used to obtain summary information about the distribution, variability, and central tendency of continuous variables. Possibilities for Descriptives include mean, sum, standard deviation,

variance, range, minimum, maximum, S.E. mean, kurtosis and skewness. For this example, we will examine the distributions of age and education for the General Social Survey sample. Since both these variables were measured at the interval/ratio level, different statistics from our previous example will be used.

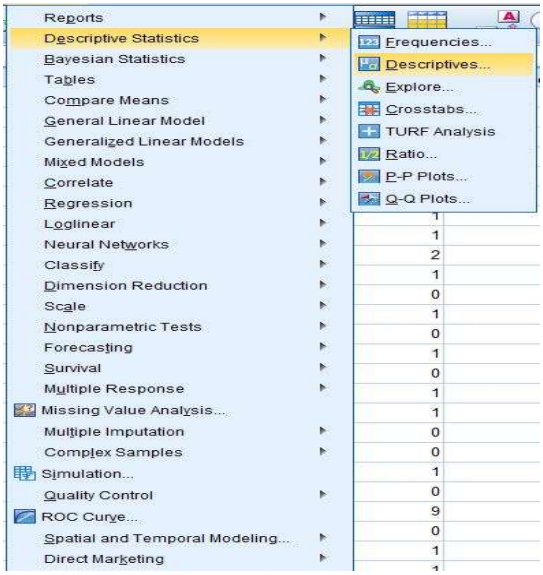


Figure 16-13

Choosing Descriptive Procedure:

First, click the Analyse menu and select Descriptive Statistics, then move across to the sub-menu and select Descriptives (see Figure 16-13). The Variable Choice dialogue box will appear (see Figure 16-14).

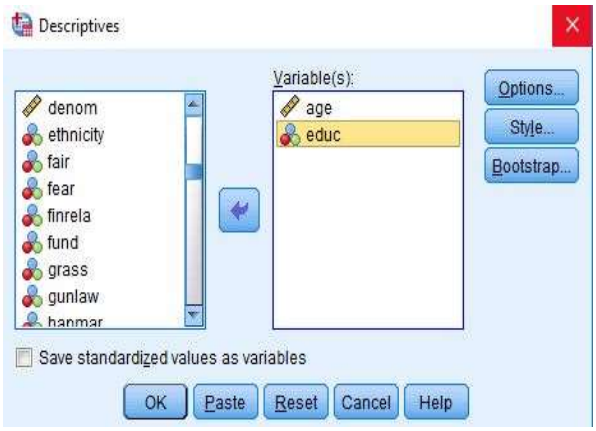


Figure 16-14

Selecting Variables for Analysis:

First click on *age*, the variable name for AGE OF RESPONDENT. Click the select arrow in the middle and SPSS will place *age* in the Variable(s) box. Follow the same steps to choose *educ*, the variable name for HIGHEST YEAR OF SCHOOL COMPLETED. The dialog box should look like Figure 16-14.

We could click OK and obtain a frequency and percentage distribution, but we will click the Options button and decide on statistics for our output. The Options dialog box, Figure 16-15, will open.

Figure 16-15

Since these variables are interval/ratio measures, choose: Mean, Std. deviation, Minimum and Maximum. We will leave the defaults for the Distribution and Display Order.

Next, click the Continue button to return to the main Descriptives dialog box, (Figure 16-14). Click OK in the main Descriptives dialogue box, and SPSS will calculate and display the output seen in Figure 16-16.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
age AGE OF RESPONDENT	2855	18	89	47.56	17.589
educ HIGHEST YEAR OF SCHOOL COMPLETED	2859	0	20	13.68	2.956
Valid N (listwise)	2848				

Figure 16-16

Interpretation of the Descriptives Output:

In the interpretation of Figure 16-16, AGE OF RESPONDENT has a mean of 47.56 and a standard deviation of 17.589. The youngest respondent was 18 and the oldest was 89. Look at your SPSS output for HIGHEST YEAR OF SCHOOL COMPLETED. It has a mean of 13.68 (a little more than 1 year beyond high school) and a standard deviation of 2.956. Some respondents indicated no "0" years of school completed. The most education reported was 20 years.

Explore

Explore is primarily used to visually examine the central tendency and distributional characteristics of continuous variables. Explore statistics, including M-estimators, outliers, and percentiles. Grouped frequency tables and displays, as well as stem-and-leaf and box plots, are available. Explore will aid in checking assumptions using Normality plots and the Levene test for Spread vs. Level.

Choosing the Explore Procedure:

From the Analyze menu choose Descriptive Statistics, drag to the sub menu and select Explore.

Selecting Variables:

As in the other procedures, find and click the variable you want to explore, and then click the select arrow to include your variable in the Dependent List box. Choose the variable *educ* and move into the Dependent List box. The dialog box should look like Figure 16-17.

Figure 16-17

Selecting Displays:

In the Display box on the bottom left, you may choose either Both, Statistics, or Plots. We left the default selection, Both, to display statistics and plots.

Selecting Statistics:

Click the Statistics button and the Explore: Statistics dialog box will open, Figure 16-18.

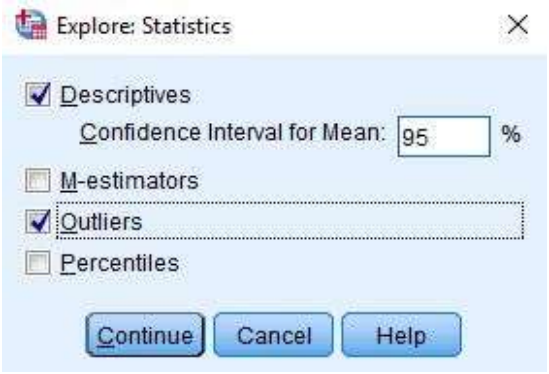


Figure 16-18

Leave checked the Default box for Confidence Interval for the Mean 95%, and click the **Outliers** box so we can look at the extreme observations for our variable. Click Continue to return to the main explore dialog window.

Selecting Plots:

Click the Plots button on the main Explore Dialogue box, Figure 16-17, and the Explore: Plots dialogue box, Figure 16-19, will open.

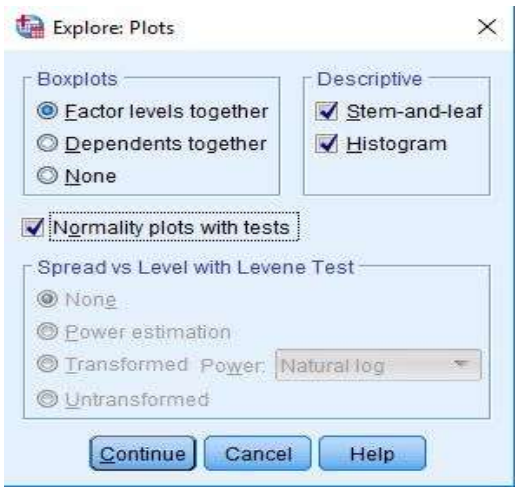


Figure 16-19

Click Stem-and-leaf and Histogram in the Descriptive box. Click on Normality Plots with Test so we can see how close the distribution of this variable is to normal. Leave the default for Spread vs. Level with Levene Test. Click Continue to return to the main explore dialog box, Figure 16-17.

Selecting Options:

Click the Options button in the main explore dialog box, Figure 16-17, and the Explore: Options dialog box, Figure 16-20, will be displayed.

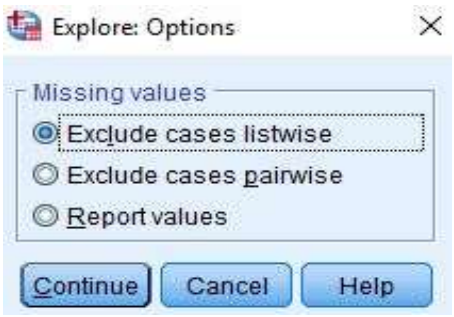


Figure 16-20

No changes are needed here since the default of Exclude cases listwise is appropriate. Now click Continue to return to the main Explore dialog box, Figure 16-17. Click OK in the main Explore dialog box and SPSS will perform the chosen tasks and display the data in the SPSS Output.

Interpretation of Explore Output:

Use the scroll bar to view any part of the output. The first part of the output is the Case Processing Summary, Figure 16-21.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
educ HIGHEST YEAR OF SCHOOL COMPLETED	2859	99.7%	8	0.3%	2867	100.0%

Figure 16-21

We see that 2859 (99.7%) of our respondents answered this question. Only eight, 0.3% of the sample, were missing. The GSS often uses a split sample where not all respondents are asked the same questions. This is a question where all respondents were asked, so the total sample size was 2859.

The Descriptive statistics output should resemble Figure 16-22.

Descriptives

			Statistic	Std. Error
educ HIGHEST YEAR OF SCHOOL COMPLETED	Mean		13.68	.055
	95% Confidence Interval for Mean	Lower Bound	13.57	
		Upper Bound	13.79	
	5% Trimmed Mean		13.72	
	Median		13.00	
	Variance		8.738	
	Std. Deviation		2.956	
	Minimum		0	
	Maximum		20	
	Range		20	
	Interquartile Range		4	
	Skewness		-.193	.046
	Kurtosis		.872	.092

Figure 16-22

We can observe all the typical descriptive statistics in this output: mean (13.68), lower bound (13.57), and upper bound (13.79) for a 95% confidence interval of the mean. (In polling terminology, this indicates that we are 95% confident the population mean lies between 13.57 and 13.79). Also shown are the median (13.00), variance (8.738), standard deviation (2.956), minimum (0), maximum (20), range (20), interquartile range (4.00), skewness (−.193), and kurtosis (0.872). A narrative describing the education level of the sample respondents might be similar to the following:

Our sample from the 2014 General Social Survey indicates that the average education for those over 18 was 13.68 years, with 95% confidence that the population average would fall between 13.57 and 13.79 years. The minimum years of education reported was 0, and the maximum was 20. The median, the middle point where 50% of the population falls below and 50% above, was 13.00.

The extreme values are shown in Figure 16-23. This figure displays the five highest and five lowest values for our variable. More than five respondents reported their years of education as 20. At the lower end, two respondents indicated they had zero years of schooling. The Test of Normality is shown next (see Figure 16-24). This shows that this distribution is not significantly different from the expected normal distribution. This is a pretty stringent test; most researchers would not require the distribution to be this close to normality.

Extreme Values

			Case Number	Value
educ HIGHEST YEAR OF SCHOOL COMPLETED	Highest	1	44	20
		2	59	20
		3	77	20
		4	92	20
		5	99	20[a]
	Lowest	1	1520	0
		2	714	0
		3	2385	1
		4	1435	1
		5	190	1

a. Only a partial list of cases with the value 20 are shown in the table of upper extremes.

Figure 16-23

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
educ HIGHEST YEAR OF SCHOOL COMPLETED	.151	2859	.000

a. Lilliefors Significance Correction

Figure 16-24

The histogram, Figure 16-25, shows a rough bell-shaped distribution. SPSS divided our distribution into twenty-one groups with a width of one year of education for each group.

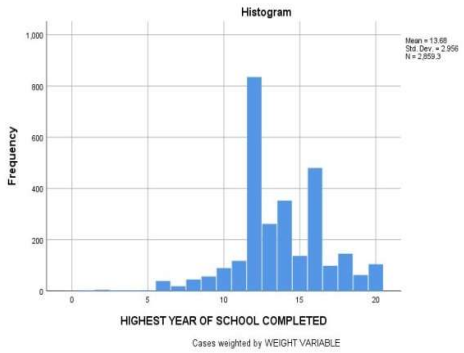


Figure 16-25

The largest group has just over 800 cases, based on a visual estimate. The smallest group has very few cases; we know there were only two respondents reporting 0 years of education from our Extreme Values table. The statistics on the histogram indicate that the standard deviation is 2.956, with a mean of 13.68, for a total N of 2859. The Stem-and-Leaf plot follows. Figure 16-26 again shows a distribution that is close to, but not quite, normal, with outliers at the ends and a higher number of observations above the mode. We observed this in our earlier output. Additionally, we notice that 12, high school; 14, junior college; and 16, college are clear stopping points.

```
Frequency    Stem &  Leaf

  16.40 Extremes    (=<5.0)
  38.83        6 .  0000
    .00        6 .
  18.57        7 .  00
    .00        7 .
  44.67        8 .  00000
    .00        8 .
  56.79        9 .  000000
    .00        9 .
  89.33       10 .  0000000000
    .00       10 .
 117.81       11 .  0000000000000
    .00       11 .
 834.74       12 .  000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
    .00       12 .
 261.99       13 .  00000000000000000000000000000
    .00       13 .
 352.53       14 .  000000000000000000000000000000000000000
    .00       14 .
 136.89       15 .  000000000000000
    .00       15 .
 480.17       16 .  00000000000000000000000000000000000000000000000000000
    .00       16 .
  98.34       17 .  00000000000
    .00       17 .
 145.72       18 .  0000000000000000
    .00       18 .
  62.27       19 .  0000000
    .00       19 .
 104.29       20 .  000000000000

Stem width:   1
Each leaf:        9 case(s)
```

Figure 16-26

Interpretation of the Q-Q Plot of Education:

Continue scrolling down the SPSS Output Navigator to the Normal Q-Q Plot of HIGHEST YEAR OF SCHOOL COMPLETED (see Figure 16-27).

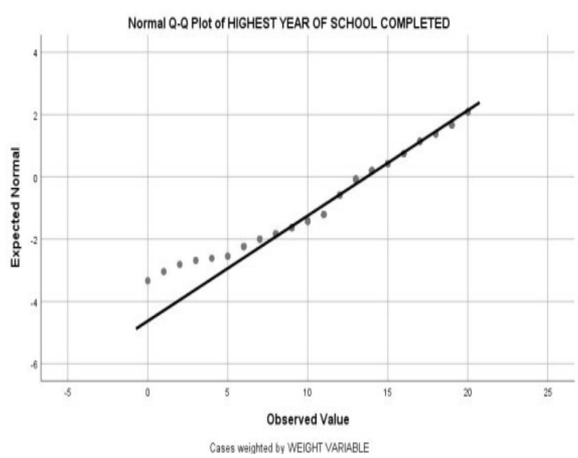


Figure 16-27

A Q-Q plot charts observed values against a known distribution, in this case a normal distribution. If our distribution is normal, the plot would have observations distributed closely around the straight line. In Figure 16-27, the expected normal distribution is the straight line and the line of little boxes is the observed values from our data. Our plot shows the distribution deviates somewhat from normality at the low end. The high end of the distribution is pretty much normal.

The Detrended Normal Q-Q plot shows the differences between the observed and expected values of a normal distribution. If the distribution is normal, the points should cluster in a horizontal band around zero with no pattern. Figure 16-28, of HIGHEST YEAR OF SCHOOL COMPLETED, indicates some deviation from normal, especially at the lower end. Our overall conclusion is that this distribution is not normal. Most researchers would consider this close enough to treat as a normal distribution.

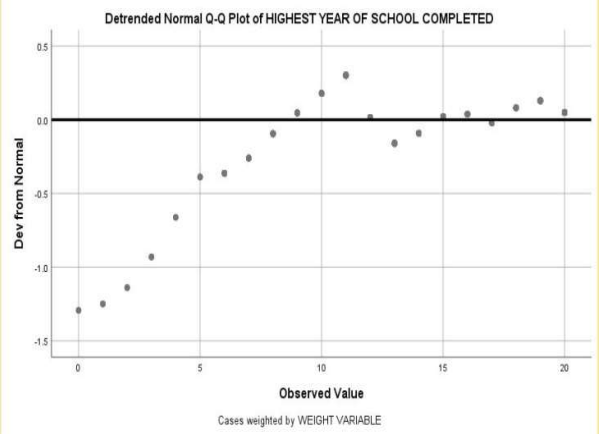


Figure 16-28

Interpretation of the Boxplot:

In the SPSS Output, scroll to the boxplot of HIGHEST YEAR OF SCHOOL COMPLETED.

Once again the major part of our distribution deviates from normal. There are significant outliers, the cases beyond the lower line of our boxplot. Our outliers are at the lowest end of the distribution, people with little or no education.

Conclusion

When performing univariate analysis, the level of measurement and the resulting distribution determine the appropriate analysis and influence subsequent multivariate analysis with the studied variables. The specific output from SPSS used in a report is selected to clearly present the distribution and central tendencies of the analysed variables. Sometimes, you report a particular output to facilitate comparison with other studies. In any case, choose the minimal output that effectively achieves this purpose. Do not report every SPSS output you obtained.

16.4 SUMMARY

Univariate analysis focuses on analysing one variable at a time and is generally the first step in data analysis. It helps researchers understand the distribution, central tendency, dispersion, and shape of the variable. In SPSS, univariate analysis can be conducted using Frequencies, Descriptives, and Explore procedures. Frequencies are primarily used for categorical variables to display counts and percentages, along with simple statistics and charts. Descriptives are used for continuous variables to obtain key summary statistics such as mean, standard deviation, minimum, and maximum. Explore provides both numerical and graphical summaries and assists in examining outliers and normality assumptions. Univariate analysis also helps identify data entry errors, missing values, and the need to regroup categories before proceeding to more advanced statistical methods.

16.5 TECHNICAL TERMS

1. **Univariate Analysis:** Refers to the examination of a single variable at a time to understand its basic characteristics, such as distribution, central tendency, and variability.
2. **Frequency Distribution:** It is a summary showing how often each value or category of a variable appears in the dataset.
3. **Central Tendency:** It refers to the measures that describe the centre or typical value of a distribution. These commonly include the mean, median, and mode.
4. **Dispersion:** It indicates how spread out the data values are, and is measured using the range, variance, and standard deviation.
5. **Bar Chart:** It is a graphical representation used for categorical variables, where the height of each bar indicates the frequency or percentage of each category.
6. **Histogram**: It is a graphical display used for continuous variables, showing the frequency distribution across intervals or ranges of values.
7. **Normal Distribution:** It is a symmetrical, bell-shaped distribution in which most values cluster around the mean and the probabilities for values further from the mean taper off equally on both sides.
8. **Outliers**: Extreme data values that differ significantly from other observations and may indicate unusual cases, errors, or special conditions.
9. **Stem-and-Leaf Plot:** It is a method of displaying numerical data that helps in quickly assessing the shape of the distribution while preserving exact data values.
10. **Boxplot**: It is a graphical summary of data showing the median, quartiles, and potential outliers, helping to visualize the spread and skewness of the data.
11. **Q-Q Plot (Quantile-Quantile Plot):** It is a graph used to compare the observed distribution of a variable with the expected normal distribution to assess normality.
12. **Missing Values:** Data points that are not recorded, incorrectly entered, or not applicable. These must be identified and handled before analysis.

16.6 SELF-ASSESSMENT QUESTIONS:

1. Define univariate analysis and explain its importance in data analysis.
2. Differentiate between Frequencies and Descriptives procedures in SPSS.
3. What is the significance of checking missing values during univariate analysis?
4. Explain the purpose of a histogram in univariate analysis.
5. Describe the steps to perform the Frequencies procedure in SPSS for a categorical variable. Illustrate how to interpret the output.
6. Explain how the Explore procedure helps in understanding distribution and identifying outliers. Support your explanation with interpretation guidelines for boxplots and Q-Q plots.

16.7 SUGGESTED READINGS

1. General Social Survey (GSS) Data Documentation.
2. IBM Corp. (2020). *SPSS Statistics User Guide.*
3. Field, A. (2018). *Discovering Statistics Using SPSS Statistics*. Sage Publications.
4. Pallant, J. (2020). *SPSS Survival Manual*. McGraw-Hill Education.

Dr. G. MALATHI

LESSON-17

CROSS TABULATIONS

OBJECTIVES OF THE LESSON

After studying this lesson, students will be able to:

1. Construct and interpret cross-tabulations (crosstabs) to analyse relationships between categorical variables in SPSS.
2. Apply Chi-Square and measures of association (Phi, Cramer's V, Kendall's tau c) to evaluate significance and strength of relationships.
3. Compare group means using the Independent-Samples t test, Paired-Samples t test, and the Means procedure.
4. Perform and interpret One-Way ANOVA to determine whether mean differences exist across three or more groups.

STRUCTURE OF THE LESSON

17.1 INTRODUCTION TO CROSS TABULATION

In this lesson, we'll look at how SPSS can be used to create contingency tables, sometimes called cross tabulations (or crosstabs), bivariate, or two-variable tables. A contingency table helps us look at whether the value of one variable is associated with, or "contingent" upon, that of another. It is most useful when each variable contains only a few categories. Usually, though not always, such variables will be nominal or ordinal. Some techniques for examining relationships among interval or ratio variables are presented in later lessons.

To make it easier to follow the instructions in this Lesson , it is recommended that you set certain options in SPSS in the same way that we have. First, click on Edit in the menu bar, then on Options, and General. Under Variable Lists, click on Display names, and Alphabetical. These choices will ensure that the variables in dialog boxes will look like they do in our examples (see Figure 17-1).

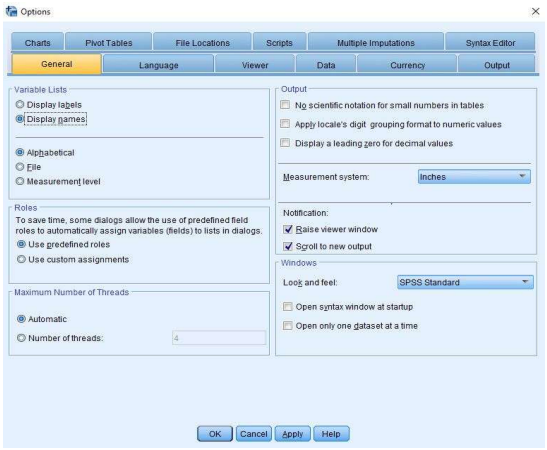


Figure 17-1

Now click on Pivot Tables in the tabs. SPSS offers a number of different "looks" for contingency tables. You might want to experiment with the different choices. For now, however, we're using the System Default under TableLook (see Figure 17-2). Then click on OK.

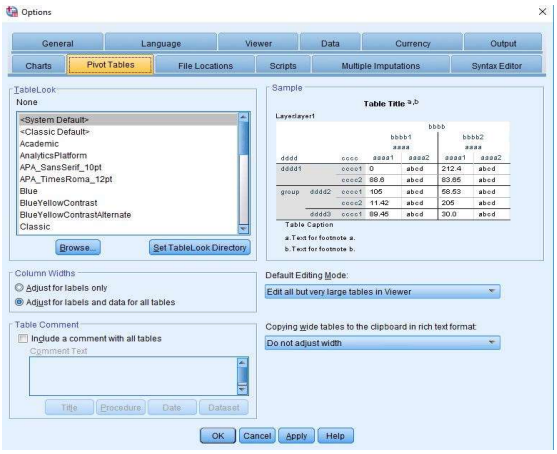


Figure 17-2

To illustrate the Crosstabs technique, we'll use the General Social Survey subset - GSS16A (Questionnaire in Lesson 15).

Crosstabs are particularly useful for exploring the relationship between variables. We're going to see if there is any difference between men and women in their attitudes towards abortion." To create a contingency table (crosstabs), from the menu, click on Analyze, Descriptive Statistics, and Crosstabs. This will open the dialog box shown in Figure 17-3.

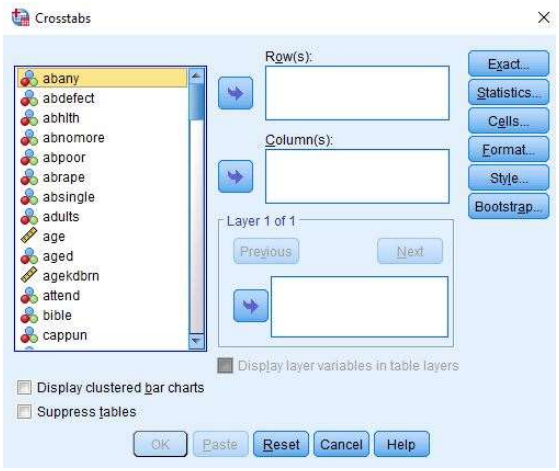


Figure 17-3

You then choose the row (usually the dependent) variable and column (usually the independent) variable.[1] In Appendix A, you will see that there are seven variables that deal with opinions about abortion. Let's choose *abhlth* (abortion if the woman health is endangered) for our row variable and *sex* (respondent's sex) for the column variable. To do this, select the variable you want from the list and click on it to highlight it, then use the arrow keys to the right of the List box to move the variable into either the Row or the Column box. If you've done everything correctly, your screen will look like Figure 17-4, but don't click OK yet!

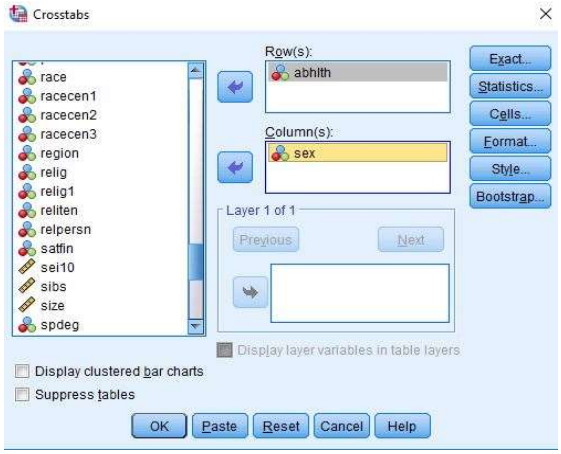


Figure 17-4

In the buttons within the Crosstabs dialog box, click on Cells. Here you have a number of choices for the information you would like to have in each cell of your table. The Observed box should already be selected—it shows the actual number of cases in each cell. You will also want to see percentages as well as raw numbers so that you can easily compare groupings of

[1] The independent variable is the causal variable.

different sizes. You should always make sure that each category of the independent variable totals 100%; our general rule is to have the dependent variable in the row and the independent variable in the column. So choose Columns for the percentages as in Figure 17-5.

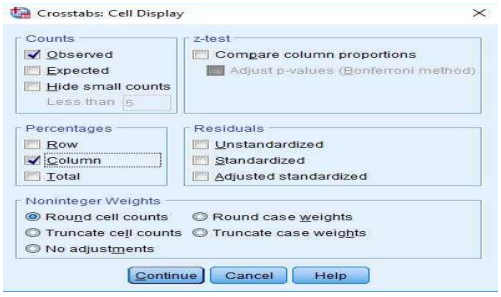


Figure 17-5

Now click on Continue to get back to the Crosstabs dialog box. Once you are back there, click OK. SPSS will now open the Output window, which will show you your table (see Figure 17-6).

abhlth ABORTION IF WOMAN'S HEALTH SERIOUSLY ENDANGERED * sex RESPONDENT'S SEX Crosstabulation

			sex RESPONDENT'S SEX 1 MALE	2 FEMALE	Total
abhlth ABORTION IF WOMAN'S HEALTH SERIOUSLY ENDANGERED	1 YES	Count	737	892	1629
		% within sex RESPONDENT'S SEX	90.0%	88.5%	89.2%
	2 NO	Count	82	116	198
		% within sex RESPONDENT'S SEX	10.0%	11.5%	10.8%
Total		Count	819	1008	1827
		% within sex RESPONDENT'S SEX	100.0%	100.0%	100.0%

Figure 17-6

The Case Processing Summary shows the Valid, Missing, and Total cases. The high percent of missing cases here reflects the people who were not asked this particular question in the survey. Only the valid cases are used in the table.

The Crosstabs shows the 1,827 valid cases arranged in a table that shows what percent of men and women said either Yes or No to the *abhlth* question. Note that 90.0% of the men and 88.5% of the women said Yes, a difference of only 1.5 percentage points.

Your initial conclusion here might be that on abortion issues, there's virtually no difference between men and women in their responses. Is this correct or did you stop your analysis a little too soon? Let's look at a different abortion question. Repeat the steps above, but use *abnomore* as your dependent variable this time. Your results should look like Figure 17-7.

abnomore ABORTION IF MARRIED--WANTS NO MORE CHILDREN * sex RESPONDENT'S SEX Crosstabulation

			sex RESPONDENT'S SEX		
			1 MALE	2 FEMALE	Total
abnomore ABORTION IF MARRIED--WANTS NO MORE CHILDREN	1 YES	Count	397	437	834
		% within sex RESPONDENT'S SEX	48.5%	43.1%	45.5%
	2 NO	Count	421	576	997
		% within sex RESPONDENT'S SEX	51.5%	56.9%	54.5%
Total		Count	818	1013	1831
		% within sex RESPONDENT'S SEX	100.0%	100.0%	100.0%

Figure 17-7

Now we see that 48.5% of the men and 43.1% of the women said Yes to "Abortion if a woman is married and wants no more children." When we compare Figure 17-6 with Figure 17-7, we see there is a large difference between total Yes answers (89% compared with 45%), which indicates that abortion as an issue needs to be broken down into specific conditions if you want to study it in depth. But there still isn't much of a difference between men and women in these two variables. Even though this difference is small, we still might wonder if it is a significant difference. To answer this we will need to do some more statistical analysis.

For our next cross tabulation, again go to the menu and choose Analyze, Descriptive Statistics, and Crosstabs. In the Crosstabs dialog box, place abnomore as the row variable and sex as the column variable. Now click on the Statistics button, then select Chi-Square to obtain a measure of statistical significance, and also select Phi and Cramer's V, which are measures of the strength of association between two variables when one or both are at the nominal level of measurement. Phi is suitable for tables with two rows and two columns, while Cramer's V is appropriate otherwise. Your dialog box should look like Figure 17-8.

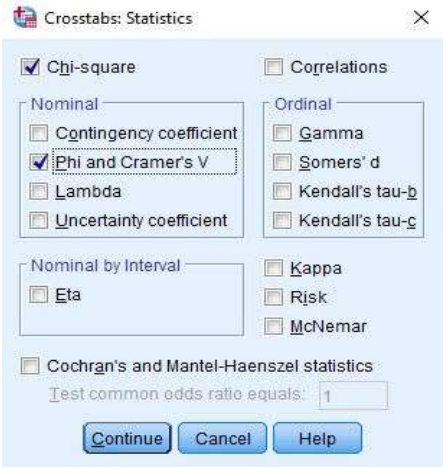


Figure 17-8

Click on Continue, then OK. The table in Figure 17-7 reappears, but with some additional information (you might have to scroll down to see it)—look for "Chi-Square Tests" (Figure 17-9).

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5.309[a]	1	.021		
Continuity Correction[b]	5.093	1	.024		
Likelihood Ratio	5.308	1	.021		
Fisher's Exact Test				.023	.012
Linear-by-Linear Association	5.306	1	.021		
N of Valid Cases	1831				

a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 372.59.

b. Computed only for a 2x2 table

Figure 17-9

The Pearson Chi Square test indicates that the relationship is statistically significant. It would occur by chance approximately 2 times out of 100. The Cramer's V of .054 in Figure 17-10 (Symmetric Measures) indicates that there is virtually no relationship.

Symmetric Measures

		Value	Approximate Significance
Nominal by Nominal	Phi	.054	.021
	Cramer's V	.054	.021
N of Valid Cases		1831	

Figure 17-10

Let's look at a somewhat different table. We're going to consider the relationship between education and political views. Click on Analyze, Descriptive Statistics, and Crosstabs. If the variables you used before are still there, click on the Reset button, then move *polviews* to the Row box and *degree* to the Column box. Since both of these variables are ordinal, we'll want to obtain different statistics to measure their relationship. Click on Statistics and then on Chi-square and Kendall's tau c. (Tau c is a measure of association that is appropriate when both variables are ordinal and do not have the same number of categories.)

Now click on Continue, and then on Cells, and then on Column percents. Now click on Continue and then click on OK. What do the results show? While the Chi-square statistic is statistically significant, the value of Kendall's tau c is quite low, indicating that there is virtually no relationship between these two variables. The pattern of the percentages shows the same lack of relationship.

17.2 COMPARING MEANS

Cross tabulation is a useful method for exploring the relationship between variables with only a few categories. For example, we could compare how men and women feel about abortion. Here, our dependent variable has only two categories—approve or disapprove. But what if we wanted to find out whether the average age at first childbirth is lower for women than for men? In this case, the dependent variable is a continuous one with many values. We could recode it into a few categories (e.g., under 20, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 and older), but that would lead to a significant loss of information. A better approach would be to compare the mean age at first childbirth for men and women

Open GSS16A.sav to answer this question. Click on Analyze, point your mouse at Compare Means, and then click on Means. We want to put age at birth of first child (*agekdbrn*) in the Dependent List and *sex* in the Independent List. Highlight *agekdbrn* in the list of variables on the left of your screen, and then click on the arrow next to the Dependent List box. Now click on the list of variables on the left and use the scroll bar to find the variable *sex*. Click on it to highlight it and then click on the arrow next to the Independent List box. Your screen should look like Figure 17-11. Click on OK and the Output Window should look like Figure 17-12. On the average, women are a little less than 2.5 years younger than men at the birth of first child.

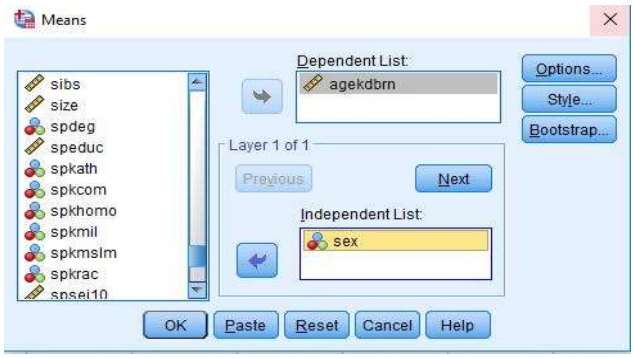


Figure 17-11

Report

agekdbrn RESPONDENT'S AGE WHEN FIRST CHILD BORN

sex RESPONDENT'S SEX	Mean	N	Std. Deviation
1 MALE	25.84	863	5.836
2 FEMALE	23.45	1184	5.739
Total	24.46	2047	5.898

Figure 17-12

17. 2.1 Independent-Samples T Test

If women are, on average, slightly less than 2.5 years younger than men at the birth of their first child, can we conclude that this is also true in our population? Can we infer this from our sample (around 2,000 people selected from the population)? To answer this question, we need to perform a t test. This test will evaluate the hypothesis that men and women in the population do not differ in their mean age at the birth of their first child. By the way, this is called a null hypothesis. The specific type of t test we will use is called the independent-samples t test, since our two samples are completely independent of each other. In other words, the selection of cases in one sample does not influence the selection in the other. We will explore later a situation where this is not the case.

We want to compare our sample of men with our sample of women and then use this information to make an inference about the population. Click on Analyze, then point your mouse at Compare Means and then click on Independent-Samples T Test. Find *agekdbrn* in the list of variables on the left and click on it to highlight it, then click on the arrow to the left of the Test Variable box. This is the variable we want to test so it will go in the Test Variable box. Now click on the list of variables on the left and use the scroll bar to find the variable *sex*. Click on it to highlight it and then click on the arrow to the left of the Grouping Variable box. *Sex* defines the two groups we want to compare so it will go in the Grouping Variable box. Your screen should look like Figure 17-13. Now we want to define the groups so click on the Define Groups button. This will open the Define Groups box. Since males are coded 1 and females 2, type **1** in the Group 1 box and **2** in the Group 2 box. (You will have to click in each box before typing the value.) This tells SPSS what the two groups are that we want to compare. (If you don't know how males and females are coded, click on Utilities in the menu bar, then on Variables and scroll down until you find the variable *sex* and click on it. The box to the right will tell you the values for males and females. Be sure to close this box.) Now click on Continue and on OK in the Independent-Samples T Test box. Your screen should look like Figure 17-14.

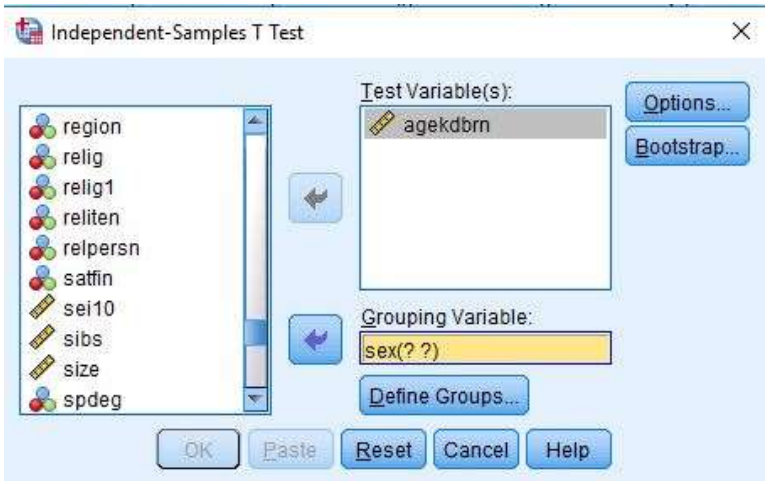
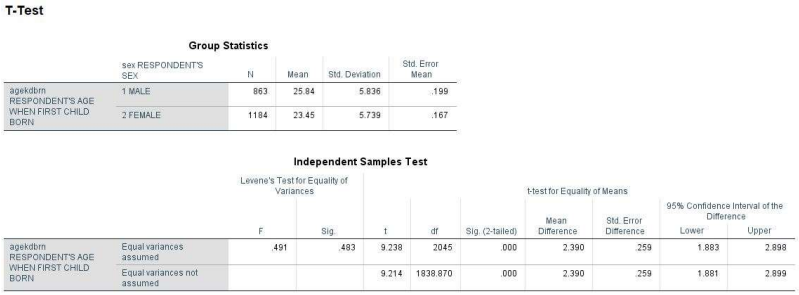


Figure 17-13



T-Test

Group Statistics

	sex RESPONDENT'S SEX	N	Mean	Std. Deviation	Std. Error Mean
agekdbrn RESPONDENT'S AGE WHEN FIRST CHILD BORN	1 MALE	863	25.84	5.836	.199
	2 FEMALE	1184	23.45	5.739	.167

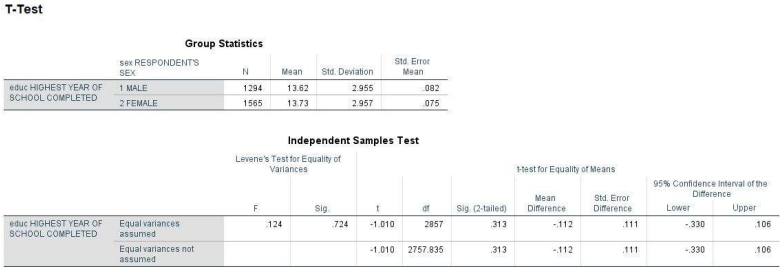
Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
agekdbrn RESPONDENT'S AGE WHEN FIRST CHILD BORN	Equal variances assumed	.491	.483	9.238	2045	.000	2.390	.259	1.883	2.898
	Equal variances not assumed			9.214	1839.870	.000	2.390	.259	1.881	2.899

Figure 17-14

This table shows the mean age at the birth of a first child for men (25.84) and women (23.45), with a mean difference of 2.39. It also presents the results of two t-tests. Remember, these tests assess the null hypothesis that men and women have the same average age at the birth of their first child in the population. There are two versions of this test: one assumes equal variances for the populations of men and women (for agekdbrn), while the other makes no assumption about the variances. The table also provides the degrees of freedom and the observed significance level. The significance value is .000 for both versions of the t-test. In fact, this indicates less than .0005, as SPSS rounds to the nearest third decimal place. This significance level is the probability that the t-value would be this large or larger purely by chance if the null hypothesis were true. Since this probability is very small (less than five in ten thousand), we reject the null hypothesis and conclude that there is likely a difference between men and women regarding the average age at the birth of their first child in the population. Note that this is a two-tailed significance level. To obtain the one-tailed significance level, simply divide the two-tailed value in half.

Let's work another example. This time we will compare males and females in terms of average years of school completed (*educ*). Click on Analyze, point your mouse at Compare Means, and click on Independent-Samples T Test. Click on Reset to get rid of the information you entered previously. Move *educ* into the Test Variable box and *sex* into the Grouping Variable box. Click on Define Groups and define males and females as you did before. Click on Continue and then on OK to get the output window. Your screen should look like Figures 17-15. There isn't much of a difference between men and women in terms of years of school completed. This time we do not reject the null hypothesis since the observed significance level is greater than .05.



T-Test

Group Statistics

	sex RESPONDENT'S SEX	N	Mean	Std. Deviation	Std. Error Mean
educ HIGHEST YEAR OF SCHOOL COMPLETED	1 MALE	1294	13.62	2.955	.082
	2 FEMALE	1565	13.73	2.957	.075

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
educ HIGHEST YEAR OF SCHOOL COMPLETED	Equal variances assumed	.124	.724	-1.010	2857	.313	-.112	.111	-.330	.106
	Equal variances not assumed			-1.010	2757.835	.313	-.112	.111	-.330	.106

Figure 17-15

17.2.2 Paired-Samples T Test

We said we would look at an example where the samples are not independent. (SPSS calls these paired samples. Sometimes they are called matched samples.) Let's say we wanted to compare the educational level of the respondent's father and mother. *Paeduc* is the years of school completed by the father and *maeduc* is years of school for the mother. Clearly our samples of fathers and mothers are not independent of each other. If the respondent's father is in one sample, then his or her mother will be in the other sample. One sample determines the other sample. Another example of paired samples is before and after measurements. We might have a person's weight before they started to exercise and their weight after exercising for two months. Since both measures are for the same person, we clearly do not have independent samples. This requires a different type of t test for paired samples.

Click on Analyze, then point your mouse at Compare Means, and then click on Paired-Samples T Test. Scroll down to *maeduc* in the list of variables on the left and click on it and click on the arrow to the left of the paired Variables box to move it to variable 1 in the paired Variables box. Now click on *paeduc* in the list of variables on the left and click on it and click on the arrow to the left of the paired Variables box to move it to variable 2 in the paired Variables box.

Your screen should look like Figure 17-16. Click on OK and your screen should look like Figure 17-17. This table shows the mean years of school completed by mothers (11.92) and by fathers (11.87), as well as the standard deviations. The t-value for the paired-samples t test is 0.766 and the 2-tailed significance value is 0.444. (You may have to scroll down to see these values.) This is the probability of getting a t-value this large or larger just by chance if the null hypothesis is true. Since this probability is more than .05, we do not reject the null hypothesis. This tells us that there probably isn't a difference between men and women in terms of years of school completed in the population. Notice that if we were using a one-tailed test, then we would divide the two-tailed significance value of .444 by 2 which would be .222. For a one-tailed test, we would also not reject the null hypothesis since the one-tailed significance value is more than .05.

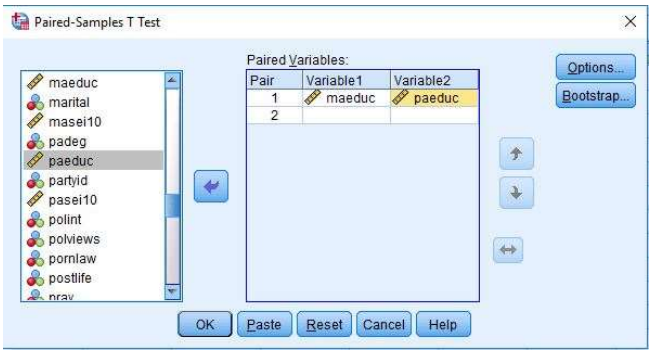


Figure 17-16

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER	11.92	1993	3.747	.084
	paeduc HIGHEST YEAR SCHOOL COMPLETED, FATHER	11.87	1993	4.080	.091

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	maeduc HIGHEST YEAR SCHOOL COMPLETED, MOTHER - paeduc HIGHEST YEAR SCHOOL COMPLETED, FATHER	.052	3.030	.068	-.081	.185	.766	1992	.444

Figure 17-17

17.3 ONE-WAY ANALYSIS OF VARIANCE

In this Lesson we have compared two groups (males and females). What if we wanted to compare more than two groups? For example, we might want to see if age at birth of first child (*agekebrn*) varies by educational level. This time let's use the respondent's highest degree (*degree*) as our measure of education. To do this we will use One-Way Analysis of Variance (often abbreviated ANOVA). Click on Analyze, then point your mouse at Compare Means, and then click on Means. Click on Reset to get rid of what is already in the box. Click on *agekdbrn* to highlight it and then move it to the Dependent List box by clicking on the arrow to the left of the box. Then scroll down the list of variables on the left and find *degree*. Click on it to highlight it and move it to the Independent List box by clicking on the arrow to the left of this box. Your screen should look like Figure 17-18. Click on the Options button and this will open the Means: Options box. Click in the box labeled Anova table and eta. This should put a check mark in this box indicating that you want SPSS to do a One-Way Analysis of Variance. Your screen should look like Figure 17-19. Click on Continue and then on OK in the Means box and your screen should look like Figure 17-20.

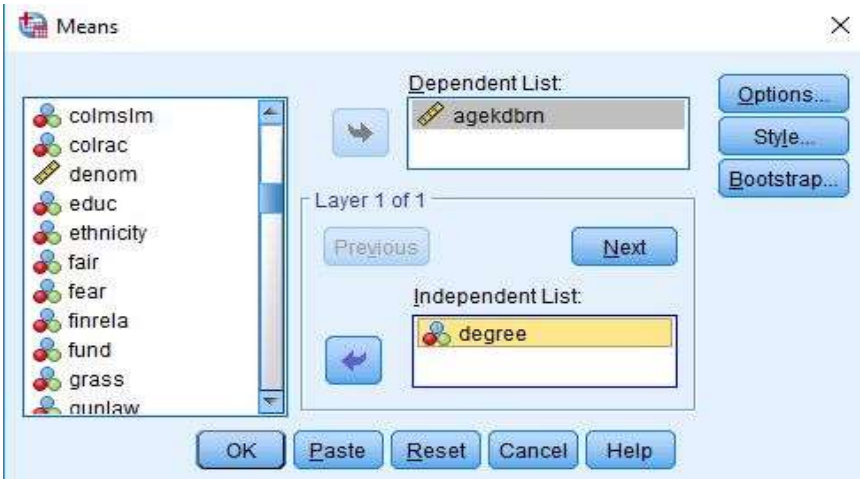


Figure 17-18

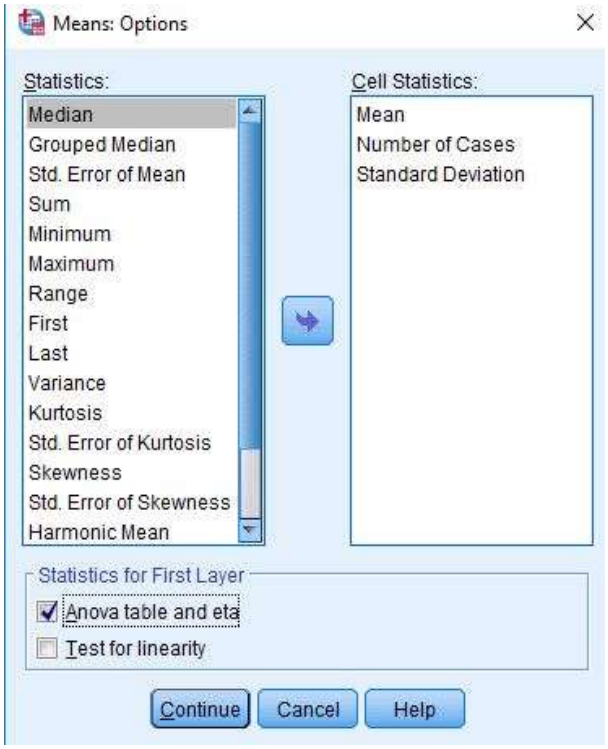


Figure 17-19

Report

agekdbrn RESPONDENT'S AGE WHEN FIRST CHILD BORN

degree RESPONDENT'S HIGHEST DEGREE	Mean	N	Std. Deviation
0 LT HIGH SCHOOL	20.87	280	4.781
1 HIGH SCHOOL	23.27	1043	5.083
2 JUNIOR COLLEGE	24.26	154	5.800
3 BACHELOR	28.16	351	5.797
4 GRADUATE	29.06	214	5.525
Total	24.46	2042	5.903

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
agekdbrn RESPONDENT'S AGE WHEN FIRST CHILD BORN * degree RESPONDENT'S HIGHEST DEGREE	Between Groups	(Combined)	14415.271	4	3603.818	129.442	.000
	Within Groups		56721.774	2037	27.841		
	Total		71137.045	2041			

Measures of Association

	Eta	Eta Squared
agekdbrn RESPONDENT'S AGE WHEN FIRST CHILD BORN * degree RESPONDENT'S HIGHEST DEGREE	.450	.203

Figure 17-20

In this example, the independent variable has five categories: less than high school, high school, junior college, bachelor, and graduate. Figure 17-20 shows the mean age at birth of first child for each of these groups and their standard deviations, as well as the Analysis of Variance table including the sum of squares, degrees of freedom, mean squares, the F-value and the observed significance value. (You will have to scroll down to see the Analysis of Variance table.) The significance value for this example is the probability of getting a F-value of 129.442 or higher if the null hypothesis is true. Here, the null hypothesis is that the mean age at birth of the first child is the same for all five population groups. In other words, the mean age at birth of the first child for all people with less than a high school degree is equal to the mean age for all with a high school degree, all those with a junior college degree, all those with a bachelor's degree and all those with a graduate degree. Since this probability is so low (<.0005 or less than 5 out of 10,000), we would reject the null hypothesis and conclude that these population means are probably not all the same.

There is another procedure in SPSS that does One-Way Analysis of Variance and this is called One-Way ANOVA. This procedure allows you to use several multiple comparison procedures that can be used to determine which groups have means that are significantly different.

17.4 SUMMARY

This lesson explains how to use SPSS to analyse relationships among variables through cross tabulation and mean comparison techniques. Crosstabs are employed when both variables have a small number of categories, often nominal or ordinal. The Lesson shows how to create contingency tables, display percentages, and interpret differences between groups. It also introduces the Chi-Square test to assess statistical significance and measures of association such as Phi, Cramer's V, and Kendall's tau c.

Next, the Lesson discusses comparing means when the dependent variable is continuous. The Independent-Samples t-test is used when two groups are unrelated, while the Paired-Samples t-test applies when two measurements are linked (e.g., mother and father education levels). Finally, the One-Way ANOVA is introduced to compare means across more than two groups, with the F-test assessing whether group differences are statistically significant.

17.5 TECHNICAL TERMS

1. **Cross Tabulation (Crosstabs):** A table used to examine the relationship between two categorical variables.
2. **Observed Frequency:** Actual number of cases in each cell of a crosstab.
3. **Column Percentages:** Percentages calculated within each column to compare groups accurately.
4. **Chi-Square Test:** A statistical test used to assess whether a significant association exists between two categorical variables.
5. **Phi Coefficient:** A measure of association used for 2×2 crosstab tables.
6. **Cramer's V:** A measure of association used when a table has more than two rows or columns.

7. **Kendall's tau c:** A measure of association suitable for ordinal variables with unequal category numbers.
8. **Mean:** The average value of a continuous variable.
9. **Independent-Samples T Test:** Compares the mean of one group with the mean of another independent group.
10. **Paired-Samples T Test:** Compares means of two related or matched samples.
11. **One-Way ANOVA:** A procedure to compare means across three or more independent groups.
12. **F-statistic:** A value used in ANOVA to determine whether group means differ significantly.
13. **Null Hypothesis:** The assumption that there is no difference or no relationship in the population.

17.6 SELF-ASSESSMENT QUESTIONS

1. Define cross tabulation and state its purpose.
2. Why are column percentages used when interpreting crosstabs?
3. What is the purpose of the Chi-Square test in crosstabs?
4. Differentiate between Independent-Samples and Paired-Samples t-tests.
5. What does a statistically significant F-value in ANOVA indicate?
6. Create a crosstab between gender and support for a social issue. Explain how you would interpret both observed counts and column percentages.
7. A researcher finds no significant difference between men and women in average years of schooling. Explain how the t-test supports this conclusion.
8. Conduct a One-Way ANOVA comparing mean income across four education levels. Explain the interpretation of the F-value and significance level.
9. Provide an example where a paired-samples t-test is more appropriate than an independent-samples t-test and explain why.

17.7 SUGGESTED READINGS

1. General Social Survey (GSS) Data Documentation.
2. IBM Corp. (2020). *SPSS Statistics User Guide.*
3. Field, A. (2018). *Discovering Statistics Using SPSS Statistics*. Sage Publications.

Dr. G. MALATHI

LESSON-18

CORRELATION AND REGRESSION

OBJECTIVES OF THE LESSON

After studying this lesson, students will be able to:

1. Understand the concept, purpose, and interpretation of correlation and regression in examining relationships among variables.
2. Conduct Pearson's r and Spearman's rho correlation tests in SPSS and interpret their statistical outputs.
3. Construct and interpret scatterplots with regression lines, regression equations, and R-square values.
4. Perform simple linear regression in SPSS and evaluate model results, including coefficients, predicted values, and residuals.

STRUCTURE OF THE LESSON

18.1 INTRODUCTION TO CORRELATION AND REGRESSION

Correlation and regression analysis (also called "least squares" or "ordinary least squares (OLS)" analysis) helps us examine relationships among interval or ratio variables. In this Lesson, we'll explore techniques for conducting correlation and bivariate regression.

To illustrate these techniques, we'll use the "COUNTRIES.sav" file, which is sourced from various origins and contains data on countries worldwide. See Appendix 2B(Lesson 15) for a codebook explaining the variables included in this file. Open the file following the instructions in Lesson 15 under "Getting a Data File."

We'll begin by considering the relationship between perceived lack of government corruption (in other words, perceived honesty in government) and Internet freedom. Our hypothesis will be that in countries where Internet freedom is high, people will have a greater sense that they can hold government accountable (the technical term for this sense is called "political efficacy") and they will tend to regard their system as less corrupt.

Perceived honesty in government (*honestgov*) is taken from a measure devised by Transparency International. The Internet Freedom Index (*ifreedom*) and the Political Rights Index (*polrights*) are taken from measures devised by Freedom House. The Civil Liberties Index (civillib), which is not used in this Lesson but which you may choose to use in an exercise, is also from Freedom House. The honestgov and the ifreedom indexes range from 0 to 100; the *polrights* and *civillib* indexes are on scales that range from 1 to 7[1].

18.2 USING SPSS FOR CORRELATION ANALYSIS

How close are the relationships among Internet freedom, political rights, and perceived honesty in government? To find out, click on Analyze, Correlate, and Bivariate. A dialog box will appear on your screen. Click on *honestgov* and then click the arrow to move it into the box. Do the same with *ifreedom* and *polrights*. The dialog box should look like Figure 18–1.

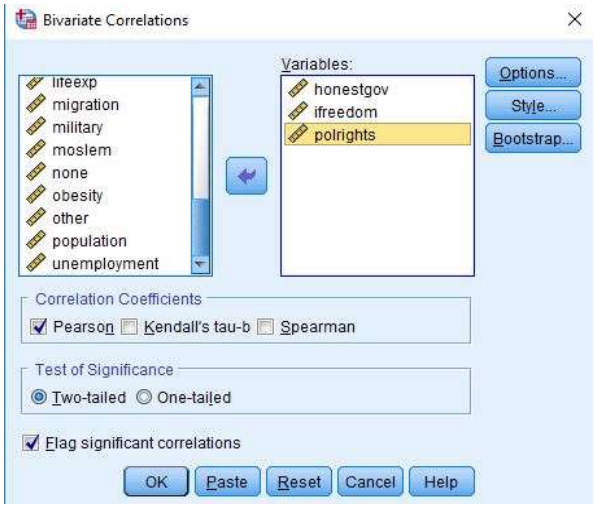


Figure 18-1

The most widely used bivariate test is the Pearson's r correlation coefficient. It is designed for use when both variables are measured at either the interval or ratio level and each variable is normally distributed. However, sometimes these assumptions are violated. If you examine histograms of our three variables, you'll notice that none are truly normally distributed. Additionally, the variables we are using could arguably be considered ordinal rather than interval measures. We'll use Pearson's r, but with caution. SPSS also offers another correlation test, Spearman's rho, which is suitable for variables that are not normally distributed or are ranked (i.e., ordinal rather than interval). We will conduct both tests to compare how much the results differ depending on the test used — in other words, whether relying on Pearson's r for these variables might be misleading.

[1] To make using them more intuitive, and more consistent with honestgov, ifreedom, polrights, and civillib have been recoded from the original sources so that the **higher** the number, the **higher** the level of perceived honesty in government, Internet freedom, political rights, and civil liberties.

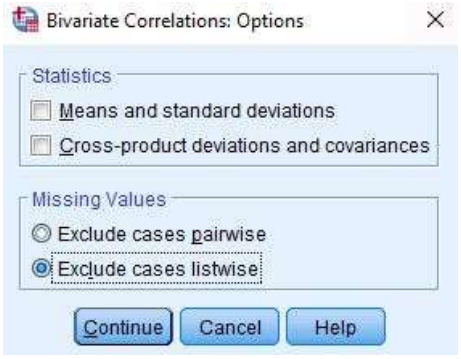


Figure 18-2

In the dialog box, click on Options and, in the resulting box, on Exclude cases listwise. The result should look like Figure 18–2). The reason for doing this is that, as we've noted, *ifreedom* is based on many fewer cases than the other two variables, and we want to be able to make "apples to apples" comparisons. Click on Continue.

The box next to Pearson is already checked, as this is the default. Click in the box next to Spearman. Click the button next to One-tailed test of significance. (This is because we will be testing "directional" hypotheses, that is, not just the idea that two variables are related but, for example, that the higher the value of the *ifreedom* index, the higher the value of the *honestgov* index.) Therefore, we would expect the correlation to be positive. Your dialog box should now look like the one in Figure 18–3. Click OK to run the tests.

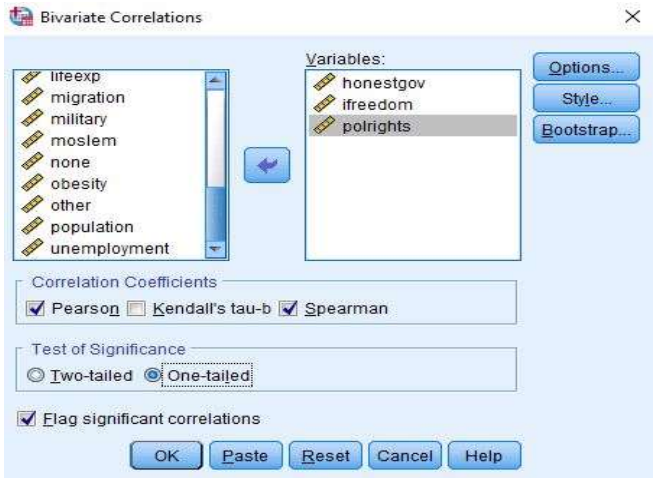


Figure 18-3

Your output screen will show two tables (called matrices): one for Pearson's r and one for Spearman's rho. The Pearson's correlation matrix should look like the one in Figure 18–4.

The cells of the table show the Pearson's r correlation between each variable and each other variable, the level of statistical significance of the relationship (that is, the likelihood that it could have occurred by chance), and the number of cases on which the correlation is based.

Correlations[b]

		honestgov Lack of Perceived Corruption Index	ifreedom Internet Freedom Index	polrights Political Rights Index
honestgov Lack of Perceived Corruption Index	Pearson Correlation	1	.467**	.601**
	Sig. (1-tailed)		.001	.000
ifreedom Internet Freedom Index	Pearson Correlation	.467**	1	.830**
	Sig. (1-tailed)	.001		.000
polrights Political Rights Index	Pearson Correlation	.601**	.830**	1
	Sig. (1-tailed)	.000	.000	

**. Correlation is significant at the 0.01 level (1-tailed).

b. Listwise N=46

Figure 18-4

The correlation coefficient can range from -1 to 1, where -1 or 1 signifies a "perfect" relationship. The farther the coefficient is from 0, whether positive or negative, the stronger the relationship between the two variables. Therefore, a coefficient of .467 is equally strong as a coefficient of -0.467. Positive coefficients indicate a direct relationship: as one variable increases, so does the other. Negative coefficients indicate an inverse relationship: as one variable increases, the other decreases. Notice that the Pearson's r for the relationship between Internet freedom and perceived honesty in government is .467. This suggests, as predicted, that as Internet freedom increases, perceived honesty in government also increases.

The correlation matrix also gives the probability that the relationship we have found could have occurred just by chance. (Labeled as Sig. [1-tailed]). The probability value is .001, which is well below the conventional threshold of $p \leq .05$. Thus, our hypothesis is supported. There is a relationship (the coefficient is not 0), it is in the predicted direction (positive), and is statistically significant.

Correlations[b]

			honestgov Lack of Perceived Corruption Index	ifreedom Internet Freedom Index	polrights Political Rights Index
Spearman's rho	honestgov Lack of Perceived Corruption Index	Correlation Coefficient	1.000	.414**	.539**
		Sig. (1-tailed)	.	.002	.000
	ifreedom Internet Freedom Index	Correlation Coefficient	.414**	1.000	.851**
		Sig. (1-tailed)	.002	.	.000
	polrights Political Rights Index	Correlation Coefficient	.539**	.851**	1.000
		Sig. (1-tailed)	.000	.000	.

**. Correlation is significant at the 0.01 level (1-tailed).

b. Listwise N = 46

Figure 18-5

Recall that we had some concerns about using the Pearson's r coefficient. Figure 18–5 shows the results using Spearman's rho. Notice that the coefficient for the relationship between *ifreedom* and *honestgov* is .414, or about the same as the value of Pearson's r for this relationship. Similarly, the other values of Spearman's rho are similar to those for Pearson's r. This is reassuring.

18.3 SIMPLE LINEAR REGRESSION ANALYSIS IN SPSS

Let's look more closely at the relationship between *ifreedom* and *corruption* graphically by creating a scatterplot. Click on Graphs, Chart Builder. This will open up the dialog box shown in Figure 18-6. (If you get a message telling you to be sure that the measurement levels of each variable have been set properly, click on OK, since this has already been done for you for the "COUNTRIES.sav" file.)

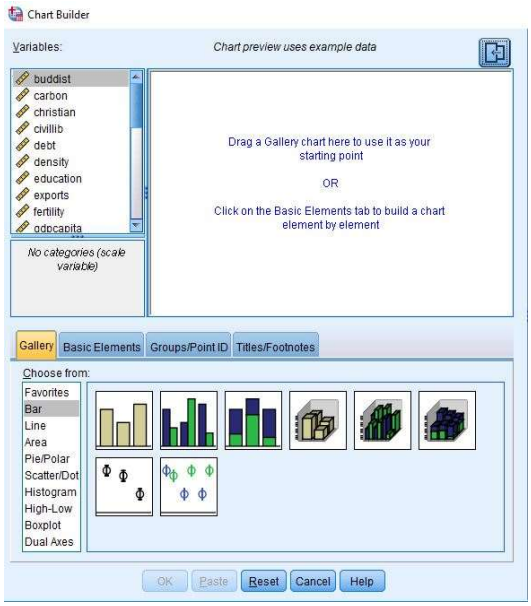


Figure 18-6

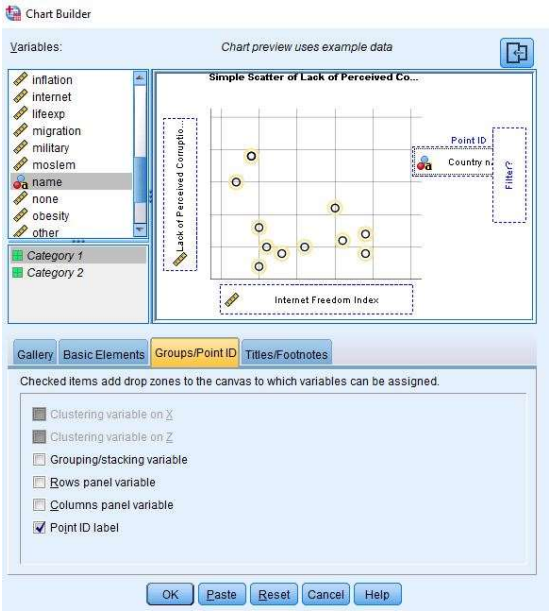


Figure 18-7

Next, in the "Choose from," list at the lower left, click on Scatter/Dot. Then, shift your attention to the sample graph patterns, and click on the first one (upper left). Holding down the mouse button, drag the sample to the large chart preview window. Then, add variables to the chart preview window. From the list of variables, click on *ifreedom* and drag it to the box located on the horizontal (X) axis (because it is the independent variable in our hypothesis and the independent variable belongs on the horizontal axis). Next, click on *honestgov* and drag it into the box located on the vertical (Y) axis. Finally, add data labels: from the menu in the middle of the Chart Builder, click on Groups/Point ID, select Point ID label and, from the list of variables, click on *name* and drag it to the box on the chart called "Point Label Variable?" (Note: Point ID labels aren't a good idea if you have a large number of cases, but will work well here.) Your dialog box should now look like the one in Figure 18–7. Then, click OK. What you see is a plot of perceived honesty in government for each country included in the chart by each

country's level of Internet freedom. Your scatterplot should look like the one in Figure 7–8.

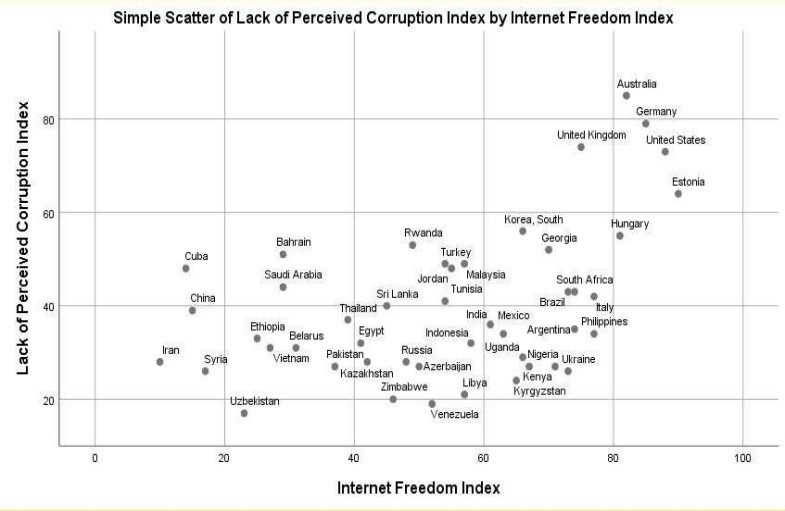


Figure 18-8

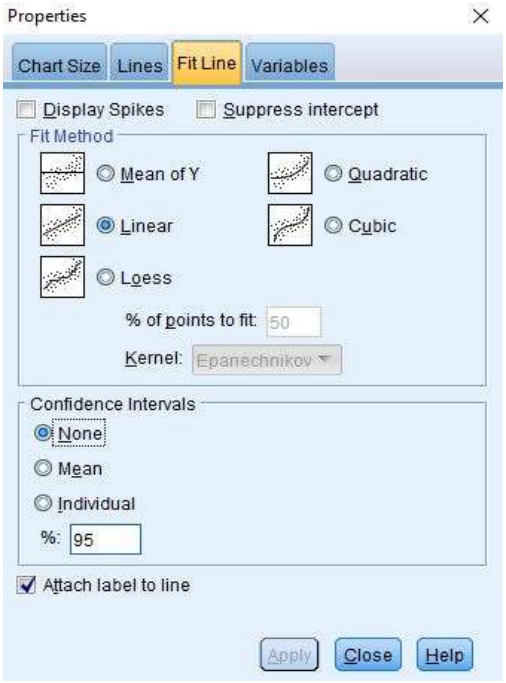


Figure 18-9

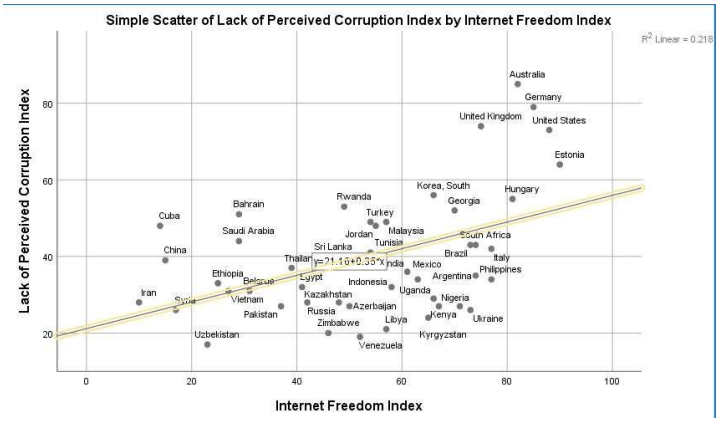


Figure 18-9

You can edit your graph to make it easier to interpret. First, double-click anywhere in the graph. This will cause the graph to open in its own window. On the menu bar, click on Elements, then Fit Line at Total. You will get a dialog box that looks like the one in Figure 18–9. In the Fit Method section, click on Linear (it is the default) and then click on Apply and close the box. (If the Apply button is not active, select a different Fit Method, then change back to Linear before clicking on Apply. If your graph doesn't show country names, click on Elements again, then on Show Data Labels.) Your graph now looks like the one in Figure 18–10. Notice the line variously known as the "least squares line," the "line of best fit," or the "regression line"—we'll go with the last of these— that is now drawn on the graph. Regression and correlation analyze linear relationships between variables, finding the regression line that best fits the data (that is, keeps the errors, the squared distances of each point from the line, to a minimum). Also notice the formula (y=21.16+35*x), called the "regression equation," superimposed on the line, and the R-square Linear statistic (.218) to the right of the graph. We'll return a bit later to the regression equation and the R-square linear statistic (usually just called "r^2").

In general, countries to the right on the graph (that is, those that have freer Internet access) tend also to be higher on the graph (that is, have more perceived honesty in government). This is just what we hypothesized. We can now do some "deviant case analysis." Countries that appear above the regression line are those with more perceived honesty in government than we would expect given their level of Internet freedom, while those below the line have less.

Some countries are pretty much where we'd expect (in that they are close to the line), while some others are well above or below.

Multiplied by 100, r^2 tells us the percentage of the variation in the dependent variable (*honestgov*, on the Y-axis) that is explained by the scores on the independent variable (*ifreedom*, on the X-axis). Thus, Internet freedom explains 21.8% of the variation in perceived honesty in government. Recall that the Pearson's r coefficient was .467. If you take the square root of .218, you get .467, the same as the value of r. (If the relationship were negative, you'd take the negative square root.) Though the r statistic is the one most commonly reported, r^2 is extremely

useful, since it tells us the "proportional reduction in error" we achieve in "predicting" the value of the dependent variable by knowing that of the independent variable.

How strong a relationship is this? There's no firm answer to this question. One scholar (Karl Deutsch) once suggested that, if you can explain at least 10% of the variance of a variable, you have something worth talking about. If your r^2 exceeds .5 (that is, it explains over 50% of variance), then your knowledge exceeds your ignorance! We would probably consider anything between an r^2 of .1 and .5 (or an r between about ±.3 and ±.7) to be a moderately strong relationship.

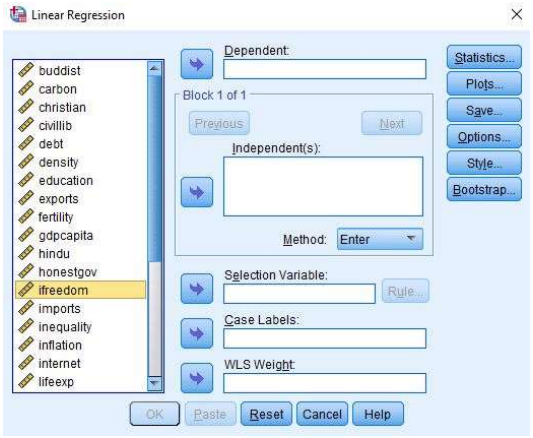


Figure 18-11

Doing a regression analysis can help us to understand the regression line in more detail. Close the SPSS Chart Editor. Click on Analyze, Regression, and Linear. This opens up the dialog box shown in Figure 18-11. Move *honestgov* to the Dependent box, and *ifreedom* to the Independent(s) box. Click OK. The results should look like those shown in Figure 18-12.

The first table just shows the variables that have been included in the analysis. The second table, "Model Summary," shows the R-square statistic, which is .218 (Where have you seen this before? What does it mean?) (Note: the "Adjusted R Square," .200, is slightly lower because it takes into account the number of independent variables in the equation.) The third table, ANOVA, gives you information about the model as a whole. ANOVA is discussed briefly in Lesson 6. Note that if you take the Regression Sum of Squares (the variance explained by the relationship) and divide by the Total Sum of Squares, the result is equal to R^2. The final table, Coefficients, gives results of the regression analysis that are not available using only correlation techniques. Look at the "Unstandardized Coefficients" column. Two statistics are reported: "B," which is the regression coefficient, and the standard error. Notice that there are two statistics reported under B, one labeled as "(Constant)," the other labeled as "ifreedom Internet Freedom Index". The latter is the regression coefficient, which is the slope of the line that you saw on the scatterplot. (Note that in scholarly reports, it is conventional to refer to the regression coefficient using the lower case, "b.") The one labeled as (Constant) is not actually a regression coefficient, but is the Y-intercept (SPSS reports it in this column for convenience only).

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.467[a]	.218	.200	14.410

a. Predictors: (Constant), ifreedom Internet Freedom Index

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2544.374	1	2544.374	12.253	.001[b]
	Residual	9136.431	44	207.646		
	Total	11680.804	45			

a. Dependent Variable: honestgov Lack of Perceived Corruption Index

b. Predictors: (Constant), ifreedom Internet Freedom Index

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	21.164	5.768		3.669	.001
	ifreedom Internet Freedom Index	.348	.099	.467	3.500	.001

a. Dependent Variable: honestgov Lack of Perceived Corruption Index

Figure 18-12

What do these numbers mean? You may recall from your statistics course that the formula for a straight line is:

$Y = a + bX$

Y refers to the value of the dependent variable for a given case, a is the Y-intercept (the point where the line crosses the Y-axis, listed as Constant on your output), b is the slope of the line which describes the relationship between the independent and dependent variables, and X is the value of the independent variable for a given case.

We know that the linear relationship between X and Y (*ifreedom* and *honestgov*) is not perfect. The correlation coefficient was not 1 (or –1), and the scatterplot showed plenty of cases that did not fall directly on the line. Thus, it is clear to us that knowing a country's level of Internet freedom will not tell us without fail its level of perceived honesty in government. It is clear that there is some error built into our findings. This is the reason that the regression line is also called the "Best Fit Line." For these reasons, it is conventional to write the formula for the line as:

$\hat{Y} = a + bX + e$, where "e" refers to error. $\hat{Y}$ ("Y hat") indicates the value of Y predicted by the equation for a given case. We could also write it as "Y'" (Y prime) or "Y^c" (the calculated value of Y).

What can we do with this formula? One thing we can do is make predictions about particular values of the dependent variable, using just a little arithmetic. All we have to do is plug the values from our output into the formula for a line. For now, we will ignore the error terms ("e"), but will come back to them shortly. Plugging the numbers from Figure 18–12 into the formula for a straight line, we obtain $\hat{Y}$=21.164+.348*X, the same equation we saw earlier in Figure 18–10, except that, here, numbers have been carried out to three decimal places. We can then plug in the value of X (*ifreedom*) for any given country, multiply by.348, and add that to 21.164. The result will be the predicted value of the *honestgov* variable for that country.

For example, looking at the file in Data View mode, we see that South Africa, the United Kingdom, and Ukraine all have similar *ifreedom* scores (74, 75, and 73 respectively). Plugging these values into the equation we obtain:

- For South Africa, $\hat{Y}$=21.164+.348*74=46.916.
- For the United Kingdom, $\hat{Y}$=21.164+.348*75=47.264.
- For Ukraine, $\hat{Y}$=21.164+.348*73=46.568.

These numbers represent the predicted values of *honestgov* for these three countries, that is, what the values would be if all three countries fell right on the regression line. In other words, we would predict that, since all three countries have similar *ifreedom* scores, they will also have similar *honestgov* scores. Going back to Data View, however, we see that the actual scores are 43, 74, and 26 respectively. If we subtract the predicted scores from the actual scores ($Y-\hat{Y}$), we obtain the "residual," which is a measure of the error in our prediction for a given case. In this example, the residuals are:

- For South Africa, $Y-\hat{Y}$=43-46.916=-3.916.
- For the United Kingdom $Y-\hat{Y}$=74-47.264=26.736.
- For Ukraine, $Y-\hat{Y}$=26-46.568=-20.568.

In other words, as can be seen in Figure 18-10 above, perceived honesty in government in South Africa is about what we would expect, whereas it is much higher than predicted in the United Kingdom, and much lower than predicted in Ukraine.

We won't go into it here, but you can, for all cases, add the predicted values of the dependent variable and the residuals as additional variables in the data file. To do this, click on SAVE in the regression dialog box, and select Unstandardized Predicted Values and Unstandardized Residuals.

18.4 SUMMARY

Correlation and regression are key statistical techniques used to examine and describe relationships between variables measured at the interval or ratio level. Correlation assesses the strength and direction of association between two variables, with coefficients ranging from -1 to +1. Pearson's r is used for normally distributed interval/ratio variables, while Spearman's rho is appropriate for ordinal or non-normal data.

Regression further examines the nature of this relationship by estimating an equation ($Y = a + bX$) that predicts the dependent variable based on the independent variable. The regression line or "line of best fit" minimises prediction errors, and the R-square (r^2) indicates how much of the variation in the dependent variable is explained by the independent variable. Using SPSS, correlation matrices, scatterplots, regression coefficients, and residuals can be generated to support interpretation and hypothesis testing.

18.5 TECHNICAL TERMS

1. **Correlation:** A Statistical measure of the relationship between two variables.
2. **Pearson's r:** Correlation coefficient for interval/ratio variables assuming normal distribution.
3. **Spearman's rho**: Rank-based correlation for ordinal or non-normal variables.
4. **Positive Correlation:** Both variables increase together.
5. **Negative Correlation:** One variable increases while the other decreases.
6. **Scatterplot:** Graph showing the relationship between two variables.
7. **Regression Line / Line of Best Fit:** The straight line best representing the data trend.
8. **Regression Equation (Y = a + bX):** Predicts values of the dependent variable.
9. **Intercept (a):** Value of Y when X = 0.
10. **Slope (b):** Change in Y for a one-unit change in X.
11. **R-Square (r^2):** Proportion of variance explained by the independent variable.
12. **Residual:** Difference between actual and predicted values ($Y - \hat{Y}$).

18.6 SELF-ASSESSMENT QUESTIONS

1. Distinguish between Pearson's r and Spearman's rho.
2. Explain the meaning of R-square in regression analysis.
3. What is a residual? How is it calculated?
4. Why is the regression line referred to as the "line of best fit"?
5. Explain the process of conducting and interpreting correlation analysis in SPSS using Pearson's r and Spearman's rho.
6. Discuss simple linear regression with an example. Explain how to interpret the regression equation, slope, intercept, and R-square value.
7. Describe how a scatterplot is constructed and interpreted, including the use of the regression line and deviant case analysis.

18.7 SUGGESTED READINGS

1. General Social Survey (GSS) Data Documentation.
2. IBM Corp. (2020). *SPSS Statistics User Guide.*
3. Field, A. (2018). *Discovering Statistics Using SPSS Statistics*. Sage Publications.

Dr. G. MALATHI