Indira Gandhi
National Open University
School of Social Sciences

MEC-109
Research Methods in Economics

Block

1

RESEARCH METHODOLOGY: ISSUES AND PERSPECTIVES

UNIT 1
Research Methodology: Conceptual Foundation **7**

UNIT 2
Approaches to Scientific Knowledge: Positivism and Post Positivism **21**

UNIT 3
Models of Scientific Explanation **43**

UNIT 4
Debates on Models of Explanation in Economics **57**

UNIT 5
Foundations of Qualitative Research: Interpretativism and Critical Theory Paradigm **72**

Expert Committee

Prof. D.N. Reddy
Rtd. Professor of Economics
University of Hyderabad, Hyderabad

Prof. Harishankar Asthana
Professor of Psychology
Banaras Hindu University, Varanasi

Prof. Chandan Mukherjee
Professor and Dean
School of Development Studies
Ambedkar University, New Delhi

Prof. V.R. Panchmukhi
Rtd. Professor of Economics
Bombay University and Former
Chairman ICSSR, New Delhi

Prof. Achal Kumar Gaur
Professor of Economics
Faculty of Social Sciences
Banaras Hindu University, Varanasi

Prof. P.K. Chaubey
Professor, Indian Institute of Public
Administration, New Delhi

Shri S.S. Suryanarayana
Rtd. Joint Advisor
Planning Commission, New Delhi

Prof. Romar Korea
Professor of Economics
University of Mumbai
Mumbai

Dr. Manish Gupta
Sr. Economist
National Institute of Public
Finance and Policy
New Delhi

Prof. Anjila Gupta
Professor of Economics
IGNOU, New Delhi

Prof. Narayan Prasad **(Convenor)**
Professor of Economics
IGNOU, New Delhi

Prof. K. Barik
Professor of Economics
IGNOU, New Delhi

Dr. B.S. Prakash
Associate Professor in Economics
IGNOU, New Delhi

Shri Saugato Sen
Associate Professor in Economics
IGNOU, New Delhi

Course Coordinator: Prof. Narayan Prasad

Block Preparation Team

Units	Resource Person	IGNOU Faculty (Format, Language and Conent editing)	Block Editor (Content)
1	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	Prof. Prem Vashistha NCAER, New Delhi
2	Prof. S.G. Kulkarni Professor of Philosophy University of Hyderabad, Hyderabad	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	
3,4	Prof. D. Narshimha Reddy Rtd. Professor of Economics University of Hyderabad, Hyderabad	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	
5	Dr. Nausheen Nizami Assistant Professor in Economics Gargi College, University of Delhi New Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	

Print Production

Mr. Manjit Singh
Section Officer (Pub.)
SOSS, IGNOU, New Delhi

October, 2015

© *Indira Gandhi National Open University, 2015*

ISBN-978-81-266-

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at :

MEC-109 RESEARCH METHODS IN ECONOMICS

After completion of Master's degree in Economics, many of you may intend to start your career as professional economist. As professional economists, you would be required to carry out the task of analyzing many specific economic situations and indicate their impact on economic policy framework. Some of you may also pursue research degree programmes. In order to perform these tasks, you need to be quipped with the various constituents of Research Methods and the different techniques applied in data collection/analysis. The present course aims to cater to this need. The theoretical perspectives that guides research, tools and techniques of data collection and methods of data analysis together constitute the research methodology. Quantitative Research and Qualitative Research are associated with different paradigms. Accordingly approaches to social enquiry varies. Mixed methods by combining of at least one quantitative method and one qualitative method in measurement, collection or analysis of data is increasingly being used to strengthen the validity of research findings. The present course deals with all these aspects. The course comprises of 6 Blocks.

Block 1: This block deals with the theoretical perspectives/foundations guiding both quantitative and qualitative research. The essence of three paradigms i.e. positivism and post positivism. Interpretitivism and critical theory associated with different research strategies and methodologies have been discussed. This block covers the entire breadth of main trends of development in the philosophy of science and debates in the methodology of economics. The block has been divided into 5 units. **Unit 1** deals with the conceptual foundation related to Research Methodology and its constituents, approaches to social enquiry, research strategy, research process, an elementary idea of hypothesis and measurement scales of variables. **Unit 2** is devoted to scientific methodology covering Positivists' verification approach, Pauper's critical rationalism, and Thomas Kuhan's paradigm's shift approach. The fundamental differences between Popper's and Positivists' views at the one hand, and differences between Popper's view and Kuhan's views on the other, have also been explained. Model of scientific explanation broadly within the framework of positivist theory of scientific method constitutes the core of **Unit 3.** Basic rules of formal reasoning, models of explanation, role of logical reasoning in the formulation of research problem, and the problems involved in systematic explanation of phenomenon have been discussed in this unit. **Unit 4** presents the view of different economists about the models of explanation in economics in the positivists mainstream framework and a brief discussion on the alternative methods of explanation. **Unit 5** throws light on essentials of interpretativism and critical theory paradigm guiding the frameworks of qualitative research. Based upon different ontological and epistemological positions, these paradigms analyse the nature of reality differently.

Block 2 on research design and measurement issues sets the tone and context to provide balanced treatment to quantitative and qualitative research. The block comprises of three units. Research design and mixed methods, the characteristics of quantitative methods and qualitative methods, sampling design, the various issues relating to measurement of variables have been covered in this block.

Block 3: In order to validate the economic theory, we need their empirical verification. Further, in order to examine disparity in income, the inequality measures are important. Similarly in order to describe the development status of an area in terms of several dimensions, we need to develop skill to construct composite index numbers. For all these purposes, good exposure to methods of regression models, in-equality measures and composite index is required. **Block 3** meets this requirement.

Block 4: In order to analyse the quantitative and qualitative data, more analytical techniques like Multi Variate Analysis: Factor Analysis, Canonical Correlation Analysis, Cluster Analysis, Correspondence Analysis, and Structural Equation Models are being increasingly used under either mono method or mix methods. **Block 4** explains all these techniques.

Block 5: A large range of data analysis techniques are used in carrying out qualitative research. However, three important methods namely participatory methods, content analysis, and action research – relevant for conducting the qualitative research studies in the area of Economics have been covered in this Block.

Block 6: The availability of data is crucial for undertaking research. For this, one is expected to be familiar with the different databases, and sources, the concepts used in data compilation and the agencies involved in data collection. **Block 6** focuses on these aspects of database of Indian Economy.

UNIT 1 RESEARCH METHODOLOGY: CONCEPTUAL FOUNDATION

Structure

1.0 Objectives

1.1 Introduction

1.2 Research Methodology and its Constituents

1.3 Theoretical Perspectives

1.4 Approaches to Social Enquiry

1.5 Research Strategies

1.6 Research Process

1.7 Hypothesis: Its Types and Sources

1.8 The Nature, Sources and Types of Data

1.9 Measurement Scales of Variables

1.10 Let Us Sum Up

1.11 Key Words

1.12 Some Useful Books

1.13 Answer or Hints to Check Your Progress Exercises

1.0 OBJECTIVES

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

- Explain the research methodology and its constituents;
- State the various philosophical perspectives that guide research in social sciences;
- Describe various research approaches and strategies that are applied to answer the research questions;
- Describe the various steps involved in the research process; and
- Discuss the various types of research designs appropriate for different types of research.

1.1 INTRODUCTION

Research plays a significant role in human progress. It inculcates scientific and inductive thinking and promotes the development of logical thinking and organization. The role of research in several fields of applied economics has increased in modern times. As an aid to economic policy, it has gained importance for policy makers in understanding the complex nature of the functioning of the economy. It also occupies special significance in analysing problems of business and industry. Social scientists study the relationships among many variables seeking answers to complex issues through research. In short, research aids the process of knowledge formation and serves as an important source of providing policy suggestions to different business, government and social organization. The knowledge of the critical perspectives or philosophy of science, techniques

of data collection and tools or methods of analyzing data are essential for undertaking research in a systematic manner. In this unit, we shall focus our attention on important constituents of research methodology i.e., research perspectives, research approaches and strategies and data types and its sources, hypothesis formulation and measurement scale of variables. At the initial stage of research, several questions may arise in your mind – what do we mean by research? How is the term 'research methodology' distinct from 'research techniques' or 'research methods'? Which type of research approach can be applied to a particular situation or context? What are the steps involved in the research process and how to design a research project? We shall take up these issues in this unit. Let us begin by explaining the term research methodology and its constituents.

1.2 RESEARCH METHODOLOGY AND ITS CONSTITUENTS

Research in common parlance refers to a search for knowledge. It can be defined as a scientific and systematic enquiry either to discover new facts or to verify old facts, their sequences, interrelationships, causal explanation and the adherence to natural laws governing them. It thus aims to discover the truth by applying scientific methods.

Research Methodology is a wider term. It consists of three important elements:

i) theoretical perspectives or orientation to guide research and logic of enquiry,

ii) tools and techniques of data collection, and

iii) methods of data analysis.

Research Methods, comprises of **research techniques** and **tools. Research techniques** refer to the practical aspects of collecting data and the way the information/data obtained/collected is organized and analysed. **Tools** are the instruments that are used for data collection and its analysis. It includes questionnaire/schedules, dairies, check lists, maps, photos, drawings etc. Census and survey methods are mainly used to collect quantitative data. In qualitative research, data is generated/complied by way of participant observation, semi structured interviews, life histories, experiments, pilot studies, scenarios etc. Data analysis involves a set of statistical techniques used in establishing relationships between the different variables and in evaluating the accuracy of the results.

Thus, methodology, methods and tools/techniques are three distinct elements of the research process. Any one of these three elements by itself may not be adequate in many situations. For instance, no data can be systematically collected without adequate knowledge of techniques of data collection. Similarly, data can not be explained without comprehending the philosophy or perspective behind the characteristics underlying the variables to which the data relates. A sound knowledge of statistical techniques is also necessary to analyse the data efficiently.

1.3 THEORETICAL PERSPECTIVES

Theoretical perspectives relate to theories of knowledge which lies within the domain of philosophy of social science. The key concept associated with the perspectives is the paradigm. Let us start to discuss with the concept of paradigm.

Paradigm: A paradigm is a comprehensive belief system, world view or framework that guides research and practice. It consists of

- At the basic or fundamental level, a philosophy of science that makes a number of assumptions about fundamental issues relating to nature and characteristics of truth or reality (ontology) and the theory of knowledge dealing with how can we know the things that exist (epistemology).
- World view, conceptual and theoretical framework that guides research and practices used in the field.
- General methodological prescriptions including tools to be used for information/data collection and data analysis to conduct the work within the paradigm.

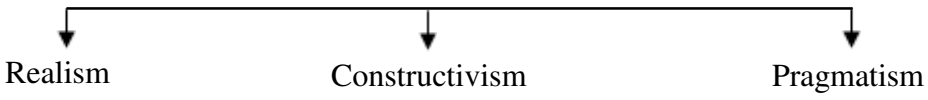
The exact number of world views and the names associated with a particular paradigm vary from author to another but four paradigms in the context of research approaches in social sciences, are important:

- Positivism and post positivism
- Critical theory
- Interpretivism
- Pragmatism.

The peculiar features of these paradigms are:

- They differ on the question of reality,
- They offer different reasons or purposes for doing research,
- They use different types of tools for data collection and sources of information also vary,
- They resort different ways of deriving meaning from the collected data,
- They vary in the relationship between research and practice.

Some authors have classified these paradigms into three categories by re-naming them as:



- Realism begins by assuming that there is a real world that is external to the experience of any particular person and the goal of research is to understand that world.
- Constructivism begins by assuming that everyone has unique experience and beliefs and it posits that no reality exists outside of those perceptions.
- Pragmatism considers realism and constructivism as two alternate ways to understand the world. However, the questions about the nature of reality are less important than questions about what is meant to act and experience the consequence of those actions.

The knowledge of all these perspectives enable a researcher to make a meaningful choice about

1) the research problem;
2) the research questions to investigate this problem;
3) the research strategies to answer these questions;
4) the approaches to social enquiry that accompany these strategies;
5) the concept and theory that direct the investigation;
6) the sources, forms and types of data;
7) the methods for collecting and analyzing the data to answer these questions.

The characteristics of positivism and post positivism as theoretical and methodological perspectives to scientific knowledge have been provided in the next unit i.e. Unit 2 of this course. Similarly the characteristics of interpretivisms and critical theory perspective to conduct social research will be taken up in Unit 5 of this course.

1.4 APPROACHES TO SOCIAL ENQUIRY

Broadly two types of approaches are used in conducting research in social sciences: quantitative and qualitative. The studies conducted within the perspective (framework) of positivism/post-positivism/realism generally resort the quantitative approach and are termed as 'quantitative research'. Quantitative research integrates purposes and procedures that are deductive, objective and generalized. Emphasis is laid on the construction of general theories which are applied universally. Well controlled procedures with large number of cases are followed in conducting the studies.

On other hand, the studies conducted within the perspective of critical theory and interpretivism paradigms are termed as qualitative research. By using induction as a research strategy, qualitative research creates the theory and discovery through flexible, emergent research designs. It tries to evolve meaning and interpretation based on closer contacts between researchers and the people they study. Thus qualitative research consists of purposes and procedures that integrate inductive, subjective and contextual approaches. Based on the above outlines on the types of research – we may say that there are two basic approaches to research viz., quantitative approach and qualitative approach.

Mixed methods research by combining quantitative methods and qualitative methods are being used in social sciences. Hence, mixed method design by integrating quantitative and qualitative approach has also emerged as an approach to social enquiry.

Quantitative approach can be further sub-classified into: inferential approach; experimental approach; and simulation approach.

In **Inferential approach**, database is established through survey method and inference is drawn about characteristics or relationship of variables. In **experimental approach**, greater control is exercised over research environment. Some variables are manipulated to observe their effect on other variables.

Simulation approach refers to the operation of a numerical model that represents the structure of a dynamic process. It involves the construction of an artificial environment in which relevant information/data can be generated. Given the values of initial conditions, parameters and exogenous variables, a simulation is run to represent the behaviour of the process over time.

Qualitative approach deals with the subjective assessment of attitudes, opinions and behaviour of respondents in the field. Results are generated either in non-quantitative form or in a form which are subjected to relatively less rigorous quantitative treatment. Various techniques like group discussions, projective techniques, in-depth interviews etc., are used. The typical characteristics of both these approaches may be summarized in the following Table:

Table 1.1: Typical Characteristics of Qualitative and Quantitative Approach

Characteristics	Qualitative Approach	Quantitative Approach
1) Typical data collection methods	Participant observation, semi-structured interviews, group discussion report cards etc.	Laboratory observations, questionnaire, schedule or structured interviews.
2) Formulation of questions and answers	Open/loosely specified questions and possible answers. Questions and answers are exchanged in two way communication between researcher and respondent.	Closed questions (hypothesis) and answer categories to be prepared in advance.
3) Selection of respondents	Information maximization guides the selection of respondent. Every respondent may be unique (key person).	Representativeness as proportion of population N. Randam sample selection, sample size according to assumptions about distribution in population N.
4) Timing of analysis	Parallel with data collection	After data collection
5) Application of standard methods of analysis	Descriptive methods of analysis are used. Mixed methods are also used.	Standard statistical methods are frequently used.
6) The role of theories in the analysis	Existing theories are typically used only as point of departure for the analysis. Theories are further developed by forming new concepts and relations. The contents of the new concepts are studied and illustrated. Practical application of theory is illustrated by cases.	A-priori deducted theories are opeationalised and tested on data. The process of analysis is basically deductive.

1.5 RESEARCH STRATEGIES

Four basic strategies can be adopted in social research depending upon the researcher's belief/reliance on perspective/paradigm about the nature of reality.

Paradigm	Research Strategy
Positivism	Induction
Post positivism/Realism	Deduction
Critical realism	Retroduction
Interpretativism	Abduction

Each strategy has different starting points (Lewis-Beck, 2004):

1) The inductive strategy begins with collection of data from which generalisation is made and can be used as an elementary explanation;

2) The deductive strategy starts out with a theory that provides a possible answer. The theory is tested in the context of a research problem by collection of relevant data;

3) The retroductive strategy starts out with a hypothetical model of a mechanism that could explain the occurance of a phenomenon under investigation;

4) The abductive strategy starts with laying the concepts and meanings that are contained in social quarters account of activities related to a research problem.

Check Your Progress 1

1) Distinguish between Research Methodology and Research Methods.

.....

.....

.....

.....

.....

2) How does the knowledge of research perspectives help a researcher to undertake research studies in social sciences?

.....

.....

.....

.....

.....

3) How is retroductive research strategy different from abductive strategy?

.....

.....

.....

.....

.....

4) Explain the term 'ontology' and 'epistomology'.

.....

.....

.....

.....

.....

1.6 RESEARCH PROCESS

Research process refers to different steps involved in a desired sequence in carrying out research. However, this does not mean that these steps are always in a given sequence. The various steps involved in research can be illustrated in a flow chart as shown in the figure below:

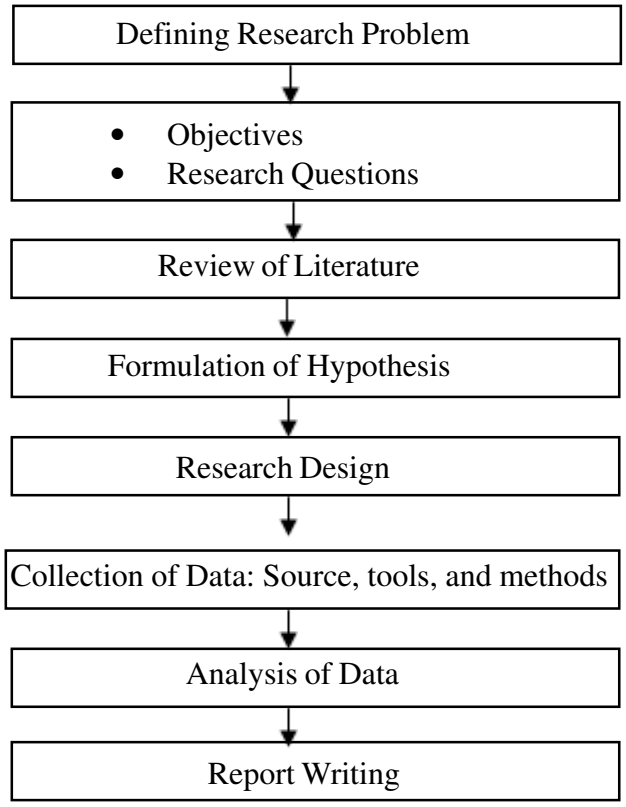


Fig. 1.1: Flowchart of Research Process

The above activities or steps overlap continuously rather than following the prescribed sequence strictly. The steps are not mutually exclusive. The order illustrated is meant only to provide a procedural guideline for research. The steps are briefly elaborated below:

a) **Defining the research problem:** Selecting and properly defining the research problem is the first foremost step. The problem to be investigated must be defined categorically. It is important to identify the general area of interest or a particular aspect of a subject matter desired to be studied. Initially, the problem may be stated in a borad way and later it can be narrowed down in operational terms. Essentially two steps are involved in formulating the

research problem: (i) Understanding the problem thoroughly; and (ii) rephrasing it into meaningful terms from operational/analytical point of view.

It is better to select the subject that is familiar with easy access to research material and data sources. Apart from the topic, following points need to be stated clearly in the research problem:

1) rationale behind the research problem;

2) the aims and objectives as per the requirements of the research questions. The statement of the objectives determines the data to be collected, hypothesis to be tested, techniques for data collection and analysis to be adopted, and the relations intended to be explored;

3) the research questions in the light of the objectives and the theoretical arguments/foundation on which it rests;

4) developing the ideas through discussions; and

5) re-phrasing the research problems identified in (i) above into a working proposition.

The **different steps** to be followed while defining the research problem, therefore are:

- statement of the problem first in a general way to be later sharpened with the help of literature review,
- understanding the nature of the problem, and
- surveying the available literature.

In addition to the above, the following points should also be observed while defining the research problem:

- Technical terms and words or phrases used in the research problem should be explicitly defined.
- Basic assumptions or postulates relating to the research problem need to be clearly stated.
- A clear and unambiguous statement of the investigation should be provided.
- The time-period required and the scope of the study must be duly stated.
- The sources of data and its limitations must be explicitly mentioned.

b) **Review of Literature:** The review of literature is meant to gain insight on the topic and gain knowledge on the availability of data and other materials on the theme of proposed area of research. The literature reviewed may be classified into two types viz. (i) literature relating to the concepts and theory

and (ii) empirical literature consisting of findings in quantitative terms by studies conducted in the area. This will help in framing research questions to be investigated. Academic journals, conference proceedings, government reports, books etc. are the main sources of literature. With the spread of IT, one can access a large volume of literature through internet.

c) **Formulation of Hypothesis:** Specification of working hypothesis (or hypotheses) is the next step of research process. A hypothesis is a tentative statement made which needs to be tested for its logical and empirical confirmation. Hypothesis can be formulated as a proposition or set of propositions providing most probable explanation for occurrence of some event or specified phenomenon. Hypotheses when empirically tested may either be accepted or rejected. A hypothesis must, therefore, be capable of being tested. A hypothesis stated in terms of a relationship between the dependent and independent variables are suitable for econometric treatment. The manner in which hypothesis is formulated is important as it provides the required focus for research. It also helps in identifying the method of analysis to be used.

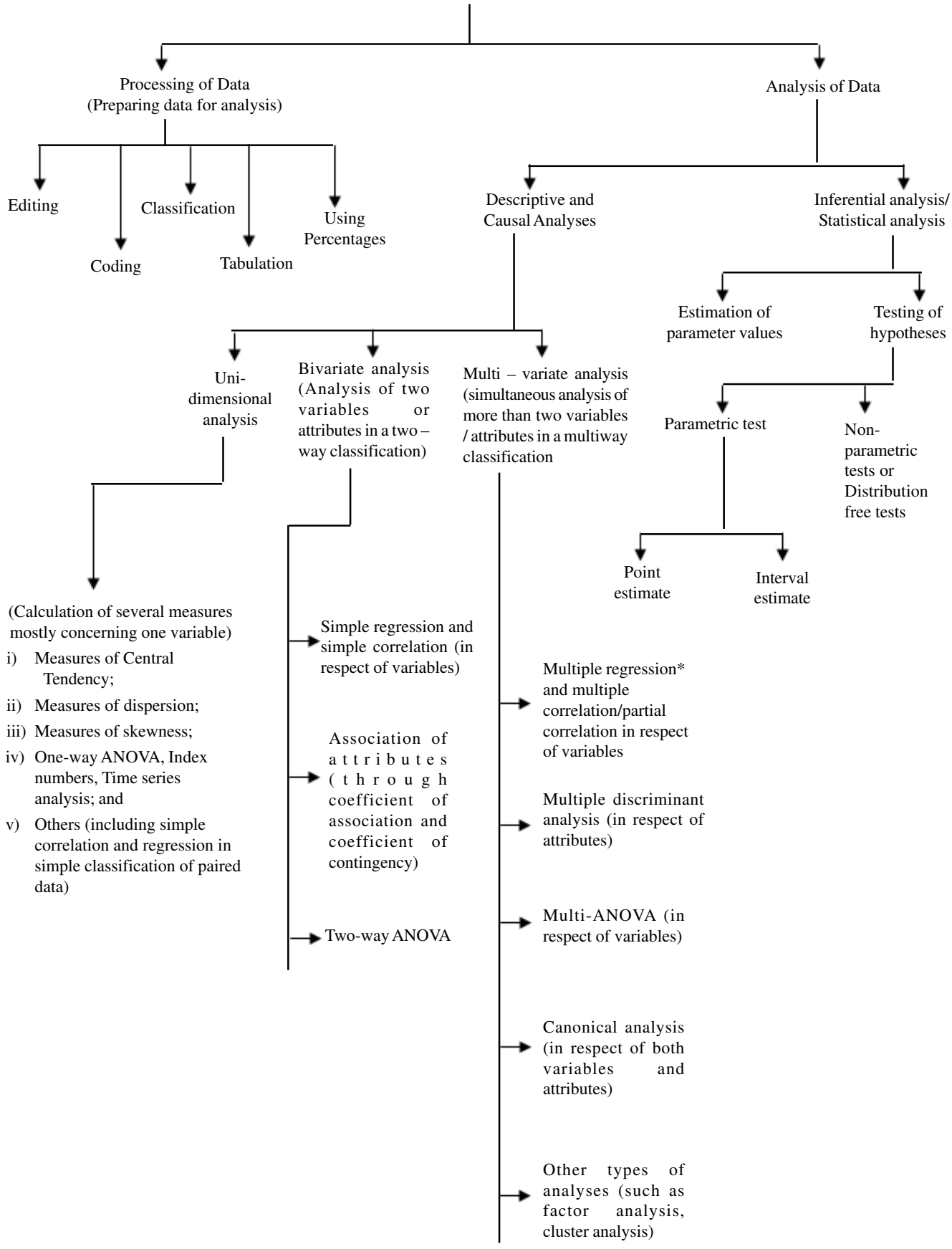
Prior thinking about the subject, examination of the available data and material related to the study, discussion with colleagues and experts help the researcher in formulation of hypothesis. Exploratory or descriptive research can be carried out even without hypothesis.

d) **Research Design:** Research design is the logical conceptual structure within which research is conducted. It is the blueprint for the collection, measurement and analysis of data. Detailed discussion on Research Design has been provided in Unit 6 of Block 2.

e) **Collection of Data:** Collection of data is an essential part of the research process. Data can be primary or secondary. Data collected by the researcher, say by a survey, is primary. The data already collected by some agency and available in some published form is secondary. There are two main techniques of data collection – (i) census and survey (ii) observation method. Primary data can also be collected by means of experiments (e.g., yield under certain conditions, observation at many time points of a certain phenomena, etc.). Intensive fieldwork methods include observation, interview, case study, etc. A survey is usually conducted by the canvassing of a questionnaire. Issues of data collection and sampling design have been discussed in detail in Unit 7 of Block 2.

f) **Analysis of Data:** Analysis involves steps like categorization, coding, tabulation, etc. The principle for classification or categorization of data has to be based on the problem under study or the hypothesis formulated. The category must be exhaustive and sufficient for classifying all responses. They must be distinct, separate and mutually exclusive. **Coding** involves grouping of responses falling under a particular category. **Tabulation** is a means of organizing the responses to facilitate comparisons bringing up the inherent relations between two or more variables. It is an orderly arrangement of data in columns and rows. Analysis and inference is usually aided by the application of different statistical and econometric techniques. Some of the major techniques usually employed in research have been reflected in the **summary chart 1.2**

Summary Chart 1.2
ANALYSIS OF DATA



Most of these analytical techniques have been covered in Units 9,10,11 and 12 of Block 3 and Units 13,14,15,16 and 17 of Block 4 of this course.

Report Writing: Originality and clarity are the two vital components of research report. It is the ultimate test of one's analytical ability and communication skills. It is an exercise involving the organization of ideas. The research report needs to be presented in such a manner that the readers can grasp the context, methodology and findings easily. The report comprise of two parts: the preliminary pages and the main text. In the preliminary pages, the report should indicate the title of the research study, name of the researcher (and his team members) and the name of the institution and/or the month/year of preparation of the report. This should be followed by a 'preface' in which the main context of preparing the report along with key findings must be presented. Towards the end of the 'preface', the important sources/persons can suitably be acknowledged.

The main text begins with an introductory chapter followed by the major aspects of the study organized into different chapters. The introductory chapter should contain a clear statement of the objectives of the study, rationale behind the study, a brief summary of the literature review, hypotheses tested (if any) and the definitions of the major concepts employed in the study. The methodology adopted in conducting the study must also be fully explained along with an explicit mention of the limitations of the study. The subsequent parts of the main text, should present the major aspects of the study arranged in a logical sequence splited into appropriate sections and subsections. The inter-connection between different sections should be properly maintained so that the report reads smoothly.

The implications of the results of the study should be stated towards the end of the report. The implications may comprise of:

i) the inferences drawn from the study;

ii) the conditions which may limit the extent of generalizations of the inferences drawn; and

iii) the questions that remain unanswered along with new areas for further research identified. The conclusion drawn from the study should be clearly related to the objectives/hypotheses stated in the introductory section.

The report may also include an 'executive summary' outlining the context and methodology, and major findings of the study. The 'executive summary' is placed right at the beginning (i.e. before the introductory chapter) so as to provide a concise picture of the entire report.

1.7 HYPOTHESIS: ITS TYPES AND SOURCES

Hypothesis are potential explanations that can account for our observations of the external world. They usually describe cause and effect relationships between a proposed mechanism or process (the cause) and our observations (the effect).

In quantitative research, there are two methods of hypothesis generation: deductive method and inductive method.

If hypothesis is generated from a theory, it is called deductive approach. On the other hand, if it is generated from observation, it is termed as inductive approach.

Deductive approach: Theory → Hypothesis→Observation→Confirmation.

Inductive approach: Observation→Pattern→Tentative Hypothesis→Theory.

In qualitative research, a hypothesis might be framed in terms of social setting having certain features, which through observation, can be confirmed or falsified. In survey or experimental research or otherwise hypothesis testing establishes the statistical significance of a finding.

1.8 THE NATURE, SOURCES AND TYPES OF DATA

For undertaking any meaningful research in terms of situational assessment, testing of models, development of theory, evaluation of economic policy, data is essential. The availability of data therefore, determines the scope of analysis. In any research, the researcher is expected to state the sources of the data used in the analysis, their definitions, and methods of collection.

The data may be of three types; Time series, Cross-section and Pooled.

1) **Time Series Data:** It is a set of observations on the values that a variable takes at different times. Such data may be collected at regular time intervals such as daily (i.e. prices, whether reports etc.), weekly (like money supply figures), monthly (i.e. consumer price index etc.) quarterly (i.e. GDP), annually (i.e. government budget etc.).

2) **Cross Section Data:** Cross-section data are data on one or more variables collected at the same point of time. For example the data on the census of population collected by the Registrar General of India.

3) **Pooled Data:** In pooled data, the elements of both time series and cross section are clubbed. For example, over a period of time say from 2000 to 2013, we have data on saving, investment and GDP across Indian states.

Panel, Longitudinal or Micro Panel Data: This is a special type of pooled data in which the same cross-sectional (say a family or firm) is surveyed overtime.

The Sources of Data: The data used in empirical analysis may be collected by a governmental agency (e.g. CSO, NSSO, RBI, Labour Bureau etc.), an international agency (e.g. International Monetary Fund (IMF) or a private organization. Such data is called secondary data because these are collected from secondary sources. The details about the kind of secondary data compiled by the different data collecting agencies have been presented in Block 6 of this course.

The data collected by the investigator or researcher through field work is termed primary data. Such data is collected by using different tools like questionnaire, schedule, interview etc. under quantitative approach and participant observation, open ended interview, group discussion, key information etc. (under qualitative approach). The methods for collecting the primary data have been discussed in Unit 7 of Block 2.

1.9 MEASUREMENT SCALES OF VARIABLES

Measurement is the process of observing and recording the observation that are collected. A variable can be measured at four levels: ratio scale, interval scale, ordinal scale and nominal scale. The suitability of analytic technique depends on the measurement scale. A particular econometric/statistical technique that may be suitable for ratio scale variables may not be suitable for nominal scale variables.

It is therefore desirable to know the distinctions among the four types of measurement scales. These are discussed below:

1) **Nominal Scale:** Under the nominal scale, the data is recorded into categories, without any order or structure. In other words, if against any question/ statement, response is recorded simply yes/no, then the scale will be nominal. It has no order and there is no distance between yes and no.

   The statistical techniques that can be used with nominal scales data are: Mode, cross tabulation – with chi-square. Other highly sophisticated modeling techniques available for nominal scale data are – logistic Linear Regression Model, Principal component analysis, factor analysis etc.

2) **Ordinal Scale:** In terms of power of measurement, ordinal scale comes next to nominal scale. Recording Ranks against various options/choices is the simplest ordinal scale. If you are asked by a researcher to rank 5 fruits from most favorable to least favorable, he is essentially asking you to create an ordinal scale of preference.

   Median, Mode, rank order correlation, non-parametric correlation and modeling techniques are used with ordinal data.

3) **Interval Scale:** In a situation when we not only talk about **differences in order** but also **differences in the degree of order**, it is referred to as interval scale. For example, if we are asked to rate our satisfaction with a piece of software on a 7 point scale, from dissatisfied to satisfied, we are using an interval scale.

   Mean and stardard deviation, correlation, regression, analysis of variance, factor analysis techniques can be used with interval scale data.

4) **Ratio Scale**: A ratio scale is the top level of measurement and satisfies the following properties:

   i) Measurement of each observation of a variable in numerals (quantitative terms) and hence possible to work out the ratio of two observations. For a variable X taking two values X_1 and X_2 the ratio will be X_1 / X_2.

   ii) Measurement of distance between two observation X_1 and X_2 i.e. $(X_2 - X_1)$.

   ii) Indication of the natural ordering (ascending or descending) of the elements of a variable. Therefore, comparison such as $X_2 \geq X_1$ or $X_2 \geq X_1$ are meaningful.

The statistical techniques used in interval scale can easily be used in ratio scale also. You will find elaborate discussion on measurement scales and scaling techniques in Unit 8 of Block 2.

Check Your Progress 2

1) In what way review of literature helps a researcher?

   .....

   .....

   .....

   .....

2) What is meant by a hypothesis?

.....

.....

.....

3) Distinguish between time series data and cross section data.

.....

.....

.....

4) State the various measurement scales of data.

.....

.....

.....

1.10 LET US SUM UP

Research plays a significant role in human progress. It inculcates scientific temper and logical thinking. Research refers to a scientific and systematic enquiry aiming either to discover new facts or to verify the old facts. The theoretical perspectives, tools and techniques of data collection and methods of data analysis together constitute the Research Methodology. Broadly there are two basic approaches of research; quantitative and qualitative. Quantitative approach generates the data in quantitative terms. This can, therefore, be subjected to quantitative analysis. Subjective assessment of attitudes, opinions and behaviour are tackled by the qualitative approach to research. Research process involves seven steps – identification of research problem, review of literature, objectives, formulating the hypothesis, finalizing research design, collection of data, analyzing the data and report writing. Data can be of three types: time series, cross section and panel data. A variable can be measured at four levels: ratio scale, interval scale, ordinal scale and nominal scale. The suitability of analytic technique depends on the measurement scale. Hypothesis is a tentative explanation of any event or phenomenon in the external world. There are two methods of hypothesis generation: deductive method and inductive method. Thus, this unit provides an overview of the conceptual foundation of Research Methodology and enables the learner to comprehend the various processes involved in carrying out the research study.

1.11 KEY WORDS

Coding : A system of symbols, letter or words used in transmitting messages.

Epistemology : Epistemology refers to the theory of knowledge of how human beings come to have knowledge of the world around them – of how we know what we know. Broadly there are two theories: Rationalism and Empiricism.

Rationalism : Rationalism is based on the idea that reliable knowledge is derived from the use of "pure" reason.

Empiricism : Empiricism envisages that the knowledge of the world can be obtained only through direct sense-experience.

Experimental Testing Research : Research in which independent variables are manipulated.

Ontology : It is a branch of philosophy that is concerned with the nature of reality. It deals with the theories about what makes up reality.

Induction : Induction is a process for moving from particular statements to general statements. This logic is used in social sciences to produce theory from data.

Deduction : Deduction is a process used to derive particular statements from general statements. A hypothesis is deduced from a theory and is tested by empirical data.

Abduction : Abduction refers to the process of moving from the way social actors describe their way of life to technical, social scientific description of that social life. It has two stages: (a) describing these activities and meaning and (b) deriving categories and concepts that can form the basis of an understanding or an explanation of the problem at hand.

Retroduction : Retroduction refers to the process of going back from, below, or behind observed patterns or regularities to discover what produces them. This logic of enquiry focuses to locate the structures and mechanisms that have produced the regularity. These structures and mechanisms envisage the tendencies or powers of things to act in a particular way.

Research Hypothesis : Research hypothesis is a predictive statement that relates an independent variable to a dependent variable. Such a hypothesised relationship is usually meant to be tested by research method. Predictive statements, which cannot be objectively verified, or the relationships that are assumed which cannot be tested do not qualify as research hypothesis.

Tabulation : Setting out the data/information in tabular form.

1.12 SOME USEFUL BOOKS

1) Kothari C.R. (1985): *Research Methodology – Methods and Techniques*, Wiley Eastern Publication Chapter 1-3 pp 1 to 67.

2) LeWis-Beck Michael S; Bryman Aan, Tim Futing, Liao (Ed) (2004): *The Sage Encyclopedia of Social Sciences Research Method,* Volumes 1,2 and 3, Sage Publications, New Delhi.

3) David L. Morgan (2014): *Integrating Qualitative and Quantitative Methods,* Sage Publications, New Delhi

1.13 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See section 1.2

2) See section 1.3

3) See section 1.5

4) See section 1.3

Check Your Progress 2

1) Review of literature enables the researcher to know about the availability of relevant material and data. It also helps in framing the research questions.

2) See section 1.7.

3) A set of values of a variable at different times is time series data whereas a set of values of one or more variables collected at a point of time is cross section data.

4) Various measurement scales of a variable are: ratio scale, interval scale, ordinal scale and nominal scale.

UNIT 2 APPROACHES TO SCIENTIFIC KNOWLEDGE: POSITIVISM AND POST POSITIVISM

Structure

2.0 Objectives

2.1 Introduction

2.2 Positivist Philosophy of Science

2.2.1 Central Tenets of Positivist Philosophy

2.3 Attack on Positivist Philosophy of Science

2.4 Karl Popper's Philosophy of Science

2.5 Criticism against Karl Popper's Philosophy of Science

2.6 Thomas Kuhn's Philosophy of Science

2.7 Popper Versus Kuhn

2.8 Let Us Sum Up

2.9 Keywords

2.10 Some Useful Books

2.11 Answer or Hints to Check Your Progress Exercises

2.0 OBJECTIVES

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

- Describe the aim of science as a cognitive enterprise;
- Explain the basic ideas of positivism;
- Appreciate the arguments or criticisms leveled against positivist philosophy of science;
- Explain the central tenets of Popper's philosophy of science;
- State the fundamental differences between Popper's and positivists' views about theory of scientific method;
- Assess Karl Popper's philosophy of science;
- Describe the essentials of Kuhn's philosophy; and
- Discuss the basic differences between the Popperian and the Kuhnian models of science.

2.1 INTRODUCTION

Science has not only shaped our mode of living in the world but also our ways of thinking about the world. It is for this reason it has acquired a central place in the intellectual life of our times. Because of its central place in the modern culture, three disciplines have emerged which has made science itself their object of inquiry. These three disciplines are: History of Science, Sociology of Science and Philosophy of Science. Whereas science studies the world natural or social

or psychological, these three disciplines study science itself. History of Science and Sociology of Science are essentially twentieth century disciplines. No doubt, Philosophy of Science has a great past built by the contributions of individual philosophies in different centuries. However, as a widespread and coordinated discipline, it is also an essentially twentieth century phenomenon.

History of Science is a study of the development of scientific ideas and institutions in a chronological order. Sociology of Science is the study of the relation between social factors and forces on the one hand and the growth of scientific ideas and institutions, on the other. In short, it is an inquiry into the relation between science and society. History of Science deals with science as a historical phenomenon and sociology of science is concerned with science as a social institution. Philosophy of Science is a study of Science as a cognitive enterprise, that is, as a knowledge-seeking activity. The enquiry we call **Philosophy of Science** addresses questions like "What is the aim of Science?", "What is the method of Science?" In what sense, if any, is Science objective, rational and progressive?"

It is obvious that answers to questions such as these are of importance in figuring out how if at all, and in what sense, if any, social sciences including Economics, can claim to be scientific.

2.2 POSITIVIST PHILOSOPHY OF SCIENCE

Positivism (which is also called "logical positivism") is a movement in Philosophy of the first half of the 20th century. Positivists debunked the whole of traditional Philosophy by attacking Metaphysics which was the most important branch of Philosophy. Metaphysics was so central to Philosophy that Philosophy was virtually identified with Metaphysics. Metaphysics sought to answer most general and basic questions like "What is Ultimate Reality?" "Does God exist? And is so what is God's relation with the world?", "Is there a soul?", "What is the relation between the mind and the body?", "Is man free or are human actions determined?" etc., Positivists maintained that Metaphysics was a spurious discipline because metaphysical statements (such as "Matter is Ultimate reality" or "Ultimate Reality is Spiritual" etc.,) are meaningless since they are not verifiable in experience. "A statement" they claimed "is meaningful if and only if it is verifiable." Apart from being anti-metaphysical, they were empiricists i.e., according to them, experience is the source of knowledge. They called themselves "Neo-Empiricists" to distinguish themselves from the Traditional Empiricists of 17th and 18th centuries like Locke, Berkeley and Hume.

2.2.1 Central Tenets of Positivist Philosophy

In science we start with particular observations regarding what is the case. Using the method of induction, we arrive at definitions which are statements about the essential nature of things. We, then, using the method of deduction, on the basis of definitions, arrive at demonstrations which show why things must be what they are. Thus, the aim of science is two fold: **Definition and Demonstration**. The method of science, broadly, is of two types: **Induction and Deduction**. All science must proceed from "What is" given to us in particular observations to "what must be" as shown by the demonstrations. The path of science is an arch whose two end points are observation and demonstration.

In the whole span of three centuries – from the beginning of the 17th century till the end of the 19th century – two views stand out prominently as answer to the question regarding the aim and method of science. The first view is called Induction. The second view is called **Hypothesisism,** according to which the method of science is the method of Hypothesis. **Francis Bacon** is the Father of Inductivism and **Rene Dercartes** is the Father of Hypothesisism. The two views provide two distinct models of scientific practice, which we may call **Baconian** and **Cartesian** models of Scientific practice.

Positivists worked out a well-knit philosophy of science. Here are some of the central tenets of the Positivist Philosophy of Science:

1) Science is qualitatively distinct from, superior to and ideal for all other areas of human endeavor (**Scientism).**

2) The distinction, superiority and idealhood that Science enjoys is traceable to its possession of method (**Methodologism**).

3) There is only one method common to all sciences, irrespective of their subject matter (**Methodological monism**).

4) That method which is common to all sciences, natural or human, is the method of Induction (**Inductivism**).

5) The hallmark of science consists in the fact that its statements are systematically verifiable.

6) Scientific observations are or can be shown to be "pure" in the sense that they are theory-free.

7) The theories are winnowed from facts or observations.

8) The relation between observation and theory is unilateral in the sense theories are dependent on observations whereas observations are theory-independent.

9) To a given set of observation based statements, there corresponds uniquely only one theory (just as from a given set of premises in an argument, only one conclusion follows).

10) Our factual judgments are value-neutral and our value judgments have no factual content (Fact-Value Dichotomy thesis); hence, science being the foremost instance of factual inquiry, does not have value commitments.

11) That all scientific explanation must have the following pattern:

$L_1 \ldots\ldots L_n$

$I_2 \ldots\ldots I_n$

Therefore, E.

Where $L_1 \ldots\ldots L_n$ is a set of laws, $I_2 \ldots\ldots I_n$ is a set of statements describing initial conditions and E is the statement describing phenomenon to be explained. That is to say, to explain a phenomenon scientifically is to deduce its description from a set of laws (which are called "Covering Laws") via a set of statements describing initial conditions. In sum, all explanation worthy to be called '**scientific**' must contain laws and involve deduction (Hence, this is called Deductive-Nomologism where 'nomological' means 'concerning laws').

12) The aim of science is either economical description of phenomena or precise prediction of facts and not providing an account of observations in terms of unobservables. Hence, scientific theories are not putative descriptions of the unobservable world. The aim of science has nothing to do with alleged reality of such a world (Anti-Realism).

13) Unlike other areas of activity, science is progressive in the sense that scientific change is always change for better, whereas other areas exhibit just change: the progress of science consists in the accumulation of observations on the one hand, and, cumulative growth of theories,on the other hand. The latter means that any new theory includes the old theory(plus something). Thus the growth of science essentially exhibits continuity.

14) Science is objective in the sense that its theories are based on 'pure' observations or facts which as theory free. Interpretations may be subjective but observations/facts are objective because they are free from interpretation/ theory.

15) Science is rational because the principle of Induction which is central to the method of science is rationally defensible, in spite of Hume's skepticism regarding its rational defensibility.

Positivists tried to justify the principle of Induction by invoking the concept of pure observation. According to them, theories are arrived at on the basis of the Principle of Induction. If we can show that theories are very closely related to pure observations, the Principle of Induction stands rationally justified. They tried to work out a whole project to demonstrate rational justification of the Principle of Induction on these lines.

2.3 ATTACK ON THE POSITIVIST PHILOSOPHY OF SCIENCE

It must be noted that the concept of pure observation is necessary for positivist philosophy of science for showing that science is objective and that science is rational. Hence, the centrality of this concept to the positivist philosophy of science. Therefore, the collapse of this concept to the positivist philosophy of science though other theses of positivist philosophy of science listed above also were demolished by its opponents. Let us briefly look at the arguments which demolished the positivist thesis of pure observations i.e., thesis at Sl.No.6 in the list above.

Firstly, observations presuppose some principle of selection. We need relevant observations. In science it is the problem that decides what is a relevant observation and thus provides the principle of relevance. Hence, there can not be observations without a – prior problem.

As Popper says "Before we can collect data, our interest in data of a certain kind must be aroused; the problem always comes first". It may be objected that we become aware of the problems because of observations and hence observations come first and therefore positivists are right. But this objection does not hold. Two persons might make same observations but one may come out with a problem and the other may not. Therefore, mere observations would not generate problems in science. Usually problems are generated when there is a clash between what

we observe and what we expect. Of the two persons making the same observations, one comes out with a problem because he sees a conflict between what he observes and what he expects whereas the other observer may have not expectation which conflicts with what he observes. The former believes in a theory which produces certain expectations which conflict with his observations and hence he comes out with a problem. In other words, a – prior belief in a theory is necessary for a problem to be generated and a – prior awareness of the problem is necessary for making relevant observations. Thus, theory precedes observations.

Secondly, in science observations are taken into account only if they are described in a language that is currently used in a particular science. An observation, however, genuine, is no observation unless it is expressed in a recognized idiom. It is the theory which provides the idiom or language to be used to describe facts or observations. It is relevant to quote the words of Pierre Duhem, a distinguished physicist and philosopher:

"Enter a laboratory. The table crowded with an assortment of apparatus: an electric cell, silk covered copper wire, small cups of mercury, spools, a mirror mounted on an iron bar; the experimenter is inserting into small openings the metal ends of ebony-headed pins: the iron bar oscillates and the mirror attached to it throws a luminous band upon a celluloid scale: the forward backward motion of this spot enables the physicist to observe the minute oscillations of the iron bar. But ask him what he is doing. Will he answer "I am studying the oscillations of an iron bar which carries a mirror?" No, he will say that he is measuring the electrical resistance of the spools. If you are astonished, if you ask him what his words mean, what relation they have with the phenomenon he has been observing and which you have noted at the same time as he, he will answer that your question requires a long explanation and that you should take a course in electricity".

Thirdly, most of the observations in science are made with the help of instruments. These instruments are constructed or designed in accordance with the specifications provided by some theories. These theories, one may say, form the software of these instruments. Belief in the reliability of these instruments implies the acceptance of these theories which has gone into the making of these instruments. This observations presuppose prior acceptance of theories.

Fourthly, observations in science need to be legitimized i.e., ratified by a theory. An example makes the point clear. We all know that Galileo used some telescopic observations to support the Helio-centric theory against the geo-centric theory of Ptolemy and his followers. His opponents did not consider the telescopic observations adequate. Why did they not? No doubt, they had belief in the reliability of telescope; they had no problem in using telescope for terrestrial purposes i.e., making observations of earthly objects. They opposed the extension of telescopic observations to the celestial sphere i.e., regarding heavenly bodies. Their argument was that the normal factors like background and neighbourhood which help our normal perceptions are absent in the sky. Further, it is impossible to directly verify whether telescopic observations of heavenly bodies are accurate. They rightly demanded from Galileo a theory of light which would justify the extension of the telescope from terrestrial to celestial sphere. Galileo had no such theory. But he rightly believed that such a theory could be provided in future so that telescopic observations would get justification. Thus, while the opponents of Galileo insisted that the telescopic observations be justified by an

optical theory at the same time as their acceptance, Galileo maintained that the justification could be provided subsequent to their acceptance. It may be noted that both sides accepted that the telescopic observations needed justification in term of a theory of light.

All this does not mean that observations are theory-dependent whereas theories are observation-independent. Observations and theory are interdependent though it is not easy to clarify what the nature of this interdependence is. However, positivists were wrong in claiming that observations are theory –independent. To say that observations are theory dependent is to say that observation is not a passive reception but an active participation of our cognitive faculties equipped with prior knowledge which we call theory. After all, observations are not 'given' but 'made'.

Check Your Progress 1

1) What is the aim of science?

.....

.....

.....

.....

.....

2) What do you mean by the term 'inductivism'?

.....

.....

.....

.....

3) What are the features of hypothesis?

.....

.....

.....

.....

4) Match the following:

A	B
i) The relation between observation and theory	i) Value judgments
ii) Factual inquiry under positivist approach does not have	ii) Precise prediction of facts
iii) The aim of Science is	iii) Objective
iv) Science is	iv) Unilateral

5) Give two counter arguments against the positivist thesis of pure observation.

.....

.....

.....

2.4 KARL POPPER'S PHILOSOPHY OF SCIENCE

Karl Popper was the first to react against the positivist philosophy of science. In fact he started attacking it quite early. But his attack on positivists and his alternative to the positivist philosophy of science came to be widely known at the beginning of the second half of the 20th century. Popper's theory of science, particularly his theory of scientific method has won a lot of admirers among scientists and philosophers. As we know, positivists tried to work out a **sophisticated version of Inductivism**. Popper worked out a **sophisticated version of Hypothesisism**. We shall briefly consider his views on the nature of science.

According to Popper, the central task of philosophy is not to solve Hume's problem or problem of Induction as thought by Positivists. This is because (1) the problem of Induction cannot be solved, and (2) it need not be solved because the method of science is not the method of Induction. The central task of philosophy of science, Popper maintains, is to solve what he calls the problem of demarcation or Kant's problem i.e., the problem of identifying the line of demarcation between science and non-science. Popper maintains that **what distinguishes science from the rest of our knowledge is the systematic falsifiability of scientific theories**. This falsifiability is the line of demarcation between science and non-science. Falsifiability is the criterion of scientificity. A statement is scientific if and only it is falsifiable.

In accordance with what he considers to be the hallmark of scientific theories, Popper puts forward what he considers to an adequate model of scientific method. He characterizes his model of scientific method as **Hypothetico-Deductive model**. According to him, the method of science is not method of Induction but the method of **Hypothetico-Deduction**. What are the fundamental differences between these methodological models"? Firstly, **the inductivist model maintains that our observations are theory-independent and therefore are indubitable. That is to say, since observations are theory-independent, they have probability Value 1**. It also says that our theories are only winnowed from observations and therefore our scientific theories have the initial probability value 1 in principle. Of course, inductivists admitted that in actual practice, the theories may contain something more than what observation based statements say, with the result that our actual theories may not have been winnowed from observations.

Hence, the need for verification arises. Popper rejects the inductivist view that our observations are theory-free and hence rejects the idea that our observation statements have probability equal to 1. More importantly, he maintains that theories are not winnowed from observations or facts, but are free creations of human mind. Our scientific ideas, in other words, are not extracted from our observations; they are pure inventions. Since **our theories are our own**

constructions, not the functions of anything like pure observations, which according to Popper are anyway myths, the initial probability of our scientific theories is zero.

From this it follows that whereas according to the inductivists what scientific tests do is to mrely find out whether our scientific theories are true, according to Popper scientific tests cannot establish the truth of scientific theories even when the tests give positive results. If a test gives a positive result, the inductivists claim that the scientific theory is established as true, whereas according to Popper all that we calim is that our theory has not yet been falsified. In Popper's scheme no amount of positive result of scientific testing can prove our theories. Whereas the inductivists speak of confirmation of our theories in the face of positive results of the tests, Popper only speaks of corroboration. In other words, in the inductivist scheme we can speak of scientific theories as established truths, whereas in the Popperian scheme, a scientific theory, however, well supported by evidence remains permanently tentative. We can bring out the fundamental difference between verificationism (inductivism) and falsificationism (Hypothetico-Deductivism) by drawing on the analogy between two systems of criminal law. According to one system, the judge has to start with the assumption that the accused is innocent and consequently unless one finds evidence against him, he should be declared innocent. According to the other, the judge has to start with the assumption that the accused is a culprit and consequently, unless evidence goes in his favour, he should be declared to be a culprit. Obviously the latter system of criminal law is harsher than the former. The inductivist scheme is analogous to the former kind of criminal law, where as the Hypothetico-Deductive scheme is akin to the latter one.

The steps of Scientific Procedure

In the Popperian scheme we begin with a problem, suggest a hypothesis as a tentative solution, try to falsify our solution by deducing the test implications of our solution, try to show that the implications are not borne out and consider our solution to be corroborated if repeated attempts to falsify it fail. Thus, problem, tentative solution, falsification and corroboration constitute the steps of scientific procedure. Popper's theory of scientific method is called Hypothetico-Deductivism because, according to him, the essence of scientific practice consists of deducing the test implications of our hypothesis and attempting to falsify the latter by showing that the former do not obtain, whereas according to Inductivism the essence of scientific practice consists of searching for instances supporting the generalization arrived on the basis of some observations and with the principle of induction.

Popper claims that the Hypothetico-Deductive model of scientific method is superior to inductivist model for the following reasons:

Firstly, it does justice to the critical spirit of science by maintaining that the aim of scientific testing is to falsify our theories and by maintaining that our scientific theories are, however corroborated, going to permanently remain tentative. In other words, the Hypothetico-Deductivist view presents scientific theories as permanently vulnerable with the sword of possible falsification always hanging on their head. The inductivist view of scientific method makes science a safe and defensive activity by portraying scientific testing as a search for confirming instances and by characterizing scientific theories as established truths. According

to Popper, the special status accorded to science is due to the fact that science embodies an attitude which is essentially open-minded and anti-dogmatic. Hypothetico-Deductivism is an adequate model of scientific practice because it gives central place to such an attitude.

Secondly, Popper thinks that if science had followed the inductivist path; it would not have made the progress it has. Suppose a scientist has arrived at a generalization. If he follows the inductivist message, he will go in search of instances which establish it as a truth. If he finds an instance which conflicts with his generalization, what he does is to qualify his generalization saying that the generalization is true except in the cases where it has to be held unsupported. Such qualifications impose heavy restrictions on the scope of the generalization. This results in scientific theories becoming extremely narrow in their range of applicability. But if a scientist follows the Hypothetico-Dedectivist view, he will throw away his theory once he comes across a negative instance instead of pruning it and fitting it with the known positive facts. Instead of being satisfied with a theory, tailored to suit the supporting observations, he will look for an alternative which will encompass not only the observations which supported the old theory but also the observations which went against the old theory and more importantly which will yield fresh test implications. The theoretical progress science has made can be explained only by the fact that science seeks to come out with bolder and bolder explanations rather than taking recourse to the defensive method of reducing the scope of the theories to make them consistent with facts. Hence, Popper claims that the Hypothetico-Deductive model gives an adequate account of scientific progress. According to him, if one accepts the inductivists account of science one fails to give any explanation of scientific progress.

Thirdly, the **Hypothetico-Deductive** view according to Popper avoids the predicament encountered by inductivist theory in the face of **Hume's** challenge. As we have seen, Hume conclusively showed that the principal of induction could not be justified on logical grounds. If Hume is right, than science is based upon an irrational faith. According to Hypothetico-Deductivist view, science does not use the principle of induction at all. Hence, even though Hume is right, it does not matter since science follows the Hypothetico-Deductivism are so radically different that the latter in no way face any threat akin to the one faced by the former. In this connection, he draws our attention to the logical asymmetry between **verification**, the central component of the inductivist scheme, and **falsification**, the central component of the Hypothetico-Deductivist scheme. They are logically asymmetrical in the sense that one negative instance is sufficient for conclusively falsifying a theory, whereas no amount of positive instances are sufficient for conclusively falsifying a theory, whereas no amount of positive instances are sufficient to conclusively verify a theory. It may be recalled that Hume was able to come out with the problem of induction precisely because a **generalization** (all theories according to Inductivism are generalizations) cannot be conclusively verified.

How does Popper characterize scientific progress? According to him, one finds in the history of science invariable transitions from theories to better theories. What does the work 'better' stand for? It may be recalled that, according to Popper, no scientific theory, however, corroborated can be said to be 'true'. Hence, Popper drops the very concept of truth and replaces it by the concept of **Verisimilitude** (truth-likeness or truth-nearness) in his characterization of the

goal of science. In other words, through science cannot attain truth, i.e., though our theories can never be said to be true, science can set for itself the goal of achieving higher and higher degrees of Verisimilitude, i.e., successive scientific theories can progressively approximate to truth. So, in science we go from theory to better theory and the criterion for betterness is Verisimilitude. But what is the criterion for Verisimilitude? The totality of the testable implications of a hypothesis constitutes what he calls 'the empirical content' of the hypothesis. The totality of the testable implications which are borne out constitute the truth content of the hypothesis and the totality of the testable implications which are not borne out is called the falsity content of the hypothesis. The criterion of the Verisimilitude of a theory is nothing but truth content minus the falsity content of a theory. In the actual history of science we always find, according to Popper, theories being replaced by better theories, that is, theories with higher degree of Verisimilitude. In other words, of the two successive theories, at any time in the History of Science, we find the successor theory possessing greater Verisimilitude and is therefore better than its predecessor. In fact, according to him, **a theory is rejected as false only if we have an alternative which is better than the one at hand** in the sense that it has more testable implications and a greater number of its testable implications are already borne out. The growth of science is convergent in the sense that the successful part of the old theory is retained in the successor theory with the result the old theory becomes a limiting case of the new one. The growth of science thus shows continuity. In other words, it is the convergence of the old theory into the new one that provides continuity in the growth of science. It must also be noted in this connection that unlike the Inductivists or Positivists, Popper is a **Realist** in the sense, according to him, scientific theories are about an unobservable world. This means that the real world of the unobservable thought can never be captured entirely by our theories, evidence that through the gap between Truth and our theories can never be completely filled, it can be progressively reduced. Consequently, the real world of unobservable will be more and more like what our theories say though not completely so.

How does Popper establish the objectivity of scientific knowledge? Inductivists sought to establish the objectivity of science by showing that scientific theories are based upon pure observations. The so-called 'pure' observations were supposed to be absolutely theory-free. They are only 'given' and hence, free from subjective interpretations. Popper, as we have seen, rightly rejects the idea of pure observations. Consequently, he cannot accept the inductivist account of the objectivity of science. What engenders scientific objectivity, according Popper, is not the possibility of pure observation, but the possibility of **inter-subjective testing**. In short, science is objective because it is public and it is public because its claims are inter-subjectively testable.

To the question, "Which comes first, observation or theory?" the Inductivist answers 'observation'. Popper answers 'earlier observation or earlier theory'. To him the question is as illegitimate as the question, "which comes first, egg or hen?" which can be answered only by saying 'earlier egg or earlier hen?'

It will be convenient if we list the main theses of Popper's philosophy of science arranged in a manner with our list of the theses of the positivist philosophy of science:

1) Science is qualitatively distinct from, superior to and ideal for all other areas of human endeavour (**scientism**).

2) The distinction, superiority and idealhood that science enjoys is traceable to its possession of a method (**Methodologism**).

3) There is only one method common to all science irrespective of their subject matter (Methodological Monism).

4) That method which is common to all sciences, natural and human, is the method of Hypothetico-Deduction (**Hypothetico-Deductivism**).

5) The hallmark of science (i.e., the distinguishing mark of science) consists in the fact that its statements are systematically falsifiable (**falsifiability**).

6) Scientific observations are not and cannot be shown to be pure; that is, they are theory-dependent.

7) Theories are not winnowed from observations or facts; they are pure inventions of human mind i.e., only conjectures and not generalizations based on 'pure observations'.

8) The relation between observation and theory is one of interdependence.

9) To a given set of observation-statements there might correspond more than one theory.

10) Our factual judgements may have value commitments and our value judgements may have cognitive content (hence fact-value dichotomy is unacceptable); science is not value neutral but the value commitments can be critically discussed and therefore they are not subjective.

11) All scientific explanations must have deductive-nomological pattern and thus the thesis of Deductive-Nomologism is acceptable.

12) The aim of science is to provide an account of observable world in terms of unobservable entities and to provide accounts of those unobservable entities in terms of further unobservable entities. Unobservable entities are, therefore, real and our theories are putative descriptions of such real entities ('Realism').

13) Unlike other areas of human activity, there is progress in science which consists in going from one theory to a better theory. Here, 'better' means 'more true'. 'More to true' means 'the world of unobservables'. In short, science is progressive in the sense our successive theories in any domain of science exhibit greater and greater verisimilitude or truth-nearness i.e. the match between our theories and reality. Unlike positivists, Popper rejects the idea that progress of science is characterized by cumulative growth of theories. According to him, a new theory is entirely new and not an old theory plus an epsilon as Positivists thought. Thus, in Popper's scheme, the growth of science is essentially discontinuous. Of course, Popper makes some room for continuity also when he says that old theory (at least true part of it) is a limiting case of the new theory.

14) Science is not objective in the sense scientific theories are based on pure observations as positivists thought because there are no pure observations. Science is objective in the sense its theories are inter-subjectively testable.

15) Lastly, science is not rational in the sense the principle of Induction can be rationally justified as Positivists thought. The principle of Induction cannot be rationally justified; nor is it used by science. Science is rational in the sense it embodies critical thinking. Apart from insisting that our theories be falsifiable, science has institutional mechanisms for practicing and promoting critical thinking. What is rationality other than critical thinking? Positivists and Popper differ from each. The theses (1), (2), (3) and (11) are common to both Positivists and Popperians. Popper rejects most of other theses of Positivists, especially their central thesis which concerns the idea of pure observation. Finally, he agrees with the Positivists that science is uniquely progressive, objective and rational; but his notions of progressiveness, objectivity and rationality of science are entirely different from those of Positivists.

2.5 CRITICISM AGAINST KARL POPPER'S PHILOSOPHY OF SCIENCE

A serious lacuna in Popper's position concerns his idea of scientific progress.

First of all, according to Popper, the growth of science is essentially discontinuous in the sense that a new theory which displaces an old theory is not the old theory plus an epsilon because it is entirely new. Yet, he seeks to make room for continuity in the growth of science by insisting that the old theory is a limiting case of the new theory. In this connection he cites an example of Newtonian mechanics and Relativistic mechanics. The former is the limiting case of the latter in the sense that in a certain domain both give the same results. Thus the former is contained in the latter. Hence there is some continuity in the growth of science. But Popper overlooks the fact that such examples of an old theory being a limiting case of the new one are rare. For example, it is absurd to say that Phlogiston theory is a limiting case of oxygen theory or that Ptolemy's theory is limiting case of Copernican theory.

Secondly, Popper says that successive theories in any domain exhibit increasing verisimilitude i.e., truth nearness. That is, reality constituted by unobservable entities is more like what a new theory says than what its immediate predecessor says. This means that following Popper we have to say that the ultimate constituents of matter are more like fields as the present physical theory says than like particles (atoms) as claimed by Newtonian theory. This is unintelligible. What does it mean to say that the ultimate constituents of matter are more like fields than particles called atoms? Either they are like fields or like particles.

Thirdly, when Popper says a new theory is better than the old one (in the sense it is more true), he assumes that the two theories can be compared. This means that they have something common which makes them comparable. But this has been ably questioned by Thomas Kuhn who sought to show that when one fundamental theory replaces another, the two theories are so radically different as to make any talk of comparison between them highly questionable. We shall now turn to his views.

Check Your Progress 2

1) On what basis, according to Popper, a line can be demarked between science and rest of knowledge:

.....

.....

.....

.....

2) Identify the main differences between Inductive Model and Hypothetico-Deduction Model.

.....

.....

.....

.....

3) What are the Central components of Hypothetico-Deductive Model?

.....

.....

.....

.....

4) In what sense, science is rationale?

.....

.....

.....

.....

2.6 THOMAS KUHN'S PHILOSOPHY OF SCIENCE

Thomas Kuhn's work **The Structure of Scientific Revolutions** is a milestone in the history of the 20th century of philosophy of science. A brief exposition of his basic ideas are given below.

According to Kuhn, in the life of every major science there are two stages (1) **Pre-paradigmatic stage**, and (2) **paradigmatic stage**. In the pre-paradigmatic stage one finds more than one mode of practicing that science. That is, there was a time in the history of Astronomy when different schools of Astronomy practised Astronomy differently. So is the case with Physics, Chemistry and Biology. In that stage their situation was similar to that which obtains today in areas like art, philosophy and even medicine wherein divergent modes of practicing these disciplines co-exist. Today, we speak of schools of Art (e.g., painting), schools of Philosophy and systems/schools of medicine. But today we do not speak of Schools of Astronomy or Physics or Chemistry or Biology.

This is, according to Kuhn, in areas like art, philosophy and medicine that did not, and cannot make a transition from pre-paradigmatic stage to paradigmatic stage, which marks the disappearance of plurality, that is, disappearance of schools. In other words, the transition means replacement of plurality by monolith i.e., a uniform mode of practice.

Such a transition is made possible, Kuhn claims, by acquisition of a **paradigm**. When a science makes such a transition, we may say, it has bcome 'mature' or 'science' in the proper sense of the term. Astronomy was the first to make such a transition followed by Physics, Chemistry and Biology in that order. Social Sciences are still, according to him, in the pre-paradigmatic stage, though Economics is showing signs of such a transition. This is evident from the fact that in Social Sciences there is no consensus on fundamentals as we can see prevalence of distinct schools in every Social Science.

So, the transition to maturity is effected by acquiring a paradigm by a science. The question is "What is a paradigm?".

We all know that Ptolemy's *Almagest* Newton's *Principia* and Darwin's *Origin of the Species* are path-breaking works in the areas of Astronomy, Physics and Biology respectively. According to Kuhn, these works provided paradigms for these disciplines. They did so by specifying the exact manner in which these disciplines ought to proceed. They laid the ground rules regarding what problems these disciplines must tackle and how to tackle them. Hence, paradigms are Universally recognized achievements that for a time provide model problems and solutions to community of practitioners. Hence, in the **first** place, a paradigm specifies what that ultimate constituents of that sphere of reality, which a particular science is inquiring into, are.

Secondly, it identifies the model problems. **Thirdly**, it specifies the possible range of solutions. **Fourthly**, it provides the necessary strategies and techniques for solving the problems. **Lastly**, it provides examples which show how to solve certain problems. In other words, a paradigm is a disciplinary matrix of a professional group. Once a science possesses a paradigm, it develops what Kuhn calls, a 'normal science tradition'. Normal science is the day-to-day research activity purporting to force of nature into conceptual boxes provided by the paradigm. The practitioners of normal science, that is, a scientist who engages in day-to-day research, internalizes the paradigm by professional education. This explains the prevalence of textbook culture in science education.

Of course, scientific practice is not exhausted in terms of day-to-day research or 'normal science'. When a paradigm fails to promote fruitful, interesting and smooth normal science, it is considered to be in a crisis. The deepening of the crisis leads to the replacement of the existing paradigm by a new one. This process of replacement is called '**scientific revolution**'. Therefore, scientific revolutions are "the tradition-shattering complements to the tradition bound activity of normal science." Thus, once a science enters the paradigmatic stage, it is characterized by (1) normal science, and (2) revolutions. In sheer temporal terms, normal science occupies much larger span than revolutionary science. That is to say, science is revolutionary once a while and mostly it is non-revolutionary or normal. Also the scientific activity engaged in by most of the practitioners can be characterized aptly in terms of noral science. On account of this temporal and numerical

magnitude we can say that much of the scientific activity as we ordinarily encounter is normal though this normal course is occasionally interrupted by revolutions which change the form, content and direction of the process of the scientific activity, which is basically normal by which we mean a non-revolutionary committed and tradition bound activity. Normal science demands a through going convergent thinking and hence is preceded by an education that involves 'a dogmatic initiation in a pre-established tradition that the student is not equipped to evaluate'. "Normal science is an activity that purports not to question the existing paradigm but to (1) "Increase the precision of the existing theory by attempting to adjust the existing theory or existing observation in order to bring the two into closer and closer agreement", and (2) "to extend the existing theory to areas that it is expected to cover but in which it has never before been tried". In other words, normal science consists of solving puzzles that are encountered in forcing nature into the conceptual boxes supplied by the reigning paradigm.

It is in this way Kuhn attempts to account for the smooth, defined and directional character of day-to-day scientific research in terms of the features of what he calls "Normal Science". Normal Science has no room for any radical thinking. It is limited to the enterprise of solving certain puzzles in accordance with the rules specified by the paradigm. These rules are never questioned but only accepted and followed. The aim of scientific education is to ensure that the paradigm is internalized by a student. In other words, the professional training in science consists in accepting the paradigm as given and equipping oneself to promote the cause of paradigm by giving a greater precision and further elaboration. The day-to-day scientific research does not aim at anything fundamentally new but only at the application of what has already been given, namely the theoretical ideas and the practical guidelines for solving certain puzzles. It is in this sense that **normal science is a highly tradition-bound activity**.

As pointed out earlier, normal science purports to force nature into the conceptual boxes provided by the reigning paradigm by solving puzzles in accordance with the guidelines provided by the paradigm whose validity is accepted without question. During this process of puzzle solving, certain hurdles may be encountered. We than speak of "anomalies". That is, an anomaly arises when a puzzle remains puzzle defying every attempt to resolve it within the framework of the paradigm. But, appearance of one or two anomalies is not sufficient to overthrow a paradigm. The ushering in of the ear of a new paradigm has to be preceded by the appearance of not one or two minor anomalies, but many major ones. In order to declare a paradigm to be crisis-ridden, what is needed is an accumulation of major anomalies. But there is no clear cut and objective criterion to decide which anomalies are major and how many anomalies must accumulate to declare a paradigm to be crisis-ridden. In other words, there is no criterion which decides whether a perceived anomaly is only a puzzle or the symptom of a deep crisis. The issue will be decided by the community of the practitioners of the discipline through the judgment of its peers. Once the scientific community declares the existing paradigm to be crisis-ridden, the search for the alternative begins. Of course the crisis-ridden paradigm will not be given up unless a new theory is accepted in its place. It is only during this transitional period of search for the new paradigm that the scientific debates become radical.

During the process of search for an alternative, the scientific community has to make a choice between competing theories. In this choice, the evaluation procedures of normal science are of no help, "for these depend in part upon a particular paradigm and that paradigm is at issue." The issue concerning the paradigm choice can not be settled by logic and experiment alone. What ultimately matters is the consensus of the relevant scientific community. Kuhn points out, "that question of value can be answered only in terms of criteria that lie outside of normal science altogether, and it is that recourse to external criteria that most obviously makes paradigm debates revolutionary." Thus a paradigm choice cannot be explicated in the neutral language of mathematical equations and experimental procedures, but in terms of specific perceptions which a scientific community as a social entity entertains about what it considers to be the basic values of its professional enterprise. In other words, the ultimate explanation of a theory choice is not methodological but sociological. Hence in Kuhn's scheme, the idea of scientific community as a social entity is axiomatic. That is to say, according to him, "If the term "paradigm" is to be successfully explicated, scientific communities must first be recognized as having an independent existence". This means that one must explain scientific practice in terms of paradigms and paradigmatic changes and the latter are to be explicated in terms of a particular scientific community which shares the paradigms and brings about paradigmatic changes. Thus, the concept of scientific community is basic to the concept of paradigm. The concept of **scientific community** can be explicated in sociological terms. Hence, the ultimate terms of explication of scientific activity are sociological.

What is the relation between the old paradigm, which is overthrown and the new paradigm which succeeds it? Kuhn's answer to this question is extremely radical. According to him, in no obvious sense one can say that the new paradigm is better or truer that the old one. Kuhn maintains that the two successful paradigms cut the world differently. They speak different languages. In fact when a paradigm changes, to put it metaphorically, the world changes. With his characteristic lucidity he says, "the transition from a paradigm in crisis to new one from which a new tradition of normal science can emerge is far from accumulative process, one that is achieved by an articulation or extension of the old paradigm. Rather, it is a reconstruction of the field from new fundamentals, a reconstruction that changes some of the field's most elementary theoretical generalizations as well as many of its,… methods and applications." This apart Kuhn contends that the two paradigms talk different languages. Even if the same terms are used in two paradigms, the terms have different meanings. What can be said in the language of one paradigm can not be translated into the other language. Based on these reasons, Kuhn claims that the relation between two successive paradigms is incommensurable. No wonder Kuhn compares paradigm shift to gestalt shift. With this, the idea of scientific progress as a continuous process and the idea of truth as the absolute standard stand totally repudiated. Kuhn advances what might appear to be an undiluted relativism according to which truth is intro-paradigmatic and not inter-paradigmatic. That is to say what is true is relative to a paradigm and there is no truth lying outside all paradigms.

2.7 POPPER VERSUS KUHN

Some of the radical implications of Kuhn's position can be brought about by juxtaposing his views with those of Popper.

Firstly, the hallmark of science according to Popper is **critical thinking**. In fact science exemplifies critical thinking at its best. Since critical thinking considers nothing to be settled and lying beyond all doubt, fundamental disagreements and divergent thinking must and in fact do characterize science. As we have seen, according to Kuhn, what constitutes the essence of scientific practice is normal science and we have also seen why normal science is a highly tradition-bound activity, an activity made possible by a consensus among the practitioners who share a paradigm. Thus if Popper sees the essence of science in divergent thinking and fundamental disagreements, Kuhn sees the essence of science in convergent thinking and consensus. In other words, the hallmark of science according to Kuhn is tradition-bound thinking. In fact, according to Kuhn, what distinguishes science from other areas of creative thinking is that whereas in science one finds institutional mechanisms of enforcing consensus, the other areas suffer from perpetual disagreements even on fundamentals.

Secondly, if **Popper** considers the individual to be the focus of scientific activity, Kuhn bestows that status upon the scientific community. Both positivists and Popper looked upon science as the sum total of the work of individual scientists working in accordance with a method though the Positivists and Popper fundamentally differed on the characterization of that method. As opposed to this individualistic account of scientific enterprise, Kuhn propounds a collectivistic view of scientific activity. In Kuhn's scheme, it is the scientific community, which constitutes the pillar of stability and locomotive of change. This is borne out by the fact that according to Kuhn the scientific community has institutional mechanism, like peer review, by which it can settle all the issues such as whether an anomaly is a symptom of crisis, how many anomalies suffice to warrant the search for an alternative paradigm, what factors are to be considered in choosing a new theory for the status of the paradigm etc.

Thirdly, Popper and Kuhn differ fundamentally in their attitude towards the transition from one theory to another theory in science. According to Popper, we can explain every case of change of theoretical framework in terms of certain norms which science always adopts and follows meticulously. In fact, scientific rationality consists in following these norms. But Kuhn contends that an adequate explanation of change of theoretical framework must be in terms of the value judgments made by a community while making the choice. According to Kuhn, recourse to the so-called methodological norms explains nothing. From the point of view of Popper, Kuhn is an irrationalist because he sets aside methodological norms and seeks to explain theory change exclusively in terms of non-rational or sociological factors like value commitments of a professional group. Whatever be the merit of Popper's attack on Kuhn as an irrationalist, we can say that Kuhn's construct of scientific practice is sociological. That is to say, according to him, scientific activity can not be understood by trying to find out the absolute standards which have guided the scientific activity in all ages. It can only be understood in terms of the specific judgments which a community makes at a particular juncture regarding what it considers to be its value commitments as a professional group.

The above juxtaposition of Popper and Kuhn brings out the radical implications of Kuhn's views regarding the nature of scientific practice. However, in one respect Kuhn is very close to Popper. Both, like positivists, content that there is something unique to science though they differ in their explanation of what that uniqueness consists of. Positivists maintain that the hallmark of science is the systematic verifiability of its claims. According to Popper, the uniqueness of science consists in the systematic falsifiability of theories. Accounting to Kuhn, it is consensus which marks out science from other areas of human endeavour. That is to say, Kuhn, like Positivists and Popper, does not question whether science is really unique. That is to say, instead of raising critical questions about the status that science has acquired in the contemporary culture, Kuhn only seeks to provide an alternative account of how it has acquired that status. In that sense Kuhn's position is quite conservative.

In this unit, we had a brief look at the 20th century thinking on the nature of science. It is very difficult to decide which view is the correct one, though Positivist view has been shown to be highly inadequate. The question is "How should we practise social sciences so as to make them scientific"?

Some social scientists take the Positivist recommendation: "collect data, extract a generalization, verify the generalization and formulate a law". Those social scientists who are inspired by the Popperian view, take seriously the Popperian advice "Formulate a problem, provide a tentative solution, try to falsify it, if the solution survives treat it as a corroborated theory but not as a confirmed one". Still others go by Kuhn's view of science and think that the task of social sciences today is to arrive at paradigms in different social scientific disciplines. This will enable social sciences to overcome ideological commitments which generate differences even at a fundamental level. According to them, the consensus so generated will bring social sciences close to natural sciences.

Check Your Progress 3

1) Which two stages come in the life of every major science?

.....

.....

.....

.....

.....

.....

2) What do you understand by the term 'Paradigm'?

.....

.....

.....

.....

.....

3) What is scientific revolution?

.....

.....

.....

.....

4) How did Popper differ with Kuhn about essence of science?

.....

.....

.....

.....

5) Match the following:

A	B
1) Individual as the locus of scientific study	i) Positivists
2) Explanation of theory in terms of value judgment	ii) Popper
3) Theory changes in terms of certain norms	iii) Kuhn
4) Verification as the hall mark of science	iv) Popper

2.8 LET US SUM UP

By virtue of shaping our mode of living and our ways of thinking about the world, science has acquired a central place in intellectual world. Three disciplines namely history of science, sociology of science and philosophy of science have made science itself their object of inquiry. Philosophy of science deals with broadly two questions: what is the aim of the science and what are the methods of science? Different scholars have answered these questions differently.

Regarding aims and methodology two views have been dominant. 'inductivism' by Francis Bacon and 'Hypothesism' by Rene Decartes. Positivism has been a movement in the philosophy of the first half of the 20th Century. Inductivitism, systematical verifiable statements, pure scientific observations, unilateral relationship between observation and theory, value neutral theory, precise prediction of facts as aim of science, progressive/objective/rational arguments as essentials of science are some of the important central tenets of positivism. The concept of pure observation in positivist philosophy has been attacked on several grounds.

Karl Popper and Thomas Kuhn are the two most important philosophers of mid 20th century. They offered alternatives to positivism and shaped the contemporary debate on theory of scientific knowledge. Popper rejected the central thesis of positivism i.e., idea of 'pure observation'. He propounded the Hypothetico-Deductive model as an adequate and superior model of scientific method. According to him, the scientific statements are systematically falsifiable. The scientific observations are theory-dependent. Theories are pure inventions of human mind. The theory and observation are interdependent. Science is not value

neutral and therefore values are not subjective. Science is progressive. Science is objective in the sense that its theories are inter-subjectivity testable. Science, according to Popper, is rational in the sense that it embodies critical thinking. Owing to commonality of views of Popper with those of positivists regarding scientism, methodologism, methodological monism and Deductive Nomologism, Popper is sometimes viewed as semi-positivist.

Karl Popper's ideas of scientific progress, verisimilitude and comparison of old theory with new theory have been criticized.

Thomas Kuhn sees the life of major sciences in two stages – Pre-paradigmatic stage and paradigmatic stage. In the pre-paradigmatic stage, more than one mode are practised. In the process of transition a discipline acquires the stage of paradigm. When a paradigm fails to promote fruitful, interesting and smooth normal science, It is considered in crisis. Deepening of crisis leads to replacement of the existing paradigm, which is termed as scientific revolution.

Thus, on the question how we should practise social sciences so as to make them 'scientific, three views have so far been propounded. Positivist recommendation is to 'collect data, extract generalization, verify the generalization and formulate a law'. Popper's view is "formulate a problem, provide a tentative solution, try to falsify it, if solution survives, treat it as a corroborated theory but not as a confirmed one'. Which one out these three is correct – is difficult to decide? Consensus on the three models will bring social science close to natural sciences.

2.9 KEY WORDS

Inductivism : A scientific method of knowledge focusing that scientific knowledge is generated by arriving at generalizations on the basis of pure observations using the principle of induction.

Hypothesisism : Hypothesisism is a method of knowledge which envisages that knowledge is created or generated by putting forth the hypothesis having test implications. Hypothesis is a statement describing the unobservable entities which lie behind observations.

Realists : Those who support Hypothesis are realist in connection with the status of theoretical/ unobservable entities.

Anti-Realists Inductivists are also called as Anti-Realists.

Paradigm : Paradigm refers to the scientific practice comprising of a particular a world view, concept, assumptions, theories, models used by the researcher in carrying out the research studies. The concept was introduced by Thomas Kuhn.

Deductive : Derivation of a conclusion by reasoning or by inference in which the conclusion follows necessarily from the premises.

Methodological Dualism : Referring to positivist belief in separation of the knower from the known or of the subject from object.

Methodological Monism : In contrast to the comparatmentalisation of dualism, monism view the world as a "seamless web". In terms of Gouldner's argument, the separation between the knower and the known must be overcome.

Pseudo Science : Pseudo science refers to any body of knowledge or practice, which purports to be scientific or supported by science but which is judged to fall outside the domain of science.

2.10 SOME USEFUL BOOKS

Mark Blang (1992): *The Methodology of Economics*, Cambridge University Press, London.

Uma Devi. S. (1994): *Economic Theory and Methods of Reasoning*, Har-Anand Publication, New Delhi.

Williams, M., & May, T. (1996): Introduction to the Philosophy of Social Research, London: UCL Press.

Kuhn, Thomas, (1970); *The Structure of Scientific Revolutions.* The University of Chicago Press: Chicago.

Popper, Karl (1972); *Conjectures and Refutations*: *The Growth of Scientific Knowledge*, Rutledge and Kegan Paul: London.

2.11 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) The aim of science is two fold: Definition and Demonstrations.

2) See Sub-section 2.2.1

3) See Sub-section 2.2.1

4)

A	B
i)	iv)
ii)	i)
iii)	ii)
iv)	iii)

5) See Section 2.3

Check Your Progress 2

1) On the basis of falsifiability of scientific theories.

2) Inductivist model's basic proposition is that observations are pure i.e. theory – independent and theories are derived from pure observations. In Hypothetico-deductive model, theories are perceived as our mental (own) constructions and are not the functions of pure observations.

3) Hypothetico-Deductive model's central components are: (i) theories are our own constructions reflecting unobservable entities, (ii) science is progressive in terms of increasing very similitude, (iii) science is rational in terms of critical rationalism and objective in terms of inter-subjectivity.

4) Science embodies critical thinking.

Check Your Progress 3

1) (i) Pre-paradigmatic stage, (ii) Paradigmatic stage

2) The constituents of reality, model problems, range of solutions, necessary strategies and techniques for solving the problems together constitute the paradigm.

3) The process of replacement of existing paradigm is known as 'scientific revolution'.

4) According to Popper, the essence of science lies in divergent thinking and fundamental disagreements whereas Kuhn sees essence of science in convergent thinking and consensus.

5) (1) ii, (2) iii, (3) iv, (4) i.

UNIT 3 MODELS OF SCIENTIFIC EXPLANATION

Structure

3.0 Objectives

3.1 Introduction

3.2 Unified View of Rules of Positivism

3.3 Search for the Criterion of Cognitive Significance

3.4 Rules of Logic or Rules of Correct Reasoning

- 3.4.1 'Categorical Syllogism'
- 3.4.2 Hypothetical Syllogism
- 3.4.3 'Logical Truth' and 'Material Truth'
- 3.4.4 Deductive Fallacies

3.5 Models of Scientific Explanation

- 3.5.1 Hypothetico-Deductive Model
- 3.5.2 Covering-Law Models
  - 3.5.2.1 Deductive-Nomological (D-N) Model
  - 3.5.2.2 Inductive-Probablistic (I-P) Model
  - 3.5.2.3 Critical Appraisal of Covering-Law (D-N,I-P) Models of Explanation

3.6 Explanation in Non-Physical Sciences

3.7 Let Us Sum Up

3.8 Key Words

3.9 Some Useful Books

3.10 Answers or Hints to Check Your Progress Exercises

3.0 OBJECTIVES

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

- Explain the basic rules of formal reasoning;
- State the role of logical reasoning in the formulation of a research problem;
- Describe the basic rules of structural relationship between assumptions, theory and conclusions; and
- Appreciate the problems involved in systematic explanation of phenomena.

3.1 INTRODUCTION

Explanation has three meanings: **One** interpretation of explanation is to remove a perplexity or solve mystery i.e., 'puzzle solving'. The **second** meaning is to change the unknown into the known. The **third** interpretation of explanation is to give causes of phenomenon to be explained. The first two meanings are too relativistic and subjective. It is the third one emphasizing causal explanation

subserving regularity, and law like explanations that fits the meaning of scientific explanation. Here, we are concerned with one mainstream approach to science viz,. Positivism. We may begin this unit by recognizing the complexity of explaining subject matter of philosophy of science. The first two units make it clear, that there is no general agreement on Positivism as the only approach to pursue knowledge. It is at least over fifty years, since when various limitations and misconception of the rules and propositions of Positivism have been brought to light. But yet, paradoxically, the mainstream notion of science and scientific explanation is dominated by the **tenets** of Positivism. This is partly because of the rigorous and prescriptive nature of positivist propositions which remain simple and attractive, and partly because of the open ended and speculative nature of the alternative approaches. The enthusiasm with which many mainstream economists still prefer to refer to Economics as a Positive Science is yet another indication that alternative approaches in the Philosophy of Science have not yet percolated to Economics. Because of these reasons this unit on 'Models of Scientific Explanation', and the next one on 'Debates on the Models of Scientific Explanation in Economics' are essentially confined to Positivist Models of Explanation.

Let us begin with a broad unified view of Positivism in terms of certain rules. It will be followed by certain rules of deduction or formal logical reasoning that underlie Positivist Models of explanation. Afterwards, the various Models of Explanation and their limitations, will be discussed.

3.2 UNIFIED VIEW OF RULES OF POSITIVISM

Our emphasis is on the Models of Scientific Explanation under Positivism. Positivism is procrustean in nature. Its propositions, often contested, evolved over a long period. It passed through several stages, from radical empiricism to logical positivism to logical empiricism. And the propositions of Positivism at various stages varied. Nevertheless, we shall confine here to a unified view of Positivism.

Kilakowski provides such a unified view by simplifying Positivism into a set of four fules:

K_1 The Rule of Phenomenalism

The Rule emphasizes phenomena as the basis of knowledge. Sensory experience is the basis. The basis of knowledge is only the record of that which is actually manifested in experience. It is phenomenon not noumina. It is existence, not essence. Simply, it is facts and objective facts only that form the basis of knowledge.

K_2 The Rule of Nouminalism

Nouminalism refers to Insight representative of facts. Any insight formulated in general terms can not have any real reference other than individual facts. Therefore every abstract science is abridging the recording of experience, and gives no extra independent knowledge.

K_3 The Rule of Value-free Statements

Knowledge is value-free. The rule refutes to call value judgments and normative statements as knowledge. There is no room for good or bad or ethical judgments in statement of science or knowledge.

K_4 The Rule of Unity of Science

There is only one method of scientific knowledge. The Unity of science refers to a single fundamental form from which all other laws are ultimately derived.

3.3 SEARCH FOR THE CRITERION OF COGNITIVE SIGNIFICANCE

In the above summation of Positivism into a set of rules, K_1, Rule of Phenomenalism, may appear to be closer to the early inductivist phase of Positivism. The inductivist view of science emphasized observation of facts and to proceed towards by inductive inference to the formulation of universal laws about these facts. But Positivism underwent considerable sophistication to accommodate deductive reasoning as a part of scientific explanation.

Yet another aspect of Positivism which is important in understanding its approach to scientific explanation is the criterion of cognitive significance. According to Positivism, the basic criterion for distinguishing 'scientific knowledge' or meaningful statements from 'non-scientific' or meaningless statements is **testability.** In the early phase the emphasis was on of **verification** – only complete verification by observational evidence to be considered empirically meaningful. But soon it was found to be too strict, and also faced the problem 'induction'. Karl Popper suggested **falsification** testing instead of verification. But both verification and falsification tests are found to be too strict criteria, and alternatively, Rudolf Camap suggested **confirmation**. Confirmation is a relative concept. Instead of a single test for rejection or acceptance of a hypothesis, a series of positive results are supposed to increase the confirmation of it. Test instances confirmed or disconfirmed hypotheses, and hypotheses could be ranked according to their degree of confirmation relative to the available evidence. Later, emphasis shifted from testing individual theories to evaluation of competing theories on the basis of their relative degree of confirmation. But facts and testing remain basic ingredients of Positivist explanation.

3.4 RULES OF LOGIC OR RULES OF CORRECT REASONING

Just as facts and testing, rules of logic have become important part of the whole structure of scientific explanation under Positivism. Here we briefly review the main stay of deductive reasoning viz. 'syllogism' or 'rules of logic' or 'rules of correct reasoning'.

3.4.1 'Categorical Syllogism'

Two categorical statements if logically formulated, taken together would lead to a conclusion. The two statements should serve as the premises. If there is no premises, then there will be a fallacy or enthymeme.

For example, the following two examples stand for categorical syllogisms:

I)	All As are B	}	→ Major Premis
		} Premises	
	C is an A	}	→ Minor Premis
	Therefore, C is B		→ Conclusion
II)	All Politicians are corrupt	}	→ Major Premis
		} Premises	
	Blogg is a politician	}	→ Minor Premis
	Therefore, Blogg is corrupt		→ Conclusion

The following is an example of a fallacy of enthymeme because it leaves out the premises:

Blogg is a politician.

Therefore, he is corrupt.

3.4.2 Hypothetical Syllogism

Hypothetical Syllogism is referred to as **Modes Poneus**, affirming the antecedent. The major premise has if…….then form.

The following are two examples:

If A is true, (Antecedent)	**then B is true** (Consequent)	Therefore, B is true.
If Blogg is a politician, (Antecedent)	**then he is corrupt** (Consequent)	Therefore, Blogg is corrupt

Hypothesis statements may have a major Hypothesis (H) and auxiliary ($A_1A_{2.....}A_n$) hypotheses.

3.4.3 'Logical Truth' and 'Material Truth'

All statements which are logically true are not necessarily "materially true". For example:

All politicians have two tongues.

Blogg is a politician

Therefore, Blogg has two tongues.

The above statement is "logically true" but "materially" not true.

3.4.4 Deductive Fallacies

Deductive fallacies are those where premises do not lead necessarily to the stated conclusions. There are three main types of deductive fallacies:

1) Logical (formal) Fallacies (Fallacy of affirming the consequent)

Correct Reasoning or 'Modus Poneus'	**'Logical (Formal) Fallacy**
(Affirming the antecedent)	(Affirming the consequent)

If A is true, then B is true.	If A is true, B is true.
A is true.	B is true.
Therefore, B is true.	Therefore, A is true.

2) Verbal Fallacies (Fallacy of Composition)

The verbal fallacy involves a statement where something which is true of the part is also made true of the whole. For example, in case of people waiting to see a procession:

If one rises on one's tip-toes, one can see better.

If people rise on their tip-toes, they can see better.

3) Material Fallacies (All others): Material fallacies are of several kinds.

a) **Post hoc ergo propter hoc:** (After this, therefore because of this).

Because event B occurs after event A, then event B is necessarily caused by event A.

b) "Argument by Analogy": It refers to a statement where, "A is similar to B, therefore whatever is true of A is also true of B". For example, Singapore was backward in 1950s. It developed faster because of "open economy policies". India was backward in 1950s. Therefore, India would have prospered by 'open economy policies'.

c) "Appeal to Authority": It refers to a statement where the truth is sought to be asserted by referring to an authority. "A is true because (so & so) say it so". For example, "India is shining because Milton Friedman has said so".

Check Your Progress 1

1) What do you mean by the term 'Scientific explanation'?

.....

.....

.....

.....

2) What is the difference between Phenomenalism and Nominalism.

.....

.....

.....

.....

3) State the rules of logic. Explain 'categorical syllogism'.

.....

.....

.....

4) What is deductive fallacy?

.....

.....

.....

.....

3.5 MODELS OF SCIENTIFIC EXPLANATION

"Men do not think they know a thing till they have grasped the 'why' of it"

- Aristotle

In the light of understanding of the basic building blocks of knowledge viz., facts, testing and formal logic, we shall now turn to the Positivist Models of Scientific Explanation. The basic building blocks of scientific explanation are theories. The status, structure and functions of theories and theoretical terms become essential part of scientific explanation. The aim of scientific research is not only to discover and describe events and phenomena in the world but also, and more importantly, to explain and understand **why** these phenomena occur as they do.

The early Positivists like Aguste Comte and Ernest Mach did not believe in any role to explanation in science. Mach believed that theories are mere heuristic devises and talked of theories being "eliminative fiction" useful only in organizing data. But Positivism has traveled a long distance since, and by mid-1950s was referring to explanation as the Chief objective of Science. "To explain the phenomena in the world of our experience, to answer the question "Why?" rather than only the question "What?" is one of the foremost objectives of all rational enquiry…". Explanation removes puzzlement and provides understanding. However, there exist considerable differences of opinion as to the function and the essential characteristics of scientific explanation.

What follows is the description of some of the basic models of scientific explanation within the positivist approach. First, we shall discuss the hypothetico-Deductive Model (H-D Model), and then turn to Covering Law Models encompassing the Deductive-Nomological (D-N) and Inductive-Probablistic (I-P) Models.

3.5.1 Hypothetico-Deductive Model

At a time when the Positivists were not ready to concede any role to theoretical explanation, **Carnap and Hempel**, in their writings, came out with a model – which later came to be known as **Hypothetico-Deductive (H-D) model** of explanation. The H-D model not only describes the structure of theories, but provides answers to the questions of the status and functions of theories, as well. H-D model explicitly addresses the problems of a theory's structure. The propositions in a deductive system are seen as being arranged in an order of levels. Higher-level hypotheses will refer to theoretical entities, all which need not be tested. Lower-level hypotheses describe observable phenomena and are tested against reality. H-D model addresses the problem of status of theoretical

terms which may not be testable directly, to gain meaningfulness indirectly by the successful confirmation of theory in which they are embedded.

In H-D model, theories as a whole are tested by comparing their deduced consequences (prediction) with data. Every theoretical term need not be given empirical counterparts via correspondence rules. H-D model also turns the old **realist-instrumentalist controversy** into a moot debate. **Realists** claim that all theoretical terms must refer to real entities and theories, which do not, are false. Instrumentalists insist that theories are only instruments for predictions. Only relevant question for them is whether theories are adequate for prediction. H-D model seems to accommodate both these concerns and defuse the controversy.

The H-D model rejects the earlier notion that theories have no role but are only instruments. H-D model emphasizes the following positive functions of theories:

1) They allow generality in the specification of scientific laws.

2) They possess 'a certain formal simplicity' which allows the use of 'powerful and elegant mathematical machinery'.

3) They can serve the practical function of allowing the scientist to discover interdependencies among observable.

4) They are convenient and fruitful heuristic (intellectual) devices, often serving an explanatory function of their own.

Thus, the H-D model without minimizing the importance of observable phenomena for scientific explanation, allowed far more substantial role for theories and theoretical terms than did their predecessors.

3.5.2 Covering-Law Models

Carl G. Hempel and **Paul Oppenheim** developed Deductive-Nomological (D-N) Model of scientific explanation and much later, Hempel extended it to include Inductive-Probablistic (I-P) Model. These two models together are referred to as Covering-Law Models of explanation.

3.5.2.1 Deductive-Nomological (D-N) Model

Hempel and Oppenheim in their paper entitled "Studies in the Logic of Explanation" advanced an account of scientific explanation, which later came to be known as deductive-nomological (D-N) model. They divided an explanation into two major constituents – the **explanandum** and the **explanans**. The **Explanandum** means the sentence describing the phenomenon to be explained (not that phenomenon itself). **The Explanans** mean the class of those sentences, which are adduced to account for the phenomenon. The explanans fall into two sub-classes, viz., certain antecedent conditions $C_1, C_2 \ldots\ldots C_k$ and certain general laws $L_1, L_2 \ldots\ldots L_3$. If a proposed explanation is to be sound, conditions of adequacy must be satisfied. The following four conditions of logical and empirical adequacy must be satisfied:

i) **Logical Conditions of Adequacy**

R_1) The explanandum must be a logical consequence of the explanans; in other words, the explanandum must be logically deduced from the

information contained in the explanans, for otherwise, the explanans would not constitute adequate grounds for the explanandum.

R_2) The explanandum must contain general laws, and these must actually be required for the derivation of the explanandum. Though not a necessary condition, the explanans must contain at least one statement which is not a law.

R_3) The explanans must have empirical content, i.e., it must be capable, at least in principle, of test by experiment or observation. The point deserves special mention because certain arguments which have been offered as explanations in the natural and in the social sciences violate this requirement.

ii) **Empirical Condition of Adequacy**

R_4) The sentences constituting the explanans must be true.

The characteristics of explanation discussed may be summarized into the following schema:

$\{C_1, C_2 \ldots\ldots C_K$ Statements of Antecedent$\}$

Logical Deduction

$\{$ conditions $\}$ Explanans

$\{$ $\}$

$\{L_1, L_{2\ldots\ldots} L_r$ General Laws $\}$

$\{E$ Description of the empirical phenomenon to be explained $\}$ Explanandum

Hempel and Oppenheim argued that the same formal analysis, including the four necessary conditions, applies to scientific predictions as well as to explanation. They went on to argue that an explanation is not fully adequate unless its explanans, if taken account of in time, could have served as a basis for predicting the phenomenon under consideration. Consequently, whatever is said about the logical characteristics of explanation or prediction will be applicable to either, even if one of them is mentioned. This resulted what is known as Covering-Law Model extended to the D-N Model of Explanation.

Extension of characteristics of explanation to prediction could formally be presented in the following schema:

$\{C_1, C_2 \ldots\ldots C_K$ Statements of Conditions$\}$

Logical Deduction

$\{$ $\}$ Predecens

$\{$ $\}$

$\{L_1, L_{2\ldots\ldots} L_r$ General Laws $\}$

$\{P$ Description of the empirical Prediction $\}$ Predecendum

The symmetry visualized between explanation and prediction has often drawn criticism.

3.5.2.2 Inductive-Probabilistic (I-P) Model

In 1960s Hempel developed Inductive-Probabilistic Model. In the I-P model the explanans consists of sentences describing the requisite initial conditions along with statistical laws and is expected to confer upon the explanandum highly logical, or inductive probability. The characteristics are similar to D-N Model. The symmetry holds and hence another Covering-Law Explanation.

Check Your Progress 2

1) Identify the basic blocks of knowledge.

.....

.....

.....

.....

2) Explain how Hypothetico-Deductive Model defuses the conflict between the realists and instrumentalists.

.....

.....

.....

.....

3) State two characteristics of Deductive-Nomoligical (D-N) Model.

.....

.....

.....

.....

.....

3.5.2.3 Critical Appraisal of Covering-Law (D-N, I-P) Models of Explanation

The Covering-Law Model of explanation offered by the D-N and I-P models are subject to a number of limitations. For convenience we shall confine here to the limitations of D-N model, some of which are also applicable to I-P model.

First is about the necessity of laws and theories in the pursuit of scientific explanation. Laws are distinct from ordinary statements. Ordinary statements or statement of mere facts refer to particular statements and express "contingent truths:, like "John went home and had dinner". Laws refer to universal statements and express "necessary truths". For example, "all objects are attracted to one another with a force inversely proportional to the square of the distances between them" is a universal and law like statement: However, there are many apparently

universal statements which are not really universal statements. These may be the ones that involve mere "accidental generalizations" like, "all coins in my pocket are nickels".

This distinction between laws and ordinary statements does not make laws as necessary requirement for explanation. Particularly in biology, astronomy and social sciences explanations need not be in the form of laws. Explanations may be in the form of questions and answers. The often cited example is: Why do Hopi Indians continue to perform rain making ceremonies even though their ceremonies do not normally produce desired result? The explanation is to reinforce group identity, but there is no law involved. Similarly: Why do all vertebrates' hearts beat regularly? The explanation is to circulate blood throughout the body which helps to maintain uniform temperature. This explanation again doesn't require any law like statement.

Second, explanation requires not only the factors one cites to be casually relevant but also that they be causes. D-N model is often satisfied by citing relevant causes. Causes explain effects, but effects do not explain their causes, and effects of common cause do not explain one another. Based on David Hume's famous guillotine of time, two contiguous events do not become cause and effect. Inferring event B as the effect of event A because it followed the later, amounts to the "Billiard Ball Model of Causation", where the ball reaching the pouch is seen, 'as if' the player knew all the laws of time and motion of the billiard ball.

Third, much of the criticism of Covering Law Model of explanation relates to the symmetry thesis, that explanation and prediction are inherently linked, and that only difference is the sequence. But critics point out that explanation need not predict. For example, Darwin's Theory of Evolution provides a great deal of explanation on the origin and evolution of species, but it does not predict anything. Similarly, prediction need not explain anything. For instance, prediction of weather does not explain anything.

Fourth, the D-N Model ignores the divergence between the norms and practices and thereby eliminates much of science from the ambit of explanation. Many functional or technological explanations get eliminated because they are not through any laws or regularities. Purposive behaviour needs explanation of motivation and in motivated behaviour future affects the present. Determining motives to be classified in antecedent conditions and not in causal relation. Hempel and Oppenheim also recognize the problems in applying their model of explanation in social sciences. Human behaviour often is characterized by uniqueness and irrepeatability. It depends upon situation and previous history of individuals.

Notwithstanding all these limitations, "many philosophers would still maintain that the D-N Model is an important starting point for studying scientific explanation and there is something right and important about it." (Hausman, 1994, p.9). Particularly in the mainstream economics, scientific explanation is strongly associated with Positivist models of explanation and thus D-N Model becomes an important reference point for the debates on explanation in economics.

3.6 EXPLANATION IN NON-PHYSICAL SCIENCES

There is a view that explanation in biology, psychology, and the social sciences has the same structure as in the physical sciences. But the causal type of explanation is essentially inadequate in fields other than physics and chemistry. And it is more so in the case of social sciences which involve explanation of purposive behaviour.

Hempel and Oppenheim point out to some of these limitation of applying models of scientific explanation to social sciences. First, events involving the activities of humans singly or in groups have a peculiar uniqueness and irrepeatability which make them inaccessible to causal explanation. Causal explanation presupposes uniformities and repeatability which may not hold for phenomena in social sciences.

Second, the establishment of scientific generalizations and explanatory principles for human behaviour is impossible because the reactions of an individual in a given situation depend not only upon that situation, but upon the previous history of the individual.

Third, the explanation of any phenomenon involving purposive behaviour calls for reference to motivations and thus for teleological rather than causal analysis. Many of the explanations which are offered for human actions involve reference to goals and motives. Motivational explanations often tend to have less cognitive significance.

Attempts to extend scientific models of explanation to non-physical sciences without closer understanding of the context often lead to inappropriate explanations. One possibility is that there are contexts or areas within social sciences like economics where there is room for application of scientific or to be specific, positivist models of explanations. But at the same time there are vast aspects of economics where such application may not be easily amenable.

Check Your Progress 3

1) State the difference between laws and ordinary statement.

.....

.....

.....

.....

2) Describe two limitations of application of scientific models to social sciences.

.....

.....

.....

.....

3.7 LET US SUM UP

Positivism as an approach to the pursuit of knowledge is procrustean in nature. It has passed through several stages from radical empiricism to logical positivism to logical empiricism. Its propositions, often contested, evolved over a long period. The main characteristics (unified view) of positivism lie in a set of its four rules-phenomenalism, nouminalism, the rule of value free statements and the rule of unity of science.

Testability under the positivism has been basic criteria for distinguishing scientific knowledge from non-scientific or meaningless statement. In early phase, emphasis was on verification. Later on Karl Popper suggested falsification testing in place of verification. Subsequently, emphasis shifted from falsification to confirmation.

Rules of logic have become important part of the whole structure of scientific explanation under positivism. There are two types of deductive reasoning-categorical syllogism and hypothetical syllogism.

Theories are the basic building blocks of scientific explanation. Hence the status, structure and functions of theories and theoretical terms are essential parts of scientific explanation.

The basic modes of scientific explanation within the positivist approach are: Hypothetico-Deductive Model (HD), Deductive-Nomological (DN) Model, and Inductive Probabilistic (IP) models. The DN and IP models are subject to several limitations. The application of scientific models to social sciences has several limitations; peculiar uniqueness of human activities, difficulty in scientific generalizations and explanatory principles for human behaviour, etc. Hence applications of scientific models to non-physical sciences without closer understanding of the context often lead to inappropriate explanation.

3.8 KEY WORDS

Deduction : Deduction is a process used to derive particular statements from general statements.

Empiricism : 'Empiricism' refers to distinctive approach that envisages that all knowledge comes from the senses.

Falsification : The concept developed by Karl Popper (1959) is a philosophy of Science also known as critical rationalism. It uses the logic of deduction to provide the foundation for the Hypothetico-Deductive Method. He rejected the idea that observation provides the formulation for scientific theories. Theories are invented to account for observation, not derived from them. Observations are used to try to reject false theories. Theories that survive this critical process are provisionally accepted but never proven to be true.

Induction : Induction is a logical process of moving from particular statements to general statements. This is used in the social sciences to produce theory from data.

Logical Positivism : An approach to philosophy of science that envisages that all meaningful statements are either empirical or logical in character. This means that they are either open to test by observational evidence or are matches of definition and therefore, tautological. According to this approach, no fundamental difference exists between natural and social sciences.

Modelling : Refers to a process of creating a simplified representation of a System or phenomena with hypotheses to describe the system or explain the phenomenon.

Theory : Refers to set of general laws that are related in networks and are generated by inductive or deductive form of logic. In other words, theory can be described as a set of propositions of different levels of generality.

Verification : Refers to the use of empirical data, observation, test or experiment to confirm the truth or rational justification of a hypothesis.

3.9 SOME USEFUL BOOKS

Brody, Baruch A. (1970); *Readings in the Philosophy of Science*, Prentice Hall Inc., Englewoodcliffs, New Jersey.

Bryant, C.G.A (1985); *Positivism in Social Theory and Research,* London, Macmillan.

Caldwell, Bruce J (1984); *Beyond Positivism: Economic Methodology in the Twentieth Century,* George Allen and Unwin, Boston.

Hausman, D.M (1994); *The Philosophy of Economics: An Anthology,* Second Ed., Cambridge University Press, Cambridge.

Popper, Karl (1959); *The Logic of Scientific Discovery,* Basic Books, New York.

Stewart, F. (1979); *Reasoning and Method in Economics,* McGraw-Hill Book Co., London.

3.10 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Scientific explanation refers to give causes of phenomenon to be explained.

2) Phenomenon refers to a situation in which facts and objective facts based on experience forms the basis of knowledge. It is existence, not essence. On the other hand, nouminalism is insight representative of facts. It envisages that any abstract concepts used in scientific explanation must also be derived from experience.

3) Broadly there are two rules of logic. Categorical syllogism, and hypothetical syllogism.Logical formulation of two categorical statements serving as premises is categorical syllogism.

4) See Sub-section 3.4.4.

Check Your Progress 2

1) Facts, testing and formal logic.

2) See Sub-section 3.5.1

3) See Sub-section 3.5.2.1

Check Your Progress 3

1) Laws refer to universal statements expressing necessary truths whereas ordinary statement refers to particular statements of mere facts and express 'contingent truths'.

2) See Section 3.6

UNIT 4 DEBATES ON MODELS OF EXPLANATION IN ECONOMICS

Structure

4.0 Objectives

4.1 Introduction

4.2 Classical Political Economy and Ricardo's Method

- 4.2.1 Ricardo's Method
- 4.2.2 N.W.Senior
- 4.2.3 J.S.Mill
- 4.2.4 J.E.Cairnes
- 4.2.5 J.N.Keynes

4.3 Robbins, Positivism and Apriorism in Economics

4.4 Hutchison and Logical Empiricism in Economics

4.5 Milton Friedman and Instrumentalism in Economics

- 4.5.1 Goal of Science
- 4.5.2 Relationship Between Significance of Theory and 'Realism' of Assumptions
- 4.5.3 Critical Appraisal of Friedman's Instrumentalism

4.6 Paul Samuelson and Operationalism

- 4.6.1 Operationalism
- 4.6.2 Descriptivism

4.7 Theory – Assumptions Debate in Economics: A Long View

4.8 Amartya Sen on Heterogeneity of Explanation in Economics

- 4.8.1 Heterogeneity of Substance and Methods of Economics
- 4.8.2 Testing and Verification
- 4.8.3 Value Judgements and Welfare Economics
- 4.8.4 Formalisation and Mathematics

4.9 Let Us Sum Up

4.10 Key Words

4.11 Some Useful Books

4.12 Answers or Hints to Check Your Progress Exercises

4.0 OBJECTIVES

The unit aims at tracing the broad contours of the debate relating to bringing explanation in Economics closer to the Models of Positive Science. Patterning explanation in Economics on the lines of Positivist Models of Explanation has been a major methodological obsession in Economics. But the subject matter of Economics is not as easily amenable as physical science for application of rules of Positivism. After going through this unit, you will be able to:

- state the efforts made by mainstream economists to adopt the model of explanation of Positivism;

- explain the limitations, difficulties and differences in adopting 'scientific' models of explanation; and
- appraise the present state of explanation in Economics.

4.1 INTRODUCTION

The urge to model Economics or Political Economy as a science on the lines of physical sciences dates back to the period of Classical Economics. David Ricardo claimed that the principles of Political Economy were as accurate as the laws of gravity. Though Ricardo did not write much on the methods or models of explanation, his theories of political economy reflected **abstract-deductive thinking** which was claimed later as constituting basic methodology of Political Economy. It is 'Ricardo's habit of thinking' equated with hypothetico-deductive model of explanation that set the tone for the classical writers on methodology to carry forward the view of Political Economy as a pure science. We shall begin by briefly examining the views on the nature of Political Economy and its Method of Explanation in the Ricardian tradition as enunciated in the writings of N.W.Senior, J.S.Mill, J.E.Cairnes and J.N.Keynes. Then, we shall turn to Lionel Robbins who asserted that Economics was a pure positive science and emphasized apriorism as its method of explanation. This will be followed by T. Hutchison's insistence on empiricism if Economics were to be called science at all. The next two sections will discuss the foundations for present day mainstream practice of Economics in the form of 'instrumentalism' and 'operationalism' in the writings of Milton Friedman and Paul Samuelson respectively. Throughout, the focus is on the limitations of these approaches and the ad-hoc means by which problems are sought to be overcome. Though the unit is confined to the explanation in Economics in the Positivist mainstream, towards the end, a brief discussion of the alternative methods of explanation will be presented with a view to expose the student to the potential of plurality of methods of explanation in Economics.

4.2 CLASSICAL POLITICAL ECONOMY AND RICARDO'S METHOD

4.2.1 Ricardo's Method

The right method of economic enquiry as such was not discussed by Adam Smith. Smith's approach to methodology was, as it was analysed by later economists, somewhat *enigmatic*. There are instances which make some to point out that he was the first to raise political economy to the state of deductive science. But in the same breath there are many instances in his writings to as the founder of inductive historical method. Yet his overall approach often tends to be moral and ethical, distancing him from the attributes of any positivist explanation. For these reasons, on issues of methodology of Classical political economy, one invariably turns to David Ricardo.

Ricardo did not write on methodology of economics but his method explaining the principles of political economy is clearly discerned as positive, abstract and deductive, and it has come to be called '**hypothetico-deductive model of explanation'**. It vigorously denies that facts can speak for themselves and contends that it is theory that can make sense. Ricardo died in 1823, leaving behind the claim that political economy was a science. His **'Principles of Political**

Economy', which was a set of laws based on deductive reasoning, measures up to that claim. Questions were raised soon whether a deductive body of laws without empirical content could claim the status of science. The task of defending Ricardo's method and the claim of political economy as a science was left to his followers of classical political economy. The defence came from the writings of N.W.Senior, J.S.Mill, J.E.Cairnes and J.N.Keynes.

4.2.2 N.W. Senior

It was N.W.Senior in his 1927 lecture, which was later elaborated as **outline of the Science of Political Economy,** (1936), that first made a statement on the distinction between a pure and strictly positive science, and an impure and inherently normative *art* of economics. He went on to claim scientific status for the Ricardian system. Science of Political Economy, according to him, rests essentially on "a very few general propositions, which are the result of observation, or consciousness, and which almost every man, as soon as he hears them, admits, as being familiar to his thoughts". From this conclusions are drawn which hold true only in the absence of "particular disturbing causes".

Senior reduced these "very few general propositions" of political economy into the following four:

1) that every person desires to **maximize wealth with as little sacrifice as possible**;
2) that population tends to increase faster than the means of subsistence;
3) that labour working with machines is capable of producing a positive net product; and
4) that agriculture is subject to diminishing returns.

4.2.3 J.S. Mill

It is senior, who for the first time introduces the concept of '**economic man**' or 'maximizing man', as could be seen from the first proposition above. J.S.Mill, though staunch advocate of inductive method and a stout defender of methodological monism in science, was nonetheless sympathetic to Ricardo's deductive method of political economy. Mill, in his essay *on the Definition of Political Economy and on the Method of Investigation Proper* to It (1936), treats political economy as an "inexact science: and a 'separate science:, and thereby defends deductive method in economics. Elaborating on Senior's notion of 'economic man', Mill observes "political economy proceeds….under the supposition that man is a being who is determined, by the necessity of his nature, to prefer a greater proportion of wealth to a smaller in all cases…" Mill does consider "economic man" as fictional man to facilitate the propositions of political economy. The pages on 'economic man' in Mill's essay are followed immediately by the characterization of political economy as "essentially an abstract science; that employs "the method of a priori". The "disturbing causes" and "the tendency laws" in political economy necessitate *ceteris paribus* clauses.

4.2.4 J.E. Cairnes

J.E.Cairnes in his **Character and Logical Method of Political Economy (1875),** goes on to argue that political economy is a hypothetical, deductive science, and its conclusions "will correspond with facts only in the absence of disturbing

causes, which is, in other words, to say that they represent not positive but hypothetical truths". By late 19th century there were growing doubts on the method of political economy. J.N.Keynes tried to argue that what is referred to as deductive reasoning in political economy was indeed preceded by inductive inferences based on observation of human behaviour.

4.2.5 J. N. Keynes

J.N. Keynes, in **The Scope and Method of Political Economy (1890),** argued that deductive reasoning was indeed preceded by concrete observations. Deductive reasoning comes at a later stage drawing upon specific individual experiences. J.N.Keynes is known for comprehensive reconciliation of the classical methodological approaches.

The following observation stands as a testimony to his comprehensive understanding of the methodological approaches underlying a complex subject like economics. He observes "…economic science deals with phenomena that are more complex and less uniform than those with which the natural sciences are concerned; and its conclusions, except in their most abstract form, lack both the certainty and the universality that pertain to physical laws. There is a corresponding difficulty in regard to the proper method of economic study; and the problem of defining preconditions and limits of the validity of economic reasonings become one of exceptional complexity. It is, moreover, impossible to establish the sight of any one method to hold the field to the exclusion of others. **Different methods are appropriate, according to the material available, the stage of investigation reached, and the object in view; and hence arises the special task of assigning to each its legitimate place and relative importance".** One could read this even to this day, as a manifesto of methodological pluralism, that is increasingly thought of as appropriate for economics. This aspect is further discussed, when we turn to Sen on economic methodology.

Check Your Progress 1

1) Explain the characteristics of Ricardo's method of explanation.

.....

.....

.....

2) State the general prepositions advanced by N.W.Senior in support of scientific status of the Ricardion system.

.....

.....

.....

3) What is methodological pluralism?

.....

.....

.....

.....

4.3 ROBBINS, POSITIVISM AND APRIORISM IN ECONOMICS

Lional Robbins in his classic tract, *An Essay on the Nature and Significance of Economic Science (*1935 ed.) presented the dominant methodological view point of the early twentieth century economic thought, which stressed **subjectivism, methodological individualism, and the self-evident nature of the basic postulates of economic theory**. His objective was to counter the criticism from the historical and institutional schools that basic economic propositions were not amenable for Quantification, therefore not testable and hence not eligible to be called a science. Robbins countered by suggesting that basic propositions of Economics were deductions from a series of postulates. Chief of these postulates include the assumption involving the experience of the way in which scarcity of goods shows itself in the world of reality and how individuals order their preferences. These are universally acknowledged facts of experience. Robbins offers a coherent account of the status of terms used in economic theory. The fundamental assumptions are combined with subsidiary hypotheses which allow us to apply the theory to actual situations.

For Robbins rationality as a postulate in economics doesn't merely mean maximizing behaviour, but implies consistency of choice. Rationality as consistency of choice is stated in the sense that if A is preferred to B and B to C, A will be preferred to C. Robbins asserts that the assumption of rational conduct, and with it the assumption of perfect foresight in Economics, are 'expository devices'. They are not realistic assumptions. According to Robbins, empirical studies may be of use for the short-term prediction of possible trends, but they do not provide the grounds for discovering 'empirical laws'. The proper uses of realistic (empirical) studies are three: to check on the applicability of theoretical constructions to particular concrete situations, to suggest auxiliary postulates to be used with the fundamental generalizations, and to bring to light areas where pure theory can be reformulated or extended.

Robbin's insistence on Economics being based on true postulates born out of experience, which need not be tested, brings him closer to *apriorism* of the Austrian School. Economic laws are derived from postulates based on a *priori* truths of experience and therefore there is no need for testing them. Robbins' claim that Economics is a pure theory and still a Positive science may be somewhat puzzling. His claim of Positivism for Economics appears to be based more on his insistence that Economic laws are value- free. He argued that there was no room for ethical considerations in Economics, and in that sense too it is 'pure theory'.

Robbins' claim of 'pure theory' or apriorism status to Economics provoked devastating criticism, especially from T. Hutchison. The point of criticism is the 'emptiness of the propositions of pure theory'. If Economics is to be called as a positive science, Hutchison insisted that it should be testable. Otherwise it would end up as a '*pseudo-science'*. Second, the assumption of '*perfect expectations'* is baseless and therefore the rationality postulate is unrealistic. Third, there is need for more extensive use of empirical techniques in Economics instead of merely claiming all postulates are prior truths. Finally, Hutchison calls the use of 'psychological method' (or introspection) as ground for asserting the fundamental postulates.

4.4 HUTCHISON AND LOGICAL EMPIRICISM IN ECONOMICS

Terence Hutchison's *The Significance and Basic Postulates of Economic Theory* (1938) marks a clear turning point in the Twentieth Century Economic methodology and laid the foundation for a clear logical positivist turn to the practice of Economics. Hutchison, in his book, sets out on search for and making clear the foundations of modern economic theory. He states that economics is a science and as such it must appeal to facts, otherwise one would be engaging in 'pseudo-science'. What sets apart the empirical propositions of science from those of other intellectual endeavours is their testability.

According to Hutchison, science contains statements which are either conceivably falsifiable by empirical observation or are not. Those which are not falsifiable are tautologies and are thus devoid of empirical content. The primary reason why the propositions of pure theory have no empirical content is that they are posed in the form of deductive inferences. In Economics, widespread use of *ceteris paribus* clause robs empirical content. Irresponsible use of *ceteris paribus* clause make microeconomic theory effectively unfalsifiable.

Hutchison examines the formal structure of economic theory which consists of a series of deductions from basic postulates. These deductions are analytical in nature. These analytical statements are logical statements without any empirical content. Therefore, pure economic theory is empty and thus there is need for more empirical content. Hutchison insisted on the need for testability, and falsification tests even for the basic assumptions. The insistence of testing not only the theory, but the assumptions as well, for validity sparked-off a wider debate in economics later. But the immediate response was to describe Hutchison as an 'ultra-empiricist'.

Hutchison's attempt to establish the analyticity of fundamental generalizations of economic theory and to wean economists away from pure theory did not succeed very much. But his proposal to make economics increasingly testable or falsifiable form; to increase empirical investigations of various aspects of economics, and the need for economists to abandon their psychological method received universal acceptance by economists.

One of the interesting aspects of Hutchison's contribution relates to the issue of testing assumptions as well as the theory or hypothesis. This provoked a debate between Hutchison and Fritz Machlup. Machlup believes that only deduced or 'lower level' hypotheses require 'verification', and he outlines how such testing could be carried out. While Hutchison does not require that every statement in a theory be tested, he does insist that each be 'testable' and one should be able to conceive how a test could be carried out. Further, Hutchison prefers that the behavioural postulates of economics reflect the actual observed and statistically recorded behaviour of economic agents. Machlup requires no such correspondence. The crucial behavioural postulate is the assumption of rational, maximizing behaviour. And Machlup is more reasonable in arguing that this assumption need not be tested directly, which was insisted by Hutchison, but indirect testing of the same could be done. Much before the debate on testing of assumptions sparked-off by Hutchison's contribution, Milton Friedman's

independent contribution suggested that testing of assumptions as such was not an important issue. However, such a proposition did lead to much controversy as we shall see in the next section.

4.5 MILTON FRIEDMAN AND INSTRUMENTALISM IN ECONOMICS

Milton Friedman's "The Methodology of Positive Economics", the lead article in his book *Essays in Positive Economics* (1953), is among the best-known pieces of methodological writings in economics. For over two decades the methodological prescriptions advanced in his essay became widely accepted among many working economists, in spite of wide controversy it generated. Rightly, as observed by Caldwell, it was a remarkable masterpiece of marketing certain notions of explanation as the basis of the methodology of positive economics. In what follows, we shall discuss in brief Friedman's contribution including some of the major points of criticism. Friedman sets out by stating that the purpose of his essay is to counter criticism of two pillars of neo-classical, economics viz.,1) the maximization behaviour and 2) The model of perfect competition. He goes on to establish the methodological foundations of neo-classical economics by posting his own version of positivism which later came to be branded as "instrumentalism".

4.5.1 Goal of Science

According to Milton Friedman, the ultimate goal of positive science "is the development of a 'theory' or 'hypothesis' that yields valid and meaningful …..predictions about the phenomenon". The criteria for acceptability of theory or hypothesis are three:

1) that theory / hypothesis be logically consistent and contain categories which have meaningful empirical counterparts;

2) that there are 'substantive hypotheses' which are testable; and

3) that "the only relevant test of the validity of a hypothesis is comparison of its predictions with experience".

Further, when an infinite number of hypotheses generally consistent with an observed set of facts generally exist, other criteria-simplicity and fruitfulness-be invoked to chose among competing hypotheses.

4.5.2 Relationship Between Significance of Theory and 'Realism' of Assumptions

Friedman is critical of those who instead of testing a theory for predictability, try to evaluate theory by the 'realism of their assumptions? For Friedman, disparaging a theory for having 'unrealistic assumptions' is ridiculous. He tries to show that most significant theories are actually characterized by descriptively inaccurate assumptions. He observes: "Truly important and significant hypotheses will be found to have "assumptions" that are widely inaccurate, descriptive representations of reality, and, *in general, the more significant the theory, the more unrealistic the assumptions*". (emphasis added).

Realism of assumptions does not matter except when certain hypotheses are derived from assumptions, and when there is no test available, realist description is in order. With these two exceptions, realism of assumptions do not matter for the following reasons:

1) Theory is supposed to lead to simplification of complex reality by abstraction. There is inverse relationship between realism and abstraction. The more realism one insists upon, the more abstract and the complex the explanation.

2) The uninformativeness of knowledge of discrepancies between assumptions and facts, makes one to treat realism of assumptions as of no importance. He cites the example of Galileo's Law of Falling Bodies:

   "*If* a body falls toward the earth in *vacuum*, its instantaneous acceleration is constant, $s=\frac{1}{2}gt^2$. The realism of vacuum need not be tested. Since the prediction is true, then the assumption be treated as *if* true. Turning to economic theory, Friedman cites the "Billiard Ball Player" example and where once the ball reaches the pouch one could assume as if the player knew all the laws of time and motion and extends it to profit maximization: 'under behave as if they were seeking rationally to maximize their expected returns……and have full knowledge of the data needed to succeed in this attempt".

3) The presence of 'undesigned' classes of implications can be avoided by *as if* assumption.

Milton Friedman concludes his essay by asserting unity of method in all positive sciences. For him there are no methodological differences between natural and social sciences. He regards economics as "objective" as physical sciences. His main conclusions are:

1) that realism of assumptions is "largely" irrelevant to validation of theories, which ought to be judged 'almost' solely in terms of their instrumental value in generating accurate predictions.

2) standard economic theory has an excellent predictive record as judged by countless application to specific problems.

3) the dynamics of competition over time accounts for this splendid track record.

4.5.3 Critical Appraisal of Friedman's Instrumentalism

Milton Friedman's methodological approach is criticized widely. The major criticism relates to the proposition that realism of assumptions does not matter. Friedman's assertion that not only the realism of assumptions is necessary but "in general, the more significant the theory, the more unrealistic the assumptions" is called by Samuelson as a flamboyant exaggeration, as the extreme version of "F-Twist" (meaning Friedman-Twist). It is one thing to say that proving realism of all assumptions is difficult, but altogether different to say that further the assumptions move away from realism, closer is the theory to predictions. Even Fritz Machlup, L. Qsoland and Nagel, who were generally in defence of Friedman's methodology, would insist on the realism of at least second order assumptions. It is pointed out that direct evidence about assumptions is not

necessarily more difficult than testing a theory. Test assumptions may yield important insights into the theory. Caldwill points out that philosophically, Friedman's insistence on prediction not explanation may result in correlation not causation, i.e., it may result in "measurement without theory". Blaug points out that accurate predictions are not the only relevant test of the validity of a theory and it would be impossible to distinguish between genuine and spurious correlations. Further, Friedman's instrumental attitude to theories, ignores 'truth value' of theories as many philosophers of science would insist upon.

Check Your Progress 2

1) Identify the important features of methodological view point of L. Rabins.

.....

.....

.....

2) What is the important contribution of Hutchison in the area of logical empiricism?

.....

.....

.....

3) Which criteria have been laid down by Milton Friedman for acceptability of theory or hypothesis?

.....

.....

.....

4) Identify the basis of the major criticism against Friedman's instrumentalism proposition.

.....

.....

.....

4.6 PAUL SAMUELSON AND OPERATIONALISM

Paul Samuelson's methodological contribution, though not comparable to his substantial work in economic theory, does offer certain insights into the turn of economics towards logical empiricism. Samuelson is against a **priorism**. He is more akin to logical empiricism of Hutchison's kind. He does show the influence of Popper's variety of 'rational reconstruction'. His contribution to economic methodology could be seen in terms of two theses, which he advances. One has come to be described as 'operationalism' and the other as 'descriptivism'. We shall elaborate these two theses.

4.6.1 Operationalism

According to him, economists should seek to discover 'operationally meaningful theorems'. Theories are strategically simplified description of observable and refutable empirical regularities. He begins with the criticism of Friedman for the latter's F-Twist on realism of assumptions and emphasizes the need for methodological clarity. He proposes the thesis of "logical equivalence" as the basis for methodological consistency. According to "logical equivalence:, theories are merely equivalent restatements of assumptions and conclusions, i.e., A=B=C, where A is defined as 'assumptions', B is defined as theory and C as consequences or predictions.

Theory (B) consists of "a set of axioms, postulates or hypotheses that stipulate something about observable reality… " The set is either confirmable or refutable in principle by observation. A theory has a set of consequences (C) which are logically implied by theory, and a set assumptions (A) which logically implies the theory. The degree of "realism", "factual conrrectness", "empirical validity" or "truth" of any one of A,B,C is shared by the other two. Referring to 'F-Twist', he observes first, it is a contradiction to maintain all consequences (C) are valid and the theory (B) and the assumptions (A) are not valid. Second, it is absurd to maintain, in case where only some of the consequences (C) are valid, that the theory (B) and assumptions (A) are important though invalid. The part of the theory set and assumptions set corresponding to the invalid part of the consequence set should be eliminated.

4.6.2 Descriptivism

Samuelson has certain exalted view of explanation, which he considers different from the usual notion in science. He observed, "scientists never" explain" any behaviour, by theory or any other hook. Every description that is superseded by a "deeper explanation" turns out …to have been replaced by still another description …." An explanation, as used legitimately in science, is a better kind of description and not something that goes ultimately beyond description. It is this emphasis and elaboration of 'description' that earns Samuelson's approach the title of 'descriptivism'. Why only description? First, a theory is just description of observable experience, a convenient mnemonic representation of empirical reality. Second, knowledge consists essentially of observational reports. A theory expressible in observational language is superior to those which are not. Explanations are ultimate. Apriorisms must be avoided; hence theories should be expressed in observational language. All known theories in science are expressible in terms of observational statements, i.e., basic statements.

Samuelson is criticized for not practicing what he preaches methodologically. Machlup, citing Samuelson's famous work on 'factor price equalisation', accuses him for not following his own norm of deriving "operationally meaningful theorem" based on unrealistic assumptions. Another famous criticism is that of Stanley Wong's against the methodological foundations of Samulson's revealed preference theory.

4.7 THEORY – ASSUMPTIONS DEBATE IN ECONOMICS: A LONG VIEW

The spectacular advance of modern science is usually attributed very largely to the development of deductively related statements known as theories. Theories in science brought improvements in order and clarity, broadened the scope of generalizations and scored extraordinary predictive success. However, economic theory does not command as much respect as theories in physical sciences. Ever since the classical economics, there has been criticism of unrealistic assumptions of economic theory. There have been reasoned replies from time to time. Jack Melitz (1965), in one of his classic papers on the theory – assumptions controversy, provides an account of the response to the criticism since late Twentieth century. Here is a brief summary, which helps us to understand the problems of explanation that persist in economics.

During 1880-1920 advocates of economic theory adopted a *moderate and conciliatory* stance. They agreed that economic theory made false assumptions, and admitted that the value of the theory depended greatly on the degree of accord between the assumptions and the facts. Yet they insisted, first, that the assumptions did correspond broadly with events. Second sacrifice of some accuracy for simplicity was justified in view of complexity of reality. Further, they emphasized importance of combining the use of simplifying assumptions with protective measures:

1) The need to pursue more inductive studies in all areas of economics.

2) The determination of reasonable proximity between major assumptions and facts before application of economic theory.

3) The alteration of assumptions to suit the particular case involved.

During 1920s and 1930s with notable advance of purely logical branch of economics, the advocates of economic theory lost appeal. The practitioners of economic theory with the support of Robbins and the Austrians turned to a *priorism*. There was a stance that economic theory is its own reward.

In 1948-1953, as we have seen earlier, Milton Friedman tried to supply a logical foundation for the developing attitude that the realism of assumptions is not a genuine, or only a secondary concern. False assumptions in economics do not result in a handicap. Assumptions must simply work, and yield reliable results. The criticism that assumptions in economics do not correspond with facts is besides the point. In 1955 Firtz Machlup joined forces with Friedman, claiming to bring with him the support of experts in the philosophy of science and the whole *tradition* of political economy. The result is that economics continues to live with a certain methodological inadequacy. The other reasons for this disquiet in explanation in economics are summed up in the next section.

4.8 AMARTYA SEN ON HETEROGENEITY OF EXPLANATION IN ECONOMICS

By now it is clear that explanation in economics is nowhere near being in a satisfactory state. In a broad based critique of contemporary economic

methodology, Amartya Sen (1989) draws attention to the heterogeneity of the subject matter of economics. Any attempt to think of mono-method for all the diverse concerns of economics is bound to cause the kind of disquiet that we experience today. He cogently argues for heterogeneity of methodological approaches in economics.

4.8.1 Heterogeneity of Substance and Methods of Economics

According to Sen, economics as a subject is concerned with many different types of problems. The diversity of the discipline of economics should be kept in view to achieve an adequate grip on the methodological issues in the subject. The subject of economics includes three diverse, though interrelated, exercises:

1) Predicting the future and causally explaining the past events.

2) Choosing appropriate descriptions of states and events in the past and the present, and

3) Providing normative evaluations of states, institutions, and policies.

These exercises are interrelated but each requires a different methodological approach. For instance, the method of scientific explanation that insists on prediction is concerned only with the first set of exercises. The 'methodology of economics' has to admit enough diversity to be able to deal with other classes of problems as well.

4.8.2 Testing and Verification

Testing and *Verification* are important for many types of economic analyses since they are concerned with causal relationships and with making predictions. But these are not suitable for all economic theories. For instance evaluative exercises are not open to testing. Normative evaluation is a different discipline from that of making predictions on the basis of causal hypotheses. Similarly, some descriptive propositions do not have predictive content, and "testing" would be the wrong operation to seek. As far as causal theories are concerned, the need for testing them with empirical information is fairly accepted in principle by economists. Conceptual and analytical issues need to be explored very substantially to understand what types of relationships might be involved. Analyses at this stage are not meant for testing and verification. Of course, one should not end at this stage and move up to the next where testing is possible.

4.8.3 Value Judgments and Welfare Economics

Sen draws attention to the fact that following Robbins value judgments are kept out of economics. "The decision to keep economics 'value-free' would, of course, militate against the subject of welfare economics as such". Welfare economics still is accepted as important and in this domain to keep economics value-free may not be a value that would be appreciated. In evaluative exercises in welfare economics descriptive methods becomes indispensable.

4.8.4 Formalisation and Mathematics

Mathematics helps to an extent in formal reasoning and helped in systematization of many economic propositions. However, there are severe limitations of formal language of mathematics. Not all economic propositions can be reduced into mathematics. Lack of balance often has resulted in certain degree of over-

formalization of economics. While recognizing the positive contribution of mathematics in lending rigour to certain economic propositions, the negative contribution is through over concentration on mathematics to the exclusion of other disciplines. Such excesses of formalization need to be corrected.

Check Your Progress 3

1) What do you mean by the term 'logical equivalence'?

.....

.....

.....

2) Do you agree that different methodological approaches are needed in Economics?

.....

.....

.....

3) To what extent mathematics should be allowed to use in systematization of economic proposition?

.....

.....

.....

4.9 LET US SUM UP

Keeping in view that the subject matter of economics is complex and heterogeneous, any single model of explanation is bound to face serious limitations. Yet there has been an ambition in the profession from the times of classical political economy, which has turned towards a kind of arrogance that economics is a science like any physical sciences and should conform to the rule of one methodological approach, viz., logical positivism. This has been subjected to severe criticism on its limitations.

The mainstream economics is not able to come out of this obsession "It is about the one methodological rule which has dominated economics since early 1960s. The rule at issue is the methodological requirement that economic models or theories, if they are going to be given serious consideration by practicing economists, must be shown to be testable, where successful test of theory is defined as a falsification of that theory. A testable theory is a falsifiable theory".

The exploratory lesson on explanation that we had should help a student of economics to move out of this obsession. It is not that abandon testing but we must abandon the habit of considering evaluation, description, and even advocacy as not being part of pursuing knowledge in economics. Heterogeneity of economics needs diversity of approaches if the disquiet of disaffection that pervades economics is to be overcome.

4.10 KEY WORDS

Determinism : A term used to describe an argument or methodology that simply reduces casuality to a single set of factors acting more or less directly to produce outcomes.

Hypothesis : Hypothesis refers to a tentative statement that can be tested by applying the methods of particular science.

Empiricism : Empiricism is generally regarded as being at the heart of the modern scientific method that our theories should be based on our observations of the world rather than on intuition or faith; that is, empirical research and a posteriori inductive reasoning rather than purely deductive logic.

Ideology : A systematic body of concepts about human life or culture; a manner or the content of thinking characteristic of an individual or group or culture; also refers to the integrated assertion and theories that constitute a socio-political programme.

Methodological Dualism : Referring to positivist belief in separation of the knower from the known or of the subject from object.

Methodological Monism : In contrast to the compartmentalization of dualism, monism views the world as a "seamless web". In terms of Gouldner's argument, the separation between the knower and the known must be overcome.

Nomothetic : Relating to discovery of general laws or relating to the discovery of universal law.

Pseudo Science : Pseudo science refers to any body of knowledge or practice, which purports to be scientific or supported by science but which is judged to fall outside the domain of science.

4.11 SOME USEFUL BOOKS

Blaug, Mark (1980); *Methodology of Economics*, Cambridge Univeristy Press, Cambridge.

Caldwell, Bruce (1982); *Beyond Positivism: Economic Methodology in the Twentieth Centrury*, George Allen and Unwin, London.

Friedman, Milton (1953); *Essays in Positive Economics*, University of Chicago Press, Chicago.

Hausman, Daniel M. (1994); *The Philosophy of Economics: An Anthology*, Cambridge University Press, Cambridge, 2nd Edition.

Melitz, Jack (1965); "Friedman and Machlup on the Significance of Testing Economic Assumptions", Journal of Political Economy, pp.37-60.

Sen, Amartya (1989); "*Economic Methodology: Heterogeneity and Relevance*", Social Research, Vol.56, No.2 (Summer 1989), pp. 299-329.

4.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Positive, abstract and deductive briefly called hyupothetico-deductive model of explanation.

2) See Sub-section 4.2.2.

3) Methodological pluralism refers to comprehensive reconciliation of different methods and acceptance of the idea that different methods are appropriate depending upon the object in view, the stage of investigation reached and the material available.

Check Your Progress 2

1) Stress on pure theory bases on a priori truth of experience and without any need for testing, no room for ethical consideration, rationality as maximizing behaviour and as consistency of choice.

2) The insistence of testing both theory as well as assumptions.

3) See Sub-section 4.5.1

4) The proposition that realism of assumption does not matter.

Check Your Progress 3

1) Theories are merely equivalent restatements of assumptions and conclusions.

2) See Section 4.8

3) See Sub-section 4.8.4

UNIT 5 FOUNDATIONS OF QUALITATIVE RESEARCH: INTERPRETATIVISIM AND CRITICAL THEORY PARADIGM

Structure

5.0 Objectives

5.1 Introduction

5.2 Interpretive Paradigm

5.3 Critical Theory Paradigm

5.4 Applications in Research: Illustrative Cases

5.5 Let Us Sum Up

5.6 Exercises

5.7 Key Words

5.8 Some Useful Books

5.9 Answers or Hints to Check Your Progress Exercises

5.0 OBJECTIVES

This unit introduces the two widely used and well-established paradigms namely, interpretive and critical theory. Unlike post-positivism, these paradigms are the pillars of qualitative research. After going through this unit, you will be able to:

- describe the essence of interpretive and critical theory paradigm;
- discuss their theoretical framework;
- elucidate the nature of reality and methodology underlying these two paradigms; and
- state the examples and applications of interpretive and critical theory paradigm in economics.

5.1 INTRODUCTION

As has been discussed earlier in Unit 1, you have come to know that a paradigm essentially refers to a comprehensive belief system or theoretical framework/ world view that guides research and practice in a particular field. At its basic level, it has a philosophy of science that makes a number of fundamental assumptions about the nature of truth and what it means to know. Unlike positivism, that is based on empiricism and deductive reasoning for verification of truth, interpretive and critical theory paradigms are based on **subjectivism** and **substantialism** for the search of truth. The underlying ontological and epistemological assumptions of interpretitivism and critical theory paradigms have been discussed in the subsequent sections. It is important to note that these two paradigms form the backbone of philosophical science of qualitative research.

Before embarking on meaningful research using qualitative approach, it is important to understand the meaning, nature of reality, methodology and the essence of these two paradigms. Let us begin with interpretative paradigm.

5.2 INTERPRETIVE PARADIGM

Interpretive research assumes "that our knowledge of reality is gained only through social constructions such as language, consciousness, shared meanings, documents, tools, and other artifacts" (Klein & Myers, p. 69). As the term indicates, interpretive paradigm looks for understanding of a particular context. Interpretivists believe that it is important to understand the context in which research is conducted for proper interpretation of the data. The interest of interpretivists lies not in the generation of a new theory, but to judge or evaluate and refine interpretive theories. Researchers using an interpretive approach aim to uncover meaning towards a better understanding of the issues involved. The underlying ontological assumption of interpretive paradigm is subjectivism as here reality is viewed as socially constructed and interpreted. Epistemological assumption is that knowledge of reality is obtained from the accounts that social actors provide. Neuman (1997) affirms that "social reality is based on people's definition of it" (p. 69). From the previous assertions, it is apparent that interpretive researchers do not recognize the existence of an objective world. On the contrary, they see the world strongly bounded by particular time and specific context. Therefore, the epistemological question, "What is the nature of the relationship between the knower or would-be knower and what can be known" (Guba & Lincoln, 1994, p. 108) must be answered in a consistent way with the ontological view. The interpretive researcher's epistemological assumption is that "findings are literally created as the investigation proceeds" (Guba & Lincoln, p. 111). Moreover, they explicitly recognise that "understanding social reality requires understanding how practices and meanings are formed and informed by the language and tacit norms shared by humans working towards some shared goal" (Orlikowski & Baroudi, 1991, p. 14).

Interpretive research focuses on identifying, documenting, and 'knowing' – through interpretation of :

- world views,
- values,
- meanings,
- beliefs,
- thoughts and the general characteristics of life events, situations, ceremonies and specific phenomena under investigation,
- the goal being to document and interpret as fully as possible the totality of whatever is being studied in particular contexts from the people's viewpoint or frame of reference.

Interpretivists assert that all research is influenced and shaped by pre-existing theories and world-views of the researchers. The terms, procedures and data of research have meaning because a group of scholars has agreed on that meaning.

Research is thus a socially constructed activity. Three other approaches support the philosophy of interpretive paradigm. They are Verstehen, hermeneutics and phenomenology. Verstehen stresses on the understanding of the particulars of a situation, hermeneutics emphasizes on the importance of language and context in understanding and phenomenology on people's perception of the world.

Methodology

Foundationalism is an approach that asserts research can begin with **self-evident truths** which can serve as the starting point for our understanding of the world. Interpretivists are anti-foundationalists as they believe that the standards that guide research are **products of a particular group or culture**. They use a broad range of qualitative methods and gather thoughtful reflections of experienced practitioners. Interpretative approaches rely heavily on naturalistic methods (interviewing and observation) and their methods generally ensure an adequate dialogue between the researchers and the people with whom they interact to be able to construct a meaningful reality. These meanings emerge from the research process. Certain methodologies used in interpretive research are:

Naturalistic inquiry, that uses first-hand observation to understand human action and studies real-life situations as they unfold. It is non-manipulative and non-controlling, hence has lack of predetermined constraints on outcomes;

Phenomenologic methodologies, rely on descriptions of conscious experiences to develop understanding of the meaning of human action;

Constructivism, makes use of perception or self-experience in making and structuring knowledge;

Ethnographic inquiry, rely on self-experiencing the culture of participants in the field;

Symbolic interactionism, understands human action as per the meanings derived by the human beings for particular objects and people.

Some other methods for data-collection are surveys, interviews, field observation, witness accounts, focus group discussion. Interpretive research adds to the understanding of different contexts and situations. Interpretivists argue that research results should be applied to higher, conceptual level. From the data, a researcher tries to understand multiple perspectives on the same topic. For analyzing the meaning of data, interpretive researchers may use and conduct research by methods from another paradigm (eg. Post-positivism) also and doing so would lead to reinterpreting the meaning of results by another perspective.

With a philosophical alignment with interpretive naturalistic orientations, interpretive description acknowledges the constructed and contextual nature of human experience that at the same time allows for shared realities (Thorne, Reimer Kirkham, & MacDonald-Emes, 1997). Key axioms of naturalistic inquiry, such as those delineated by Lincoln and Guba (1985), provide philosophical underpinnings for research design, including:

1) There are multiple constructed realities that can be studied only holistically. Thus, reality is complex, contextual, constructed, and ultimately subjective.

2) The inquirer and the "object" of inquiry interact to influence one another; indeed, the knower and known are inseparable.

3) No *a priori* theory could possibly encompass the multiple realities that are likely to be encountered; rather, theory must emerge or be grounded in the data.

Research Strategy for Interpretivism

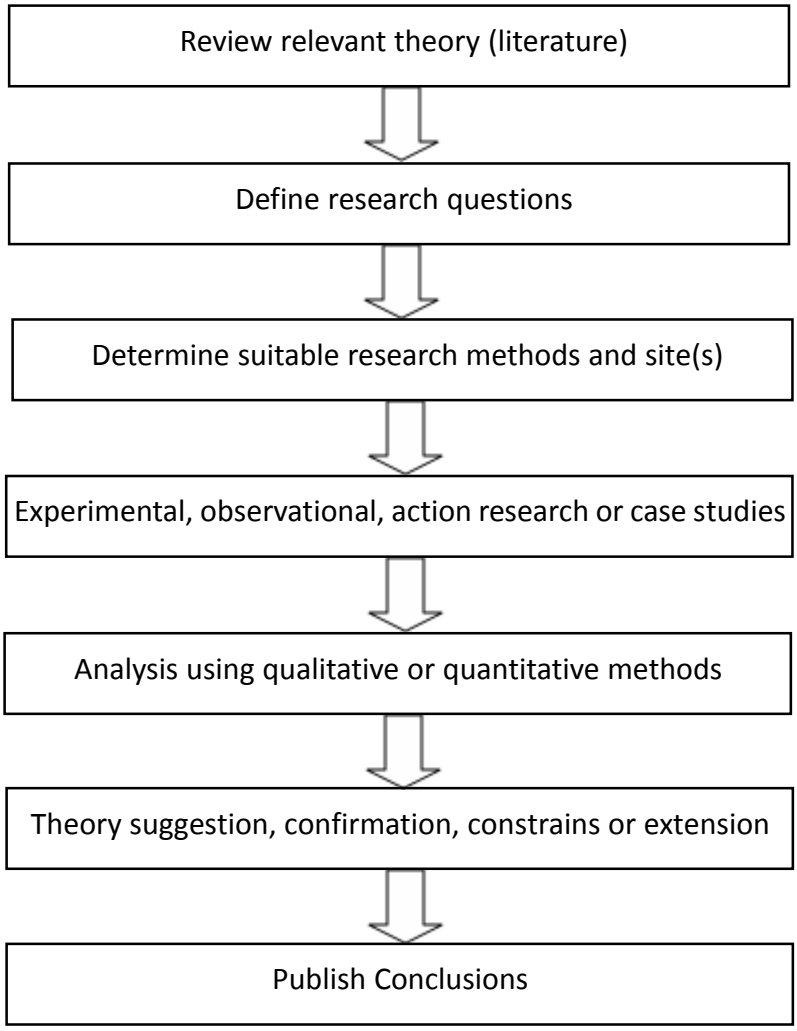
Broadly there are four distinct research strategies that work with different ontological assumptions. Abduction strategy best fulfills the needs of interpretive researcher as it starts with laying the concepts and meanings that are contained in social quarter's accounts of activities related to a research problem.

As interpretivists construct reality on the accounts of social actors present in a specific situation, it is important to take into account the different perspectives of each participant. One can understand the complexity of the issue through the following example, excerpts taken from Andrade, 2009.

Let's imagine a scenario at the beach in which a huge wave is approaching the shore. There is an excited surfer on top of the big wave and two scared children in a small inflatable boat right below the colossal wave. On the shore, a girl is admiring her boyfriend's dexterity and the petrified children's mother is watching the looming mass of water approaching the boat. On the adjacent cliff there is a relaxed monk meditating on the infiniteness of the universe, while enjoying the sea breeze and the sound of the sea. If we want to conduct research on what that wave means for beach-goers, our results will depend on who the respondent is. Interviewing one of the participants would give insights from that participant's perspective only, which may be insufficient, or even misleading, because their personal and intimate experiences with the wave are quite different from that of the others. If the interpretive researcher wants to create an integral and persuasive piece of research around this phenomenon, each participant's different perspectives should be included.

To conduct interpretive research on a certain setting, intense and long-term participant observation is required, followed by deliberate and long-term reflection on what was observed. Interpretive researchers start out with the assumption that access to reality (given or socially constructed) is only through social constructions such as language, consciousness and shared meanings. Interpretive studies generally attempt to understand phenomena through the meanings that people assign to them and interpretive methods of research in IS are "aimed at producing an understanding of the context of the information system, and the process whereby the information system influences and is influenced by the context. Interpretive research does not predefine dependent and independent variables, but focuses on the full complexity of human sense making as the situation emerges.

The life-cycle of theory-generation under interpretive paradigm has been depicted in the following flow-chart:



Criticism

Researchers who wish to use interpretive framework for their case-studies are cautioned not to lose theoretical sensitivity. Strauss and Corbin (1990, p. 41) describe theoretical sensitivity as the "awareness of the subtleties of meaning of data" and elaborate that "one can come to a research situation with varying degrees of sensitivity depending upon previous reading and experience with or relevant to that area." Ultimately, the researcher has to evaluate the relevance of their preliminary theoretical framework vis-à-vis the actual findings (Urquhart, 2001, 2007). The criteria to be kept in mind for using interpretive case studies aiming for theory-building can be seen in the appendix to the unit.

Differences of Interpretive paradigm from post-positivism can be broadly summarized in the table below:

Differences between Post-positivism and Interpretivism on Five major Issues

	Post-positivism	**Interpretivism**
Nature of reality	External to human mind	Socially constructed
Purpose of research	Find Universals	Reflect understanding
Acceptable methods and data	Scientific method	Subjective and Objective research methods are acceptable

	Post-positivism	Interpretivism
Meaning of data	Falsification Use to test theory	Understanding is contextual. Universals are deemphasized
Relationship of research to practice	Separate activities. Research guides practice	Integrated activities. Both guide and become the other.

(**Source:** Foundations of qualitative research by Jerry W. Willis (2007) p.95)

Check Your Progress 1

1) Explain how interpretive framework is a departure from post-positivism?

.....

.....

.....

2) Outline the methodology of interpretive research using a suitable example from your locality.

.....

.....

.....

3) How do the ontological and epistemological assumptions underlying interpretive paradigm guide the researcher to the methodology of research, data-collection and analysis?

.....

.....

.....

5.3 CRITICAL THEORY PARADIGM

Critical theories share some ideas of the interpretative paradigm, but what makes it different is that critical paradigm focuses on oppression. Critical social scientists believe it necessary to understand the lived experience of real people in context. Persons can perceive reality outside them and represent that reality with language. Also, reality is defined by the interaction between the knower and the known. Critical approaches examine social conditions and uncover oppressive power arrangements. 'Critical theory' is a term that can apply to a number of movements in the social sciences. As it is based on ideology of class conflict, the critical theorists whose objective is to find out power-relationships among the different sections of the society (between the doers and the oppressed), it is sometimes referred as 'ideologically oriented inquiry, neo-Marxism, materialism, the Frankfurt School and freireism. Critical theory was built on the foundation of Marxism which perceived conflicts between classes in society and believed that it could be reformed only via radical transformation of the society. The roots of the critical theory can be traced to the Frankfurt school or the Institute for Social

Research. It was established as a school of thought primarily by five Frankfurt school theoreticians; Herbert Marcuse, Theodor Adorno, Max Horkheimer, Walter Benjamin and Erich Fromm. The underlying ontological assumption of critical theory paradigm is substantialism wherein matter constitutes the reality. However, epistemologically people interpret reality differently in different times and places. Critical Theory is a theoretical tradition developed most notably by Horkeimer, Adorno, Marcuse at the Frankfort School. Their work is a critical response to the works of Marx, Kant, Hegel and Weber.

- **Historical ontology** – assumes that there is a 'reality' that is apprehendable. This is a reality created and shaped by social, political, cultural, economic, ethnic and gender-based forces that have been reified or crystallized over time into social structures that are taken to be natural or real. People, including researchers, function under the assumption that for all practical purposes these structures are real. Critical theorist believe this assumption is inappropriate.
- **Modified transactional or subjectivist epistemology** – we cannot separate ourselves from what we know and this inevitably influences inquiry. What can be known is inextricably tied to the interaction between a particular investigator and a particular object or group.

Purpose of Research

Research that aspires to be critical seeks, as its **purpose of inquiry**, to confront injustices in society. Following a tradition associated with Antonio Gramsci, critical researchers aim to understand the relationship between societal structures (especially those economic and political) and ideological patterns of thought that constrain the human imagination and thus limit opportunities for confronting and changing unjust social systems. Critical theorists are committed to understanding the complexity of such relations, however, and thus distance themselves from what they see as reductionist Marxist approaches. Critical theorists hold that these earlier approaches offered no ability to explain social change. Thus, in contrast to what they believe was an overemphasis on the determinative nature of economic and political structures, critical theorists are interested in social change as it occurs in relation to social struggle. Critical researchers assume that the knowledge developed in their research may serve as a first step toward addressing such injustices. As an approach with a definite normative dimension, the research aims for a transformative outcome, and thus is not interested in "knowledge for knowledge's sake."

Critical Theory is multi-disciplinary. It finds its applications in anthropology, economics, art criticism, education, history, psychology, political science, sociology and theology. An interesting question is why and where to apply critical theory paradigm in research. To begin with, research and practice are integrated activities in the critical paradigm. It not only studies power relationships which are critical factors in the society, the injustices and the inequities, the contradictions and the incoherencies, but also aims at helping and empowering those who are oppressed to free themselves from their oppression.

Critical theory paradigm should be applied where the objective of the researcher is to analyse the power structures in a set-up or social problems arising of such structures. For example, it can be used to understand wage inequities between male, female and child workers in the society.

Nature of reality: Critical theorists like post-positivists believe that reality is material and external to human mind but they interpret it differently. They question/critique the existing reality and analyze why such a reality exists and whether it contributes positively/negatively to human mind. Further, they also advocate evolving ways and means to reform the existing system if it is bifurcating the society into the have's and the have not's.

Methodology: Critical theorists suggest two kinds of research methodologies, namely ideology critique and action research, for undertaking research work. They rely on methods combining observation and interviewing with approaches that foster conversation and reflection. They try to challenge guiding assumptions and they begin this by asking people to reflect and question their current experience with regard to values identified. In doing research, they not only try to define a situation but change the situation.

As a research methodology, critical theory adopts an overtly critical approach to inquiry. It precedes with an attitude of suspicion, calling into question not only the data itself, but also the researcher, the research design and interpretation of findings. From a critical theory viewpoint, the task of the social scientist has three dimensions:

1) To understand the ideologically distorted subjective situation of some individual or group;
2) To explore the forces that has caused that situation;
3) To show that these forces can be overcome through awareness of them on the part of the oppressed individual or group in question.

The appropriate research strategy for critical research is **Retroduction** as it begins with a hypothetical mode of a mechanism that could explain the occurrence of a phenomenon under investigation.

- Critical theoretical approaches tend to rely on dialogic methods; methods combining observation and interviewing with approaches that foster conversation and reflection. This reflective dialogic allows the researcher and the participants to question the 'natural' state and challenge the mechanisms for order maintenance. This is a way to reclaim conflict and tension.

- Rather than naming and describing, the critical theorists try to challenge guiding assumptions.

- Critical theorists usually do this by beginning with an assumption about what is good (e.g. autonomy, democracy) and asking people in a social group, culture or organization to reflect on and question their current experience with regard to the values identified (e.g. To what extent are they an autonomous worker?)

- Critical theorists not only just try to describe a situation from a particular vantage point or set of values (e.g. the need for greater autonomy or democracy in a particular setting), but also try to change the situation.

"Objective" analysis

In their embrace of a normative perspective, Critical theorists make no claims that their analyses are "objective" in the sense usually meant by logical positivists. In fact, critical theorists argue that the subjective/objective dualism masks the ways in which both positions are limited by the social forces that inform all human action and analysis. Critical qualitative research acknowledges subjectivism in the sense that learnings and interpretations cannot be based on logic and scientific analysis only. While it affirms that knowledge can never be separated completely from the researcher's own experience, it rejects the notion that all analyses are relative. It asserts that rational analysis is fundamental to human emancipation, and hence embraces what Morrow (1994) calls critical realism

Data analysis and verification

Critical researchers assume that their task is to expose the hidden assumptions that guide both research respondent statements and often, initial analyses of data. Researchers therefore bring a level of scrutiny to their task that includes rooting out the meanings of what is left unsaid as well as that which is stated. The research is verified as other members of the research community offer corroboration that has come from their own research experiences.

Sample representativeness, typicality, and generalizability

In a response similar to that of constructivism, critical researchers employing qualitative research would note that we are not seeking to explain the "typical" person, but to analyze that person's possibilities and limits within a culture. In this approach, individuals are not seen as "types" or members of aggregate groups (although they may be both of these). Individuals instead are approached as beings that inhabit subject positions that are possible within a culture. Because individuals are members of society and must act within the society, they share certain understandings and meanings; if they did not, they could be considered insane, which in societal terms is the designation given to persons whose social realities have no seeming connection to those around them.

Validity

The test of validity in the case of critical constructivist research is directly related to its stated purpose of inquiry. The research is valid to the extent that the analysis provides insight into the systems of oppression and domination that limit human freedoms, and on a secondary level, in its usefulness in countering such systems.

Criticism

One of the charges against critical theory is its tendency toward elitism. With its proponents' commitment to the idea that research can bring about a better and more equitable world, critics charge that critical theorists tend to assume that they are not only more capable of analyzing a situation than most; they are better equipped to offer a prescriptive plan of action. Critics charge that this often brings theorists outside of their realms of expertise so that the insights they offer are naive and unworkable in the contemporary setting.

Further, critics charge that critical theorists can be unwilling to listen to the experiences of those most adversely affected by current policies and the status

quo, as they tend to focus their analyses on persons and institutions in positions of power and authority. This, critics note, causes critical theorists to be out of touch with the very persons they purport to be most interested in helping.

Differences between critical theory framework and post-positivism on five major issues are as given below:

	Post-positivism	**Critical Theory**
Nature of reality	External to human mind	Material and external to the human mind.
Purpose of research	Find Universals	Uncover local instances of universal power relationships and empower the oppressed.
Acceptable methods and data	Scientific method	Subjective inquiry based on ideology and values; both quantitative and qualitative methods are acceptable.
Meaning of data	Falsification Use to test theory	Interpreted through ideology; used to enlighten and emancipate.
Relationship of research to practice	Separate activities. Research guides practice	Integrated activities. Research guides practice.

(**Source:** Foundations of qualitative research by Jerry W. Willis (2007))

Check Your Progress 2

1) Is critical framework a departure from post-positivism and interpretivism? Give reasons in support of your answer?

.....

.....

.....

2) What points should be kept in mind while defining the methodology of critical research?

.....

.....

.....

3) Critical framework is sometimes known as the Frankfurt school. Why? Elaborate on the foundations of critical theory.

.....

.....

.....

5.4 APPLICATIONS IN RESEARCH: ILLUSTRATIVE CASES

Case 1: 'A study on 'invisible' labour-force in India

Invisible workforce comprises of care-workers who work daily for low/negligible wages however their work is critical to the success of the household/enterprise. Primarily constituting of housewives, nannies, cleaners, etc., these workers are engaged in 'care' work and their work is generally under-valued or over-looked. A news daily recently reported that as per Census 2011 figures nearly 160 million women in India aged between 15-59 years reported themselves as not working but were primarily involved in domestic work, care work and rearing families. In this case, the social actors are the housewives, nannies, cleaners, etc.

An interpretive researcher by using qualitative methods like phenomenology and symbolic interactionism would try to understand the root causes of such women not involved in economic work. Research could begin with penning down research questions like 'why the women have not taken up work outside home' or 'Are their social pressures for the women sticking to domestic work' or even 'Is the decision to work at home independent or curbed'. Next, the researcher can shortlist a region predominantly populated with invisible workers and use techniques of PRA/RRA to gather the requisite data. The perspectives on invisible work may differ among the social actors in this case too. For example if the researcher is probing the causes of invisible work he may observe that causes may be social, cultural or at times political too. It would be important to gather perspectives of not only unpaid care-workers but also their family members as to why they did not try to change their current situation, village head-men to understand whether it is due to lack of economic opportunities in the area and also to know whether illiteracy is the cause of invisibility of women workers.

Case 2: Absenteeism among students of primary school in rural areas

Universal primary education being one of the goals of Millennium Development Goals of UNDP, all developing countries are in the race of providing elementary and primary education to children in rural and urban areas. While enrolment in elementary and primary education has definitely increased in India in the last decade, but equally troublesome statistics have emerged quoting high incidence of absenteeism among students of primary school in certain rural areas. An interpretive researcher using qualitative methods and tools can take different perspectives of the social actors (read primary school children, teachers and parents) for absenteeism in primary schools.

Case 3: A study of employer's attitude towards distance/online-educated job applicants.

Distance educated youth are perceived to be less efficient and trained as compared to classroom educated youth. Instances of preferential bias have been observed in this respect often in the job market. An in-depth study on the employer's mindset can be undertaken by an interpretive researcher to explore the causes of the same. Extensive literature review is a pre-cursor to such research as that will underpin the historical causes of preferential attitude and negative perception of the employers. An interpretive researcher can employ actual observational techniques or use witness accounts of the interviews/interactions of the employer

with distance/online educated job applicants. Research questions may range from finding variability in the questions posed to such candidates from the ones who were educated in the traditional mode to quizzing grounds for negative perception of employers for the same. For example, a study quoted the following reasons for dislike of distance educated job applicants by an employer: lack of rigor, lack of face-to-face interactions, increased potential for academic dishonesty, association with diploma mills, concerns about online students' true commitment evident from regularly venturing to a college or university physical location, considered by some to be an important part of the educational experience.

Interestingly all of the above cases can also be studied in a critical framework as only the questions posed would change and power-relationships would be judged. Critical theory is also used subsequently after interpretive research as it is easier to critically analyze a problem after undertaking a deeper understanding of that problem. For example, after understanding the reasons behind the existence of invisible workers, a critical researcher may critically analyze the role of power relationships between the care workers and their families, social circle and the environmental factors leading to it. Awareness campaigns and policies may be designed to further empower them to be able to participate in the economic workforce and to allow labour statisticians to recognize their care work within the ambit of 'work'.

Check Your Progress 3

1) Using a suitable example from economics, select a theme and write a proposal in about 500 words using interpretative research methodology.

.....

.....

.....

.....

2) Explain the scenario when both interpretative and critical theory framework can be applied in the same research? Do they supplement each other? Why or why not?

.....

.....

.....

.....

5.5 LET US SUM UP

The three paradigms – postpositivism, interpretivism and critical theory – are the dominant guiding frameworks in the research literature in the social sciences. Interpretivism and critical theory are two of the contemporary frameworks of qualitative research besides positivism and post-positivism paradigms. Based upon different ontological and epistemological positions, these paradigms analyze nature of reality differently. Interpretivism is based on socially constructed reality

whereas critical research is based on the critique of social reality. These paradigms have emerged as a response to a perceived problem in society. Critical theory was a response to the complexities of modern nation states that often lead to domination and exploitation of one group by the other. Interpretivism proposed that we abandon the search for generalizable truths and laws about human behaviour and concentrate instead on local understanding.

5.6 EXERCISES

1) Consider the interpretivist philosophy of science. What to you see as the most significant departure from post-positivism and why?

2) Consider a topic of research in an area of interest to you. What would be the purpose of research on the topic if it were conducted from a post-positivist perspective? Critical or Interpretive?

3) What does interpretivism have in common with the critical paradigm? In your field which are more important: the commonalities or the differences?

5.7 KEY WORDS

Paradigm : A comprehensive belief system, world view or framework that guides research and practice.

Ontology : It is concerned with the nature of reality and inquires the characteristics of things that exist, is a major aspect of metaphysics- a branch of philosophy.

Epistemology : It is concerned how can we know? It is also a major aspect of metaphysics.

5.8 SOME USEFUL BOOKS/WEBLINKS

Willis, Jerry, W (2007), *Foundations of Qualitative Research: Interpretive and Critical Approaches*, Sage Publications, USA.

Clark, Lynn Schofield, *'Critical Theory and Constructivism: Theory and Methods for the Teens and the* New Media @ Home Project' retrieved online from http://www.ihrcs.ch/?p=92

'Critical theory paradigms' retrieved online from http://www.qualres.org/HomeCrit-3518.html

5.9 ANSWER OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See criticism, Section 5.2

2) See Methodology, Section 5.2

3) See Section 5.2

Check Your Progress 2

1) See Section 5.3
2) See Section 5.3
3) See Section 5.3

Check Your Progress 3

1) See Section 5.4
2) See Section 5.4, yes they supplement each other.

APPENDIX I

Criteria for Interpretive Case study aiming at Theory Building

Criterion	Definition	Specific case study tactic	Grounded theory[1] principles
Construct validity	Establishing correct operational measures for the concepts being studied	• Use multiple sources for evidence • Establish chain of evidence • Have key informants review draft case study report	• Corroboration • Theoretical sufficiency
Internal validity	Establishing causal relationship as distinguished from spurious relationships	• Do pattern matching • Do explanation-building • Address rival explanations • Use logic models	• Theoretical coding
External validity	Establishing the domain to which a study's findings can be generalized	• Use theory in single case studies • Use replication logic in multiple case studies	• Theoretical generalisation
Reliability	Demonstrating that a study can be repeated with the same results	• Use case study protocol • Develop case study database	• Chain of evidence as afforded by grounded theory method

(Case study methodology criteria, Yin, 2003, p.24)

[1] The discovery of theory from data is known as grounded theory and it provides the opportunity for the researcher to theorise from evidence existing in the data.

Indira Gandhi
National Open University
School of Social Sciences

MEC-109
Research Methods in Economics

Block

2

RESEARCH DESIGN AND MEASUREMENT

UNIT 6

Research Design and Mixed Methods Research **5**

UNIT 7

Data Collection and Sampling Design **25**

UNIT 8

Measurement and Scaling Techniques **53**

Expert Committee

Prof. D.N. Reddy
Rtd. Professor of Economics
University of Hyderabad, Hyderabad

Prof. Harishankar Asthana
Professor of Psychology
Banaras Hindu University
Varanasi

Prof. Chandan Mukherjee
Professor and Dean
School of Development Studies
Ambedkar University, New Delhi

Prof. V.R. Panchmukhi
Rtd. Professor of Economics
Bombay University and Former
Chairman ICSSR, New Delhi

Prof. Achal Kumar Gaur
Professor of Economics
Faculty of Social Sciences
Banaras Hindu University, Varanasi

Prof. P.K. Chaubey
Professor, Indian Institute of
Public Administration, New Delhi

Shri S.S. Suryanarayana
Rtd. Joint Advisor
Planning Commission, New Delhi

Prof. Romar Korea
Professor of Economics
University of Mumbai, Mumbai

Dr. Manish Gupta
Sr. Economist
National Institute of Public Finance and Policy
New Delhi

Prof. Anjila Gupta
Professor of Economics
IGNOU, New Delhi

Prof. Narayan Prasad (**Convenor**)
Professor of Economics
IGNOU, New Delhi

Prof. K. Barik
Professor of Economics
IGNOU, New Delhi

Dr. B.S. Prakash
Associate Professor in Economics
IGNOU, New Delhi

Shri Saugato Sen
Associate Professor in Economics
IGNOU, New Delhi

Course Coordinator and Editor: Prof. Narayan Prasad

Block Preparation Team

Unit	Resource Person	Editor (Format, Language and Content)
6	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi
7	Sai S.S. Suryanarayan Rtd. Joint Advisor Planning Commission, New Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi
8	Prof. A.K. Gaur Professor of Economics BHU, Varanasi & Prof. Narayan Prasad Professor of Economics, IGNOU, New Delhi	Prof. H.S. Asthana Professor of Psychology BHU, Varanasi

Print Production

Mr. Manjit Singh
Section Officer (Pub.)
SOSS, IGNOU, New Delhi

October, 2015

© Indira Gandhi National Open University, 2015

ISBN-978-81-266-

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi
Printed at :

BLOCK 2 RESEARCH DESIGN AND MEASUREMENT

Research design is a logical structure of an enquiry and its formulation is guided and determined by the research questions raised in the problem under investigation. Apart from specifying the logical structure of data, research design also test and eliminate alternative explanation. Broadly, the observational design sampling design and statistical design are covered in Research Design. The various attributes of people, objects or concepts are being increasingly included in explanation of human behaviour in Economics. Hence, these individual traits, attitudes need to measure for deeper analysis. Research Design and measurement issues, therefore constitute the theme of this block. The block comprises of 3 Units.

Throwing light on the concept of research design and various types of research, **Unit 6** covers the rationale and forms of mixed methods research with illustration of three case studies.

Unit 7 provides comprehensive knowledge of all the elements for collecting the quantitative data for research study. It covers the methods and tools of data collection and the various issues related to sampling design.

Unit 8 deals with the various measurement scales, scaling techniques, and criteria for good measurement and sources of errors in measurement.

UNIT 6 RESEARCH DESIGN AND MIXED METHODS RESEARCH

Structure

6.0 Objectives

6.1 Introduction

6.2 Types of Research

- 6.2.1 Theoretical and Applied Research
- 6.2.2 Descriptive and Explanatory Research
- 6.2.3 Quantitative and Qualitative Research
- 6.2.4 Conceptual and Empirical Research
- 6.2.5 Other Types of Research

6.3 Research Design

6.4 Research Design vs. Research Methods

6.5 Research Methods

- 6.5.1 Quantitative Methods
- 6.5.2 Qualitative Methods
- 6.5.3 Mixed Methods

6.6 The Rationale for Mixed Methods Research

6.7 Forms of Mixed Methods Research Designs

6.8 Case Studies of Mixed Methods Research Design

6.9 Let Us Sum Up

6.10 Some Useful Books

6.11 Answers or Hints to Check Your Progress Exercises

6.0 OBJECTIVES

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

- state the various types of research;
- explain the concept and meaning of research design;
- make a distinction between research design and research methods;
- delineate the characteristics of quantitative research and qualitative research;
- discuss the purposes of mixed methods research; and
- appreciate the application of mixed methods research in conducting research studies in social sciences in general and economics in particular.

6.1 INTRODUCTION

Due to confusion or lack of clarity between research design and research methods, we often use these two terms interchangeably. As a result, the research design is evaluated on the basis of strengths and weakness of the method rather its ability to draw relatively unambiguous conclusions. Hence

efforts have been made here to throw light on the distinction between these two terms. Traditionally two approaches – quantitative research (QN) associated with post positivist paradigm and qualitative research (QL) associated with interpretative paradigm have been followed in social sciences. However, in recent years increasing reliance on mixed methods research by combining quantitative and qualitative methods in a single research project has been observed. Hence in this unit, the concept of mixed method research, its types and application have been covered. Application of mixed method research in a subject like economics which is essentially quantitative in nature have been illustrated by case studies. Before taking up the issue of research design and research method, it is desirable to have an idea about various types of research. Hence let us begin to introduce the various types of research.

6.2 TYPES OF RESEARCH

There are different 'types of research' depending on their nature and field of specialization. While a classification based on a dichotomous distinction (like, theoretical/applied, descriptive/analytical, conceptual/empirical, etc.) is possible, it is necessary to recognize that there may be overlapping features rendering such a classification less perfect to that extent. It is, therefore, useful to first of all take a look at the underlying basic features so as to be able to identify the different types of research within their respective connotations and usage.

A variable which can be measured can take values or scores either in quantitative or qualitative terms. For instance, yield of an agricultural output, height/weight of individuals, etc. can be measured and expressed in quantitative terms. On the other hand, characteristics like one's feelings or opinion (e.g., good/bad, male/female, agreeing/disagreeing, yes/no, etc.) are attributes on which a choice can be expressed (and a score assigned) depending on the possible alternatives or choices. Thus, variables like the performance of an artist, respondent's gender, etc. on which responses can only be categorized or grouped are **'qualitative variables'**. From this angle, variables on which data can be collected and expressed in quantified terms may be called as '**quantitative variables**'. Even variables which are expressed in qualitative terms can be counted whereupon their total numbers become quantified expressions. A popular example of developmental economics which ranks different countries based on their status attained in a number of areas/variables(some of which like the political freedom/choice enjoyed by the country, though expressed in some quantitative measure, is qualitative in its nature) is the human development index. In actual research (particularly research in the field of social sciences), it is important to note that variables of both the nature normally co-exist. In view of this, the analysis pertaining to each type of variable/data needs different methods and techniques.

The type of research would also vary depending on the objectives of the study. Research design varies with the type of research one likes to pursue. With this background, we can now proceed to know the distinctions between different types of research.

6.2.1 Theoretical and Applied Research

Research can either be theoretical or applied. The former i.e., theoretical research can also be considered as 'fundamental research' as its outcome serves

as a foundation for all subsequent development in the field. Fundamental (or basic) research mainly concerns with formulation of theory with knowledge perceived as an end in itself. It, thus, aims at obtaining knowledge of the logical processes involved in a phenomenon. It pertains to the quest for knowledge about a phenomenon without concern for its practical use. Such a research may either verify the old theory or establish a new one. For example, fundamental research in economics may consist of research to develop and improve economic theories or evolve quantitative techniques to measure parameters such as multiplier effect, elasticity of demand and supply, etc. Fundamental research is essentially positive in nature.

Applied research, on the other hand, aims at finding a solution for a problem facing society or industry. It is, thus, applied to practical situations or contexts. While pure research discovers principles and laws, applied research discovers ways of applying them to solve specific problems. It is useful to test the theories developed empirically and can as a result also contribute to improving the tools and techniques of measurement. Basic research, therefore, can be treated as building blocks for applied research. Illustrations of applied research in economics can be measurement of poverty, employment, rural development, agriculture, environment, etc.

6.2.2 Descriptive and Explanatory Research

Descriptive research describes a situation, events or social systems. It aims to describe the state of affairs as it exists. Surveys and fact-finding enquiries of different kinds are part of descriptive research. Survey methods of all kinds including comparative and correlational methods are used in descriptive research studies. A survey of socio-economic conditions of rural/urban labour is an area of descriptive research. In descriptive research studies, the researchers have no control over variable. They can report only what has happened or is happening. Descriptive research deals with questions like ‘how does X vary with Y? or ‘how does malnutrition vary with age and sex?, etc.

Explanatory research, aims at establishing the cause and effect relationship. The researcher uses the facts or information already available to analyse and make a critical evaluation of the data/information. An example of explanatory research is: ‘whether increase in agricultural productivity is explained by improved rural roads?’

6.2.3 Quantitative and Qualitative Research

Quantitative Research is the systematic and scientific investigation of quantitative properties and phenomena and their relationships. The objective of quantitative research is to develop mathematical models, and to test hypotheses. Based on ontological and epistomological premises, quantitative research is characterized by the following attributes:

- A belief in a single reality.
- The pursuit of identifying universal laws.
- Separation of knower (researcher) from known (researched).
- The possibility and necessacity of value free research.
- Pursuit of universal laws (findings) beyond the limit of research/social context.

- The tendency to work with large, representative sample.
- An emphasis on deductive research via falsifiable hypothesis.
- Formal hypothesis testing.

In contrast, **Qualitative Research** captures a set of purposes associated with meaning and interpretation. Stress is laid on the socially constructed nature of reality, the intimate relationship between the research and researched and situational constraints that shape the enquiry. The key attributes associated with qualitative research are:

- A belief in a constructed reality, multiple realities, co-existence realities.
- An interdependence between the knower and the known.
- Impossibility to separate the researcher from the research subject.
- Heavy role of context in research process.
- The impossibility to generalize research findings beyond the limits of the immediate context.
- Non-separation of cause and effect.
- The explicit focus on inductive, exploratory research approaches.
- The tendency to work with small and purposely chosen sampling.
- Analyses holistic system.

6.2.4 Conceptual and Empirical Research

Conceptual research is related to abstract ideas or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret the existing theories.

Empirical research relies on experience or observation. It is data based. It is subject to verification by observation and experiment. This type of research is particularly useful when validation or verification of an aspect is required.

6.2.5 Other Types of Research

Research may be exploratory or formalized. **Exploratory** research aims at developing the hypothesis rather than testing a pre-conceived hypothetical contention or notion. **Formalised** research studies deal with a definitive structure within which specific hypotheses are tested. **Historical research** utilizes existing documents to study events of the past. Research can also be experimental or evaluative. **Experimental research** aims at identifying the causal factors by means of experiments. In **evaluative research**, the cost effectiveness of a programme is examined. Such research addresses the question of the efficiency of a programme and are useful in taking policy decisions on issues like whether the programme is effective and/or needs modification or re-orientation? Whether it should be continued?

Action research is another type of research. It deals with real world problems aimed at finding out practical solutions or answers to them. It gathers feedback which is then used to generate ideas for improvement. You will find details on Action Research in Unit 20.

Check Your Progress 1

1) Distinguish between quantitative variable and qualitative variable.

……..

……..

……..

2) State the main objectives of applied research?

……..

……..

……..

3) In what sense quantitative research is different from qualitative research?

……..

……..

……..

4) What is the distinction between explanatory and descriptive research?

……..

……..

……..

6.3 RESEARCH DESIGN

Research design is a logical structure of an enquiry. Given the research question or theory, what type of evidence is needed to answer the question (or to test the theory) in a convincing way – constitutes the essence of the research design. Let us use an analogy to understand the term 'research design'. While constructing a building, the first decision to be arrived at is: whether we need a high rise office building, a factory, a school or a residential apartment etc.? Until this is decided, we cannot sketch a plan and order material or setting critical dates for completion of the project dates. Similarly, a social researcher needs to be clear about the research questions and then the research design will flow from the research questions. The function of a research design is to ensure that the evidence obtained enables us to answer the initial research questions as unambiguously as possible. Obtaining relevant evidence entails specifying the type of evidence we need to answer the research question, to test a theory, to evaluate a programme or to accurately describe some phenomenon. The issues of sampling, method of data collection (e.g. questionnaire, observation, document analysis), design of questionnaire etc. are all subsidiary to what constitute the evidences that need to collect to answer the research questions.

Thus the research design 'deals with a logical problem and not a logistical problem. Apart from specifying the logical structure of the data, it also aims to test and eliminate alternative explanation of results.

The research design comprises of the following components:

i) the sampling design; (the type of sampling method i.e. random or non random sampling);

ii) the statistical design (i.e. the sample size and the method to draw the sampling to be adopted);

iii) the observational design (i.e. the instrument of collection of data);

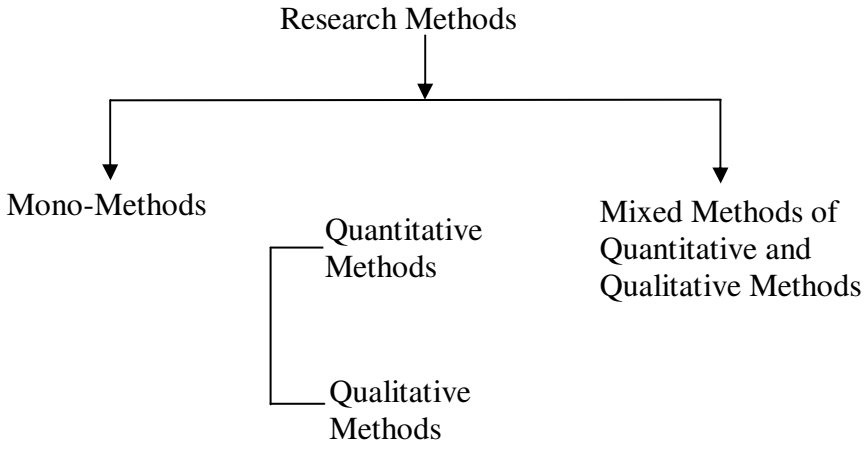
iv) the operational design i.e. the specific details by which the procedure in (i), (ii) & (iii) above are to be carried out.

6.4 RESEARCH DESIGN VS. RESEARCH METHODS

Research design is different from the research methods. The research methods are made of data collection and techniques of data analysis whereas the research design is a logical structure of an enquiry. There is nothing intransic about any research design that requires a particular method of data collection. How the data are collected is irrelevant to the logic of the research design. Confusing research designs with research methods leads to poor evaluation of designs. For example equating cross sectional designs with questionnaires or case studies with participant observation result in evaluation of research designs against the strengths and weakness of the method rather than their ability to draw relatively unambiguous conclusion or to select between rival plausible hypothesis.

6.5 RESEARCH METHODS

Research methods as explained in Unit 1 refers to tools and techniques of data collection and data analysis. Broadly Research Methods are put under two categories: (i) Mono Methods, (ii) Mixed Methods. Within mono methods, there are two approaches (i) quantitative (QN) methods (ii) qualitative (QL) methods. Since quantitative methods are associated with positivist and post-positivist paradigm, quantitative method is also termed as (QN) method research. Similarly by virtue of the fact that QL methods are associated with interpretativist paradigm and critical theory paradigm, research conducted by the qualitative method is also termed as qualitative methods research.



6.5.1 Quantitative Methods

The researcher using this approach tests theory through empirical observations and inclines to establish cause and effect relationship. Well predesigned structured questionnaire and structured interviews are resorted as tool to collect the data. The variables that can be measured are emphasized. The descriptive, analytical and inferential statistical techniques are used to analyse the data. Experimental research design is used to make the logical structure of the research study. Generalization of laws and their replication is emphasized.

6.5.2 Qualitative Methods

Under qualitative approach, a researcher starts with observation as a basis for generating theory and concentrates on meaning of observations. He studies events as they occur in naturalistic settings. He allows interview topics to emerge during conversation and listens to others' interpretation and perspectives.

Under qualitative method, open-end questionnaire, key informants, group discussion and unstructured interviews, documents, interview transcripts etc. are used as tools to collect data. Content analysis, Case study, Action research participatory method, Cluster analysis, Factor analysis, Correspondence analysis, Context analysis, are used in analyzing qualitative data. You will find details of these techniques in Unit 13, 14, 15, 16 and 17 of Block 4 and 18, 19 and 20 of Block 5 of this course.

6.5.3 Mixed Methods

The combination of at least one qualitative and at least one quantitative component related to measurement scale, tools of data collection or data analysis technique in a single research study/project is known as mixed methods research. For example it can be in the following forms:

- Employment of more than one measurement procedure in different kind of situations such as different ways of measuring levels of job satisfaction.
- Employment of two or more methods of data collection. For example we may employ observation, interview and questionnaire to collect the data about decent work.
- Employment of two or more data analysis techniques for example, content analysis and factor analysis.

Mixed methods research has gained a tremendous popularity in social, behavioural and related sciences in recent years. The rationale for mixed method design research is to take the best of qualitative (QL) and quantitative methods (QN) and combine them. However, many debates on mixed method research design are based on methodological arguments due to the ways QN method is linked to positivist paradigm and QL method to Interpretative and critical theory paradigm.

As discussed above, keeping in view the diametrical oppositional attributes of QL and QN methods, one may argue that these methods are not compatible to each other and in such a situation how a truce between two kinds of approaches be useful to conduct the meaningful research?

Notwithstanding the philosophical debate between QN and QL approach. Mixed methods research designs are justified primarily by exploiting the

strength of each method and by combining their respective strengths within one single research design. The argument of incompatibility between two methods on philosophical ground is not valid because of the following reasons: (Bergman, 2008)

i) The QL and QN methods represent large and heterogeneous families of methods which vary within their own family to such an extent that it is difficult to identify unique set of qualities or attributes that encompasses the characteristics of one family of methods, clearly distinct from the characteristics of the members of the other family. Most characteristics encompass either only a subgroup of members of the family or are also applicable to some members of other family. Hence there is a need to re-think the division of labour between QL and QN methods in order to better understand the possibilities and functions of methods to better justify and apply mixed method design.

ii) The proposition of existence of one single reality under QN method is inconsistent with research applications. The emergent structures are based on co-construction between the researchers' selection and understanding of items in a questionnaire, their choice of analytic strategy and their interpretation of the statistical output on the one hand, and the way the respondents interpreted the survey questions with the given social, political, historical and economic context, on the other hand.

   From a methodological perspective, it is not proper to declare one approach more or less valid, valuable or scientific. Instead how to understand and analyze data need to be based to a large extent on the consistency formed between how to understand data in conjunction with the specific research question, rationale, aim etc. only in connection with the specificities of the research goals it makes sense to delimit the nature of reality. Thus, the decision on whether the researcher deals with one single reality, a constructed reality, multiple realities, multiple constructed realities or a co-constructed reality is unrelated to whether patterns in the data are detected via statistical analysis or otherwise.

iii) *A small vs large samples*: The sample size for many QN studies is often quite small and sample size for some QL studies can indeed be large.

iv) Regarding hypothesis testing, many researchers engaged in QL research, pursue conjectures that are embedded either in their research questions, the kind and the way they collect data, the way they analyse the data and the way, they protect themselves from selective reportings of their findings. In many types of well-established statical analyses, the QN methods like cluster analysis, factor analysis, correspondence analysis, multi-dimensional scaling are used in exploring data structures.

In order to avoid misconceptions and mistakes while deciding the research design issues, the following points need to be kept in mind.

a) We refrain ourselves from making specific argument for or against the QL and QN method for a specific research project. We need to focus our efforts more explicitly on embedding and justifying our selected methods according to our research questions, data needs, theoretical grounding and research design.

b) Data collection methods (i.e. unstructured narrative interview, survey research based on closed-ended questions) and data analysis methods (qualitative content analysis, discourse analysis, quantitative content analysis) should be differentiated.

c) We ought to be more aware of the actual inductive and deductive analytic phases of our research projects.

Check Your Progress 2

1) Which constituent of research study does guide the research design?

……

……

……

2) State the functions of research design?

……

……

……

3) What is the distinction between Mono Method and Mixed Method?

……

……

……

……

6.6 THE RATIONALE FOR MIXED METHODS RESEARCH

Mixed methods research has gained tremendous popularity in the social, behavioural and related sciences in recent years. This increasing popularity of mixed methods research is reflected in terms of the claims by mixed method researchers, increase in the publications on this topic and the inclusion of mixed methods designs in the text books. The argument for combining qualitative and quantitative methods is put forwarded on the following grounds.

i) The research methods associated with both quantitative and qualitative research have their own strengths and weaknesses. Combining them allows the researchers to offset their weakness to draw on the strengths of both.

ii) Mixed methods designs enable the researcher to provide a comprehensive account of the area of the enquiry in which he or she is interested if both quantitative and qualitative methods are used. for example, causal explanation of any event or phenomena can be captured by quantitative methods whereas the reason explanation and pattern explanation or the explanation involving processes can better be captured by qualitative methods.

iii) Quantitative and qualitative methods be used to answer different questions in a study. Further mixed methods can be utilized in order to gain complementary views about the same phenomenon or relationship. Research question for the two strands of the mixed study address related aspects of the same phenomenon.

iv) Mixed methods design can be used to develop the research further. Questions for one strand emerge from the inferences of a previous one (sequential mixed methods), or one strand provides hypothesis to be tested in the next one.

v) Mixed methods may also be used to assess the credibility of inferences obtained from one approach (strand). Exploratory and explanatory/ confirmatory questions in such situation, may be useful.

vi) Mixed methods are also used to obtain divergent pictures of the same phenomenon. These divergent findings can ideally be compared and contrasted.

vii) Quantitative research provides an account of structures in social life but qualitative research provides sense of process.

viii) Mixed methods research entails to generate hypothesis by using qualitative method and testing the hypothesis resorting quantitative method within a single project.

ix) Mixed methods research enables to present diversity of views i.e. combining researchers' and participants' perspectives through quantitative and qualitative research respectively by way of uncovering relationships between variables through quantitative research and revealing meaning among research participants through qualitative research.

6.7 FORM OF MIXED METHODS RESEARCH DESIGNS

Connecting qualitative and quantitative methods through sequencing and prioritising, there can be four forms of mixed methods research designs.

1) **Preliminary qualitative inputs to core quantitative method (qual-QUANT)**: With a view to use the strengths of qualitative methods to contribute to core quantitative methods, preliminary qualitative methods are used for collecting and interpreting the quantitative data. Preliminary qualitative methods can also be used as an input to generate hypotheses, develop content for questionnaires etc.

2) **Supplementary qualitative method as a follow up to core quantitative method (QUANT-qual)**: Supplementary qualitative method as a follow up to core quantitative method can be used to extend what is learned from quantitative study. It can demonstrate bases for results, examine reasons for failed hypotheses, explore the theoretical significance of outliers and so on.

3) **Preliminary quantitative inputs to core qualitative methods design (quan-QUAL)**: The study undertaken by using core quantitative method design gets inputs by way of using preliminary quantitative method for collecting and interpreting the qualitative data. Preliminary quantitative

method can also locate major differences between sub-groups, guide purposive sampling, establish results to pursue and so on.

4) **Supplementary quantitative method to core qualitative method design (QUAL-quan)**: Supplementary qualitative method as a follow up to core qualitative method can be used to develop measures of key concepts and relationships across different samples.

6.8 CASE STUDIES OF MIXED METHODS RESEARCH DESIGN

i) Studies where the quantitative component is dominant in terms of the framework of positivism and post positivism paradigm and inductive and deductive logic of enquiry guides the study, quantitative approach is more dominant and the researchers specialize in quantitative work. In such studies, qualitative data and their manipulation are shaped by the nature of research questions raised in the quantitative part of the project. The qualitative data are treated and transformed in effect into quantitative variables with the purpose of fleshing out the explanations required for the quantitative results. In such cases quantitative component follows qualitative component mixed methods. An illustration of a Research Study which has used Mixed Method pre dominated by the Quantitative Method is given below in the box:

CASE 1: Mixed Method Research Design: Conducted the study within the framework of post positivist paradigm using inductive logic of inquiry and analyzing the qualitative data by quantitative technique of data analysis

Study: 'Decent Work: Concept, Measurement and Status in India': Ph.D thesis by Ms. Nausheen Nizami Ph.D Research Scholar, (Economics) IGNOU, N. Delhi

The study was conducted within the framework of positivist paradigm using inductive logic of enquiry with the objectives to (i) examine the applicability of indicators of 'decent work in India (ii) develop indicators and tools to measure 'decent work (iii) evaluate the status of decent work in IT companies and to identify the socio-economic attributes influencing status of decent work, and (iv) know the employer's perspective in provision of decent work'.

Methodology: Decent work sums up the aspirations of the people in their professional life and is a revolutionary agenda of International Labour Office. Decent work is any such work which ensures provision of fair and free employment to all men and women of economically productive age-group in conditions of fairness, equity, security and dignity. It is a multidimensional concept that applies to both formal and informal sectors of an economy. The study is based on primary data collected through well designed questionnaire and telephonic interviews. Measurement of decent work was undertaken using both quantitative and qualitative methods of data analysis. A characteristic feature of this study was to transform the qualitative data in a manner so as to be able to use quantitative techniques of data analysis.

Quantitative Techniques to analyse qualitative data

a) Composite Index construction for Decent Work

With a view to measure the extent and range of decent work provision, decent work indices were constructed. These indices have been made for all the indicators of decent work used in this study by coding, normalising and aggregating the responses as the data was qualitative in nature. Thereafter, the standard formula developed by United Nation's Development Pr*o*gramme (UNDP) for the construction of HDI and other indices was applied.

b) Factor analysis of decent work indicators

The technique of factor-analysis was adopted to short-list the most relevant and reliable indicators of decent work. The principal component method was used to outline components (i.e. indicators in this study) explaining the maximum variance in the dependent variable.

c) Phi ϕ Coefficient of Correlation Analysis

Since the variables were dichotomous in nature, correlation analysis was attempted to test whether there was statistically significant associations between the responses on different decent work indicators. Partial correlation analysis examined the relationship between few socio-demographic variables and composite Decent work Index keeping the effect of other variables constant.

d) Chi-square test for independence

Chi-square test was conducted to test the independence of the results between the two scales of Decent work index (For example, 0 and 0-1) and results of each decent work indicator for those two scales.

Thus an efficient mix of quantitative and qualitative methods were used to deal with the key issues related to decent work in this doctoral thesis.

Major Results:

1) Composite decent work index reflects deficit in the status of decent work for a vast majority (94%) of IT employees.
2) Adequacy of earnings and productive employment are directly associated with decent work which is in consonance with the features of search and matching theory of employment.
3) Longer working hours have a significant impact on the physical and emotional well-being of an employee. A vast majority of IT employees (95%) were found to be working for longer hours.
4) Work-life balance and adequate earning and productive employment were found to be correlated. Majority of the employees in IT sector were found to be earning adequate earnings but their work-life balance was disrupted owing to several reasons including longer working hours. This implies that higher earnings of the employees in the IT industry are necessary

but not a sufficient condition for the employees. Their higher earnings are not resulting into their well-being (mental) as reflected in their deteriorating health status and imbalance between professional and personal lives.

5) Social security mechanisms, social dialogue and fair treatment at work places are major aspects of decent work owing to ensure sustainability of economic well-being of the employees.

6) Decent work reflects decent work place but a decent work place may not always translate into decent work because of the gap of the nature of work and work amenities and environment. This implies that deficit in decent work to the employees is correlated with deficit in decent work places.

7) Age and social class are the prominent attributes in influencing decent work status.

8) Despite the presence of premier educational institutions like IITs, RECs, etc. demand-supply gap in the available manpower for IT industry is the major concern of the employers pushing them to face the problem of adverse selection. A direct implication of this situation is the low employment elasticity of IT industry with respect to its growth and increasing contribution to GDP.

ii) Studies where qualitative component had priority in terms of the frame work of interpretative paradigm and abductive logic of enquiry. In such situations, the researchers identify themselves primarily as qualitative researchers. In such studies, the tools used for data/information collection is mixed in nature. The modalities of the different types of data are maintained and are not treated as commensurate. Integration of data takes place as a part of interpretative process referred to as analytic or interpretative integration. In such studies quantitative techniques can be used to analyse the perceptions of the people which may be qualitative in nature. An illustration of the research study using mix of qualitative and quantitative methods in the manner discussed here in is given below:

CASE 2: Mixed Methods Research Design: Conducted the study within the framework of Qualitative Research Approach using Quantitative Methods for data analysis

Study: Gender specific impacts of climate variation in Agriculture:

By Mamta Mehar, Ph.D Research Scholar, IGNOU, N. Delhi

The Main Objective of the study was to understand the gender differences in agriculture and analyse the linkages between gender and climate variation using mixed method approach. The research questions addressed in this study were: (i) Does there exist difference in gender perception on climate variability? (ii) Whether male and female farmers follow similar coping strategies to mitigate risk and shocks in agriculture? The study focused on to understand how gender and climate variation in agriculture are interrelated with the support of empirical evidences.

Methodology

This study was conducted within the framework of qualitative approach anchored in interpretative paradigm using quantitative methods – questionnaire as a tool and survey as a method to collect the data and multivariate probit model to find out probability in deciding adaption strategy by male and female to climate variability.

This research study used case study method within a socio-cultural framework. Case study methodology derives much of its rationale and methods from ethnography characterized by "an instance in action" in a "real life context when the boundaries between the phenomenon and context are not clear". Keeping in view, the logic of enquiry, the research questions were analysed using both qualitative and quantitative methods. The qualitative analysis was done within the framework of "Interpretative paradigm of qualitative research". Since the research objective was to understand human behaviour within the surrounding context of climate change, where it was difficult to isolate the phenomena of climate variation and perception of males and females towards it, conducting of the present study under interpretive paradigm was appropriate. The associated research strategy used in this study was abduction. Abduction is the logic used to construct descriptions and explanations that are grounded in the everyday activities as well as in the language and meanings used by social actors. The social actors in the present proposal are men and women farmers, and their activities to be analysed are specific to decision making in climate change situation. The interpretative paradigm using abduction strategy was done via literature review as well as synthesis from findings of focus group discussion during the survey. In the wake of theoretical understanding, it was tried to quantify the impact using appropriate empirical analysis and using data from a baseline survey conducted by CCAFS in 2013 of 641 farm households in twelve villages of Vaishali districts in Bihar, India; it was attempted to quantify the gender differentiated impact.

Results

The linkage between gender and climate variation in agriculture was more influenced from subjective experiences of individuals or society. The movement of understanding "is constantly from the whole to the part and back to the whole" (Gadamer, 1976). According to Gadamer, it is a circular relationship. Since the study aims to understand the perception of human in context of climate variation, it cannot be explained entirely by any scientific paradigm where hypothesis are pre-structured and to some extent results are known. The interpretative paradigm points out that positivism, in its first attempt to model the social after the natural sciences, fails to see that unlike nature, social reality exists only insofar as lay members create that reality in meaningful interaction (Fuchs S., 1992: 198). But the point of interpretive paradigm is that we must first understand social worlds from within, before we develop scientific models and explanations (Fuchs, S. 1992: 205). According to Willis (1995) interpretivists are anti-foundationalists, who believe there is no single correct route or particular method to knowledge. Walsham (1995) argues that in the interpretive tradition there are no 'correct' or 'incorrect' theories. Instead, they should be judged according to how 'interesting' they are to the researcher as well as those involved in the same areas.

The linkages between gender and climate variation in context of agriculture are bi-directional. Changes in climate event effect agriculture through reduction in yield, increase in occurrence of pests or crop disease, changes in time of activities for example, changes in winter time could result in change in harvesting or sowing days of respective crops. All this affects farmers' livelihood and work activity in different ways. For example, women (and children) being the ones traditionally charged with water collection roles, have to travel longer distances after drought. Evidences suggest decrease in yield and thereby total availability of food consumption in a household basket usually has been seen as complemented by reduction in meals of female. This exposes them to health issues. Male farmers are usually seen to migrate in search of nonfarm opportunities outside their village domain, putting all burden or responsibilities on women. Moreover mostly in rural areas women are less likely to get education as compare to male, they are less allowed to mobile and do more household related activities, and they have less control over the family assets.

The other direction of gender differentiated impacts in context of climate events in agriculture can be viewed through understanding their coping strategies. The adaption of particular strategies are often influenced by the priorities and options available for women and men for coping these strategies. Constraints on women's mobility, and behavioural restrictions hinders their ability to make decision on various matters and thus they have less choices of coping strategies available. Additionally inequitable access to assets and resources such as land, knowledge, technology, limited power in decision-making, education, health care and food have been identified as determinant factors behind adaptation of different coping strategies of male versus female. The results of quantitative analysis also supports this argument.

The females are mostly illiterate (71 per cent) and have little access to resources. Only 9 per cent of farm household have reported to provide land entitlement to at least one female member. Additionally the gender power dynamics is seen to favour only one gender i.e. male. The decision making role in different household related expenditures as well as in agriculture activities is dominated by male. Considering the perception of male vis-à-vis female farmer about effects of climate change has shown a similar pattern. More than 90 per cent of farmers are aware about climate variation and feels that weather conditions adversely affect their agriculture activities. However again, the decisions to cope with the risk due to climate have suggested different behaviour, priorities by male and female. Survey results suggest that almost 60 per cent of surveyed farmers have adapted at least one coping strategy and in total adoption of more than 30 coping strategies are reported. The results of multivariate probit model suggest that male farmers have higher probability in deciding adaptations through additional jobs as well as specific agriculture strategies such as crop rotation and use of hybrid seed in farming practices.

iii) Studies presenting diversity of views i.e. combining researcher's and participants' perspectives through quantitative and qualitative research. Researchers uncover relationship between variables through quantitative methods and reveal meaning among research participants through qualitative methods. Thus a comprehensive account of the area of the enquiry is presented. An illustration of the study using this form of mixed method research is given below:

CASE 3: Mixed Method Research –used Quantitative Methods for data analysis and Qualitative Methods for revealing and understanding participants' perspective on quality of education

Study: 'Measuring Quality of Education and Its Determinants: Indian Context': Ph.D thesis submitted by Ms. Charu Jain, Ph.D Research Scholar, (Economics) IGNOU, New Delhi

The study was carried out with the objectives to (i) examine the linkages of secondary education with various socio-economic outcomes at all India level and state level, (ii) identify various determinants (including students' family background characteristics, and schools' characteristics) affecting students' performances and teaching efficiencies in senior secondary schools in Delhi. To fulfill these objectives, both quantitative and qualitative research approaches were applied.

Quantitative Method: The quantitative approach was used to test the hypothesis generated. Two set of questionnaires one related to the teachers and the other to students were got filled up through in-depth face to face interviews with the respondents. The student questionnaire was translated in Hindi language to enhance the understanding level of students in Hindi medium government schools in Delhi. Within translation procedure, the student questionnaire was first translated in Hindi and thereafter re-translated in English to re-check the accuracy and meaning of questions. The teachers' questionnaire was provided in English language only. The information thus retrieved from these questionnaires was used to measure the performance of students and teacher analyzing thereby the overall quality of education at lower secondary and senior secondary levels of education. The information collected through quantitative approach was statistically analyzed and tested using various statistical softwares like SPSS and STATA?

Qualitative Method: In the qualitative approach, direct observation method and key informer interviews were used to supplement the information/data collected through quantitative method by way of two sets of structured questionnaires. Qualitative research method enabled the researcher to approach the research topic from the perspective of teachers and students. Further, the qualitative techniques were used to examine social processes which otherwise could not have been captured by traditional quantitative measures.

The direct observation technique enabled researcher to learn about the behaviour of the people under study: students and teachers in the natural setting i.e. schools through observing. It gave the researcher the first hand information on the quality aspects of school and classrooms and working conditions for teachers, which were quite useful in doing the analysis.

The field notes, which were taken down during the period of fieldwork, were instrumental during analysis and interpretation of the results, as they helped in recalling the incidents that took place at the time of data collection. *Face-to-face interviews* with the key informants were also conducted within qualitative approach. Key informant interviews were qualitative informal interviews with people who knew what was going on in the community. Our purpose of conducting key informant interviews was to collect information from people who had firsthand knowledge about the overall functioning of school. The key informers in the sample schools were selected using purposive sampling technique and in this case the key informants were either senior teacher in school or senior administrative person in schools. Interviewing key informants enabled the researcher to look at the underlying issues and problems with varying perspectives. These interviews were conducted using an informal approach. Initially the questions were drafted in a form of well designed open ended short questionnaire for conducting face to face interviews with the respondents. One respondent or a group of 2-3 people from each school were selected purposively for gathering this information, depending upon how well they were informed about the school.

Apart from these informal interviews, the school guards/receptionists were also contacted informally to gather the information on schools. The information thus collected through qualitative approach was first entered in Excel in a pre-designed format and then coded so that it can be further analysed statistically. Moreover, the responses to few open ended questions were reviewed and inferences were drawn which were quite useful in providing suggestions and recommendation for this study.

Main Results

1) Secondary education increases income level of individuals to a significant extent. One unit increase in Gross Enrollment Ratio (GER) at secondary level leads to increase in the personal disposable income by 0.65 units.
2) The educational background of households determine their occupational pattern. The households with at least secondary education are either salaried or business class, while those with elementary level of education are more concentrated in agricultural/wage earning activities. Higher the level of education among females, greater the proportion of female work participation rate.
3) Secondary education bears highest impact on improving health indicators. Secondary education has high potential for bringing demographic transitions in terms of decline in maternal mortality rates, infant mortality rates, fertility rates, death rates etc.
4) Secondary education attainment brings behavioral changes in individuals in terms of higher savings and diversion in the expenditure pattern.
5) School resources, teachers and teaching quality, student's family background, mass media exposure and self motivational factors of students determine their outcome to significant extent.
6) Within school characteristics, cleanliness in school, well qualified teachers their friendly and supportive attitude with students and quality of school infrastructure influence students performance positively. Lack of student motivation, large class size, lack of school resources and facilities, lack of adequate teaching material affect the teaching abilities adversely.

Check Your Progress 3

1) Give two situations where Mixed Method research is superior to Mono Method?

……

……

……

……

2) Give an example of Mixed Method where qualitative method pre-dominates quantitative method?

……

……

……

……

3) State the different forms of Mixed Methods?

……

……

……

……

6.9 LET US SUM UP

Traditionally two distinct research approaches – Quantitative (QN) and Qualitative (QL) – have been followed in undertaking research in social sciences. The former is associated with positivism and post positivism paradigm and the latter principally with interpretative paradigm and critical theory paradigm. Depending upon the nature and field of specialization research studies may be of various types. Important among these are – theoretical and applied, descriptive and explanatory, conceptual and empirical, exploratory and experimental etc.

Research design is a logical structure of enquiry specifying the types of evidences needed to answer the research questions. Research methods are made of tools for data collection, type of data, sampling design, sample size, techniques for data analysis etc. There is nothing intransic about any research design that requires a particular method of data collection. Research Methods can broadly be of two types – Mono Methods and Mixed Methods. Within mono methods, it can be either quantitative or qualitative. The combination of at least one qualitative and at least one quantitative component in a single research project is known as mixed methods research. The mixed methods have gained popularity in carrying out research in social and behavioral sciences because combining them allows the researchers (i) to offset their (QN and QL methods) weaknesses, (ii) provide comprehensive account of the area of the enquiry, (iii) to answer different types of questions. Studies undertaken under mixed methods can be of four types – (i) preliminary quantitative methods

input design and core qualitative methods as core design (quan-QUAL), (ii) supplementary quantitative methods design as follow up to core quantitative methods design (QUAL-quan), (iii) preliminary qualitative methods input design and core quantitative methods design (qual-QUAN), (iv) supplementary qualitative methods design as follow up to quantitative methods design (QUAN-qual).

6.10 SOME USEFUL BOOKS

Bergman Manfred Max (2008): *The Strawmen of Qualitative-Quantitative Divide and Their Influence on Mixed Method Research in Advances in Mixed Methods Research*, edited by Manfred Max Bergman, Sage Publications Ltd. London.

Creswell John W, Clark Kicki L Plano and Amanda L. Garrett (2008): *Methodological Issues in Conducting Mixed Methods Research Designs in Advances in Mixed Methods Research*, Sage Publications Ltd. London.

Tashakkori Abbas and Teddlie Charles (2008): *Quality of inferences in Mixed Methods Research in Advances in Mixed Methods Research*, Sage Publications Ltd. London.

Morgan David L (2014): *Integrating Qualitative and Quantitative Methods- A Pragmatic Approach,* Sage Publications, London, N. Delhi, Singapore.

6.11 ANSWERS OF HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) A variable expressed in terms of numerals like 1, 2, 3,.. etc. is called quantitative variable. For instance, yield of an agricultural output, height/weight of individuals, etc. can be measured and expressed in quantitative terms. On the other hand, a variable expressed in terms of the choices or opinion about attributes/ characteristics like one's feelings or opinion is called qualitative variable.

2) See Sub-section 6.2.1

3) See Sub-section 6.2.3

4) Descriptive research describes a situation, a phenomenon or event or a social system. It aims to describe the state of affairs as it exists. On the other hand, explanatory results aim to establish cause and effect relationship.

Check Your Progress 2

1) Research questions guide the research design.

2) There are two major functions of the research design: (i) to determine the type of evidences (data/information) which are needed to answer the research questions, (ii) to eliminate the alternative explanation of the results.

3) If a research study is confined to use **either** quantitative **or** qualitative method, it is said to be a mono method research. The combination of at least one quantitative and at least one qualitative component in a single research project is known as mixed methods research.

Check Your Progress 3

1) See Section 6.6
2) See Section 6.7
3) See Section 6.7

UNIT 7 DATA COLLECTION AND SAMPLING DESIGN

Structure

7.0 Objectives

7.1 Introduction

7.2 An Overview of the Unit

7.3 Method of Data Collection

7.4 Tools of Data Collection

7.5 Sampling Design

- 7.5.1 Population and Sample Aggregates and Inference
- 7.5.2 Non-Random Sampling
- 7.5.3 Random or Probability Sampling
- 7.5.4 Methods of Random Sampling
  - 7.5.4.1 Simple Random Sampling with Replacement (SRSWR)
  - 7.5.4.2 Simple Random Sampling without Replacement (SRSWR)
  - 7.5.4.3 Interpenetrating Sub-Samples (I-PSS)
  - 7.5.4.4 Systematic Sampling
  - 7.5.4.5 Sampling with Probability Proportional to Size (PPS)
  - 7.5.4.6 Stratified Sampling
  - 7.5.4.7 Cluster Sampling
  - 7.5.4.8 Multi-Stage Sampling

7.6 The Choice of an Appropriate Sampling Method

7.7 Let Us Sum Up

7.8 Some Useful Books

7.9 Answers or Hints to Check Your Progress Exercises

7.0 OBJECTIVES

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

- appreciate different methods of collecting data;
- acquire knowledge of different tools of data collection;
- define the key terms commonly used in quantitative analysis, like parameter, statistic, estimator, estimate, inference, standard error, confidence intervals, etc.;
- distinguish between random and non-random sampling procedures for data collection;
- appreciate the advantages of random sampling in the assessment of the "precision" of the estimates of population parameters;
- acquire knowledge of the procedure for drawing samples by different methods;

- develop the ability to obtain estimates of key parameters like population and sample aggregates, proportion, mean, etc.; and of the "precision" of such estimates under different sampling methods; and
- appreciate the feasibility/appropriateness of applying different sampling methods in different research contexts.

7.1 INTRODUCTION

Studies of the behaviour of variables and of relationships amongst them necessitate measurement of the variables involved. Variables can be *quantitative variables* like GNP or *qualitative variables* like opinions of individuals on, say, ban on smoking in public places. The former set assumes quantitative values. The latter set does not admit of easy quantification, though some of these can be categorised into groups that can then be assigned quantitative values. Research strategies thus adopt two approaches, quantitative and qualitative. We shall deal with the quantitative approach in this unit. The various measurement scales and scaling techniques to measure the qualitative data will be discussed in Unit 8.

The basic ingredient of quantitative research is the measurement in quantitative terms of the variables involved or the collection of the data relevant for the analytical and interpretative processes that constitute research. The quality of the data utilised in research is important because the use of faulty data in such endeavour results in misleading conclusions, however sophisticated may be the analytical tools used for analysis. Research processes, be it testing of hypotheses and models or providing the theoretical basis for policy or review of policy, call for objectivity, integrity and analytical rigour in order to ensure academic and professional acceptability and, above all, an effective tool to tackle the problem at hand. Data used for research should, therefore, reflect, as accurately as possible, the phenomena these seek to measure and be free from errors, bias and subjectivity. Collection of data has thus to be made on a scientific basis.

7.2 AN OVERVIEW OF THE UNIT

How to assemble data on a scientific basis? There are broadly three different methods of collecting data. These are dealt with in section 7.3. The tools that one can use for collecting data – the formats and the devices that modern technology has provided – are enumerated in section 7.4. There are situations where it is considered desirable to gather data from only *a part of* the universe, or a *sample* selected from the universe of interest to the study at hand, rather than a complete coverage of the universe, for reasons of cost, convenience, expediency, speed and effort. Questions then arise as to the manner in which such a *sample* should be chosen – the sampling design. This question is examined in detail in section 7.5. The discussion is divided into a number of sub-topics. Concepts relating to population aggregates like mean and variance and similar aggregates from the sample and the use of the latter as estimates of population aggregates have been introduced in sub-section 7.5.1. There are two types of sampling: random and non random. Non random sampling methods and the contexts in which these are used are described in sub-section 7.5.2. A random sample has certain advantages over a non random sample — it provides a basis for drawing *valid* conclusions from the sample about the parent population. It enables us to state the precision of the estimates of

population parameters in terms of (a) the extent of their variation or (b) an interval within which the value of the population parameter is likely to lie with a given degree of certainty. Further, it even helps the researcher to determine the size of the sample to be drawn if his project is subject to a sanctioned budget and permissible limits of error in the estimate of the population parameter. These principles are explained in sub-section 7.5.3. Eight methods of random sampling are then detailed in sub-section 7.5.4. These details relate to (i) operational procedures for drawing samples, and (ii) expressions for (a) estimators of parameters and measures of their variation and b) estimators of such variation where the population variation parameter is not known. Different sampling procedures are also compared, as we go along, in terms of the relative precision of the estimates, they generate. Finally, the question of choosing the sampling method that is appropriate to a given research context is addressed in section 7.6. A short summing up of the Unit is given in section 7.7.

7.3 METHOD OF DATA COLLECTION

There are three methods of data collection – *the Census and Survey Method, the Observation Method and the Experimental Method.* The first is a carefully planned and organised study or enquiry to collect data on the subject of the study/enquiry. We might for instance organise a study on the prevalence of the smoking habit among high school children – those aged 14 to 17 - in a certain city. One approach is to collect data of the kind we wish to collect on the subject matter of the study from *all* such children in *all* the schools in the city. In other words, we have a ***complete enumeration*** or ***census*** of the ***population*** or ***universe*** relevant to the enquiry, namely, the city's high school children (called the *respondent units or informants* of the Study) to collect the data we desire. The other is to confine our attention to a *suitably selected part of the population* of high school children of the city, or a ***sample,*** for gathering the data needed. We are then conducting a *sample survey.* A well known example of Census enquiry is the **Census of Population** conducted in the year **2011,** where data on the demographic, economic, social and cultural characteristics of *all persons* residing in India were collected. Among **sample surveys** of note are the household surveys conducted by the National Sample Survey Organisation (NSSO) of the Government of India that collect data on the socio-economic characteristics of a sample of households spread across the country.

The Observation Method records data as things occur, making use of an appropriate and accepted method of measurement. An example is to record the body temperature of a patient every hour or a patient's blood pressure, pulse rate, blood sugar levels or the lipid profile at specified intervals. Other examples are the daily recording of a location's maximum and minimum temperatures, rainfall during the South West / North East monsoon every year in an area, etc.

The Experimental Method collects data through *well designed and controlled* statistical experiments. Suppose for example, we wish to know the rate at which manure is to be applied to crops to maximise yield. This calls for an experiment, in which all variables other than manure that affect yield, like water, quality of soil, quality of seed, use of insecticides and so on, need to be controlled so as to evaluate the effect of different levels of manure on the yield. Other methods of conducting the experiment to achieve the same objective without controlling "all other factors" also exist. Two branches of statistics – The Design and Analysis of Experiments and Analysis of Variance — deal with these.

7.4 TOOLS OF DATA COLLECTION

How do we collect data? We translate the data requirements of the proposed Study into items of information to be collected from the respondent units to be covered by the study and organise the items into a logical format. Such a format, setting out the items of information to be collected from the respondent units, is called ***the questionnaire or schedule*** of the study. The questionnaire has a set of pre-specified questions and the replies to these are recorded either by the respondents themselves or by the investigators. The *questionnaire approach* assumes that the respondent is capable of understanding and answering the questions all by himself/herself, as the investigator is not supposed, in this approach, to influence the response in any manner by interpreting the terms used in the questions. Respondent-bias will have to be minimised by keeping the questions simple and direct. Often the responses are sought in the form of "yes", "no" or "can't say" or the judgment of the respondent with reference to the perceived quality of a service is graded, like, "good", "satisfactory" or "unsatisfactory".

In the *schedule approach* on the other hand, the questions are detailed. *The exact form of the question to be asked of the respondent is not given to the respondent and the task of asking and eliciting the information required in the schedule is left to the investigator.* Backed by his training and the instructions given to him, the investigator uses his ingenuity in explaining the concepts and definitions to respondents to obtain reliable information. This does not mean that investigator-bias is more in the schedule approach than in the questionnaire approach. Intensive training of investigators is necessary to ensure that such a bias does not affect the responses from respondents.

Schedules and questionnaires are used for collecting data in a number of ways. Data may be collected by personally contacting the respondents of the survey. Interviews can also be conducted over the telephone and the responses of the respondent recorded by the investigator. The advent of modern electronic and telecommunications technology enables interviews being done through e-mails or by 'chatting' over the internet. The mail method is one where (usually) questionnaires are mailed to the respondents of the survey and replies received by mail through (postage pre-paid) business-reply envelopes. The respondents can also be asked (usually by radio or television channels or even print media) to send their replies by SMS to a mobile telephone number or to an e-mail address.

Collection of data can also be done through mechanical, electro-mechanical or electronic devices. Data on arrival and departure times of workers are obtained through a mechanical device. The time taken by a product to roll off the assembly line and the time taken by it to pass through different work stations are recorded by timers. A large number of instruments are used for collecting data on weather conditions by meteorological centres across the country that help assessing current and emerging weather conditions. Electronic Data Transfers (EDT) can also be the means through which source agencies like ports and customs houses, where export and import data originate, supply data to a central agency like the Directorate General of Commercial Intelligence and Statistics (DGCI&S) for consolidation.

The above methods enable us to collect *primary data,* that is, data being *collected afresh* by the agency conducting the enquiry or study. . The agency

concerned can also make use of data on the subject *already collected* by another agency or other agencies – *secondary data.* Secondary data are published by several agencies, mostly Government agencies, at regular intervals. These can be collected from the publications / compact discs or the websites of the agencies concerned. But such data have to be examined carefully to see whether these are suitable or not for the study at hand before deciding to collect new data.

Errors in data constitute an important area of concern to data users. Errors can arise due to confining data collection to a sample. (*sampling errors*). It can be due to faulty measurement arising out of lack of clarity about what is to be measured and how it is measured. Even when these are clear, errors can creep in due to inaccurate measurement. Investigator bias also leads to errors in data. Failure to collect data from respondent units of the population or the sample due to omission by the investigator or due to non-response (respondents not furnishing the required information) also results in errors. (*non-sampling errors*). The *total survey error* made up of these two types of errors need to be minimised to ensure quality of data.

7.5 SAMPLING DESIGN

We have looked at methods and tools of data collection, chief among which is the sample survey. How to select a sample for the survey to be conducted? There are a number of methods of choosing a sample from a universe. These consist of two categories, *random sampling* and *non-random sampling*. Let us turn these methods and see how well the results from the sample can be utilised to draw conclusions about the parent universe.

But first let us turn to some notations, concepts and definitions.

7.5.1 Population and Sample Aggregates and Inference

Let us denote population characteristics by upper case (capital) letters in English or Greek and sample characteristics by lower case (small) letters in English. Let us consider a (finite) population consisting of N units U_i (i = 1,2,….N). Let Y_i (i = 1,2,….N) be the value of the variable y, the characteristic under study, for the i^{th} unit U_i (i = 1,2,…..N). For instance, the units may be the students of a university and y may be their weight in kilograms. Any function of the population values Y_i is called a ***parameter***. An example is the population mean ‘μ’ or ‘M’ given by $(1/N) \sum_i Y_i$, where $\sum_i$ stands for summation over i = 1 to N. Let us now draw a sample of ‘n’ units u_i (i = 1,2,…..n)[1] from the above population and let the value of the i^{th} sample unit be y_i (i = 1,2,…..n)[2]. In other words, y_i (i = 1,2,….n) are the sample observations. A function of the sample observations is referred to as a ***statistic***. The sample mean ‘m’ given by $(1/n) \sum_i y_i$, $\sum_i$ (i =1 to n), is an example of a statistic.

Let us note the formulae for some important parameters and statistics.

Population total $Y = \sum_i Y_i$, $\sum_i$ stands for summation over i = 1 to N (2.1)

Population mean = ‘μ’ or ‘M’ , $= (1/N) \sum_i Y_i$, $\sum_i$, i = 1 to N (2.2)

[1] The sample units are being referred to as u_i (i = 1,2,…..n) and not in terms of U_i as we do not know which of the population units have got included in the sample. Each u_i in the sample is some population unit

[2] The same reasons apply for referring the sample values or observations as y_i (i = 1,2,…n) and not in terms of the population values Y_i . y_i wil be some Y_i.

Population variance $\sigma^2 = (1/N) \quad {}_iY_i^2 - M^2$, $\quad {}_i$, i = 1 to N (2.3)

Population SD = $\sigma = +\sqrt{[(1/N) \quad {}_iY_i^2 - M^2]}$, $\quad {}_i$, i = 1 to N (2.4)

Sample mean = $(1/n) \quad {}_iy_i$, $\quad {}_i$, (i =1 to n) (2.5)

Sample variance $s^2 = (1/n) \quad {}_i\, y_i{}^2 - m^2$, $\quad {}_i$ i = 1 to n (2.6)

$= [ss]^2 /n$, where $[ss]^2 = \quad (y_i - m)^2 = \quad y_i{}^2 - n\, m^2 =$ sum of squares of sample observations from their mean 'm' (2.7)

Sample standard deviation 's' = $+\sqrt{[(1/n) \quad {}_i\, y_i{}^2 - m^2]}$, $\quad {}_i$, i = 1 to n (2.8)

Population proportion $P = (1/N) \quad {}_iY_i = N^1 / N$, (where N^1 is the number of units in the population possessing a specified characteristic) (2.9)

$\sigma^2 = (1/N) \quad {}_i\, Y_i^2 - M^2 = P - P^2 = P(1 - P) = PQ,$

where $Q = [(N - N^1)/N] = (1 - P)$. (2.10)

m = p, (proportion of units in the sample with the specific characteristic) (2.11)

$s^2 = p(1 - p) = pq$, where p is the sample proportion and p + q = 1 (2.12)

$[ss]^2 = npq$ (2.13)

The purpose of drawing a sample from a population is to arrive at some conclusions about the parent population from the results of the sample. This process of drawing conclusions or making inferences about the population from the information contained in a sample chosen from the population is called ***inference.*** Let us see how this process works and what its components are. The sample mean 'm', for example, can serve as an estimate of the value of the population mean 'μ'. The statistic 'm' is called an ***estimator (point estimator)*** of the population mean 'μ'. The value of 'm' calculated from a specific sample is called an ***estimate (point estimate)*** of the population mean 'μ'. In general, *a function of sample observations, that is, a statistic, which can be used to estimate the unknown value of a population parameter, is an* ***estimator*** *of the population parameter. The value of the estimator calculated from a specific sample is an* ***estimate*** *of the population parameter.*

The estimate 'm_1' of the population parameter 'μ', computed from a sample, will most likely be different from 'μ'. There is thus an *error* in using 'm_1' as an estimate of 'μ'. This error is the sampling error, assuming that all measurement errors, biases etc., are absent, that is, *there are no non-sampling errors.* Let us draw another sample from the population and compute the estimate 'm_2' of 'μ'. 'm_2' may be different from 'm_1' and also from 'μ'. Supposing we generate in this manner a number of estimates m_i (i = 1,2,3,…….) of 'μ' by drawing repeated samples from the population. All these m_i (i = 1,2,3,….) would be different from each other and from 'μ'. What is the extent of the variability in the m_i (i = 1,2,3,….), or, the variability of the error in the estimate of 'μ' computed from different samples? How will these values be spread or scattered around the value of 'μ' or the errors be scattered around zero? What can we say about the estimate of the parameter obtained from the specific sample that we have drawn from the population as a means of measuring the parameter, without actually drawing repeated samples? How well do non-random and random samples answer these questions? The answers to these questions are important from the point of view of inference.

Let us first look at the different methods of non-random sampling and then move on to random sampling.

7.5.2 Non-Random Sampling

There are several kinds of non-random sampling. **A judgment sample** is a sample that has been selected by making use of one's expert knowledge of the population or the universe under consideration. It can be useful in some circumstances. An auditor for example could decide, on the basis of his experience, on what kind of transactions of an institution he would examine so as to draw conclusions about the quality of financial management of an institution. **Convenience Sampling** is used in exploratory research to get a broad idea of the characteristic under investigation. An example is one that consists of some of those coming out of a movie theatre; and these persons may be asked to give their opinion of the movie they had just seen. Another example is one consisting of those passers by in a shopping mall whom the investigator is able to meet. They may be asked to give their opinion on a certain television programme. The point here is the convenience of the researcher in choosing the sample. **Purposive Sampling** is much similar to judgement sampling and is also made use of in preliminary research. Such a sample is one that is made up of a group of people specially picked up for a given purpose. **In Quota Sampling**, subgroups or strata of the universe (and their shares in the universe) are identified. **A convenience** or **a judgement sample** is then selected from each stratum. No effort is made in these types of sampling to contact members of the universe who are difficult to reach. In Heterogeneity Sampling units are chosen to include all opinions or views. **Snowball Sampling** is used when dealing with a rare characteristic. In such cases, contacting respondent units would be difficult and costly. This method relies on referrals from initial respondents to generate additional respondents. This technique enables one to access social groups that are relatively invisible and vulnerable. This method can lower search costs substantially but this saving in cost is at the expense of the representative character of the sample. An example of this method of sampling is to find a rare genetic trait in a person and to start tracing his lineage to understand the origin, inheritance and etiology of the disease.

It would be evident from the description of the methods given above that the relationship between the sample and the parent universe not clear. The selection of specific units for inclusion in the sample seem to be *subjective* and *discretionary* in nature and, therefore, may well reflect the researcher's or the investigator's attitudes and bias with reference to the subject of the enquiry.

A sample has to be *representative* of the population from which it has been selected, if it is to be useful in arriving at conclusions about the parent population. ***A representative sample*** *is one that contains the relevant characteristics of the population in the same proportion as in the population.* Seen from this angle, the non-random sampling methods described above do not yield representative samples. Such samples are, therefore, not helpful in drawing *valid* conclusions about the parent population and *the way these conclusions change* when another sample is chosen from the population. Non-random sampling is, however, useful in certain circumstances. For instance, it is an inexpensive and quick way to get a preliminary idea of the variable under study or a rough preliminary estimate of the characteristics of the universe that helps us to design a scientific enquiry into the problem later. It is thus useful in exploratory research.

Check Your Progress 1

1) Name the three methods of the data collection

……

……

……

2) What is meant by a 'parameter'? Defining the term 'statistic', indicate the expressions for population/sample mean and population/sample mean and population/sample variance.

……

……

……

3) Mention the most important purpose of studying the population on the basis of a sample. In this context, define the terms 'estimator' and 'estimate' with a suitable example.

……

……

……

……

4) Defining the term 'representative sample'. Indicate how is 'random sampling' principally different from that of 'non-random sampling'? What could be the use of the latter despite its major drawback vis-à-vis the former?

……

……

……

……

7.5.3 Random or Probability Sampling

Random sampling methods, on the other hand, yield samples that are representative of the parent universe. The selection process in random sampling is free form the bias of the individuals involved in drawing the sample as the units of the population are selected at random for inclusion in the sample. *Random sampling is a method of sampling in which each unit in the population has a predetermined chance (probability) of being included in the sample. A sampling design is a clear specification of all possible samples of a given type with their corresponding probabilities.* This property of random sampling helps us to answer the questions we raised at the end of sub-section 2.6.1 above. That is, we can make estimates of the characteristics of the parent population from the results of a sample and also indicate the extent of error to which such estimates are subject or the *precision of the estimate*. This is better than not knowing anything at all about the magnitude of the error in our statements regarding the parent population. Let us see how random sampling helps in this regard.

A) Precision of Estimates – Standard Errors and Confidence Intervals

We noted earlier (the last paragraph of sub-section 7.5.1) that the sample mean (an estimate of the population mean 'μ') will have different values in repeated samples drawn from the population and none of these may be equal to 'μ'. Suppose that the repeated samples drawn from the population are *random* samples. The sample mean computed from a *random sample* is a *random variable*. So is the sampling error, that is, the difference between'μ' and the sample mean. The values of the sample means (and the corresponding errors in the estimate of 'μ') computed from the repeated random samples drawn from the population are the values assumed by this random variable with probabilities associated with drawing the corresponding samples. These will trace out a frequency distribution that will approach a probability distribution when the *number* of random samples drawn increases indefinitely. *The probability distribution of sample means computed from all possible random samples from the population is called the* ***sampling distribution of the sample mean***. The sampling distribution of the sample mean has a mean and a standard deviation. *The sample mean is said to be an* ***unbiased estimator*** *of the population mean if the mean of the sampling distribution of the sample mean is equal to the mean of the parent population, say, μ. In general, an estimator "t" of a population parameter "θ" is an unbiased estimator of "θ" if the mean of the sampling distribution of "t", or the expected value of the random variable "t", is equal to "θ"*. In other words, the mean of the estimates of the parameter made from all possible samples drawn from the population will be equal to the value of the parameter. Otherwise, it is said to be a ***biased estimate***. Supposing the mean of the sampling distribution of sample mean is Kμ or K+μ, where K is a constant. The bias in the estimate can be easily corrected in such cases by adopting m/K or (m – K) as the estimator of the population mean.

The variance of the sampling distribution of the sample mean is called the ***sampling variance of the sample mean***. The standard deviation of the sampling distribution of sample means is called ***the standard error (SE) of the sample mean***. It is also called the ***standard error of the estimator (of the population mean)***, as the sample mean is an estimator of the population mean. *The standard error of the sample mean is a measure of the variability of the sample mean about the population mean or a measure of the* ***precision*** *of the sample mean as an estimator of the population mean.* The ratio of the standard deviation of the sampling distribution of sample means and the mean of the sampling distribution is called the ***coefficient of variation (CV)*** of the sample mean or the ***relative standard error (RSE)*** of the sample mean. That is,

$$\text{CV or RSE} = C = \text{standard deviation / mean} \qquad (2.14)$$

CV (or RSE) is a free number or is dimension-less, while the mean and the standard deviation are in the same units as the variable 'y'. (These definitions can easily be generalised to the sampling distribution of any sample statistic and its SE and RSE.)

We have talked about the unbiasedness and precision of the estimate made from the sample. What more can we say about the precision of the estimate and other characteristics of the estimate? *This is possible if we know the nature of the sampling distribution of the estimate.*

The nature of the sampling distribution of, say, the sample mean, or for that matter any statistic, depends on the nature of the population from which the random sample is drawn. If the parent population has a normal distribution with mean μ and variance σ^2 or, in short notation, N (μ, σ^2), the sampling distribution of the sample mean, based on a random sample drawn from this, is N (μ, σ^2/n). In other words, the variability of the sample mean is much smaller than that of the variable of the population and it also *decreases* as the sample size *increases*. ***Thus, the precision of the sample mean as an estimate of the population mean increases as the sample size increases***.

As we know, the normal distribution N (μ, σ^2) has the following properties:

i) Approximately 68% of all the values in a normally distributed population lie within a distance of one standard deviation (plus and minus) from the mean,

ii) Approximately 95% of all the values in a normally distributed population lie within a distance of 1.96 standard deviation (plus and minus) of the mean,

iii) Approximately 99% of all the values in a normally distributed population lie within a distance of 2.576 standard deviation (plus and minus) of the mean.

The statement at (iii) above, for instance, is equivalent to saying that the population mean μ will lie between the observed values ($y - 2.576\ \sigma$) and ($y + 2.576\ \sigma$) in 99% of the random samples drawn from the population N(μ, σ^2). Applying this to the sampling distribution of the sample mean, which is N(μ, σ^2/n), we can say that

$$\Pr.[(m - 2.576\ \sigma / \sqrt{n}) < \mu < (m + 2.576\ \sigma / \sqrt{n})] = 0.99 \qquad (2.15)$$

or that the population mean μ will lie between the limits computed from the sample, namely, ($m - 2.576\ \sigma/\sqrt{n}$) and ($m + 2.576\ \sigma\sqrt{n}$) in 99% of the samples drawn from the population. This is an ***interval estimate,*** or a ***confidence interval,*** for the parameter with a ***confidence coefficient of 99%*** derived from the sample.

The general rule for constructing a confidence interval of the population mean with a confidence coefficient of 99% is: the lower limit of the confidence interval is given by the "estimate of the population mean minus 2.576 times the standard error of the estimate" and the upper limit of the interval by the "estimate plus 2.576 times the standard error of the estimate". (2.16)

B) **Assessment of Precision – Unknown Population Variance**

If the parent population is distributed as N (M, σ^2) and σ^2 is not known, we make use of an estimate of σ^2. The statistic 's^2 ' given in formula *2.6* can be one such, but this is not an unbiased estimate of σ^2 as E (s^2) = [(n – 1)/ n] σ^2. We, therefore, by using (2.6) and (2.7) have:

$v(y)=[ns^2/(n-1)]$ as an unbiased estimate of σ^2. (2.17)

$v(y) = [1/(n-1)]\ [ss]^2$ or, $v(y) = [(1/(n-1)][\ \ _i y_i^{\ 2} - nm^2\ ]$ (2.18)

As the sampling variance of the sample mean 'm' is σ^2/n, *an unbiased estimate v(m) of the sampling variance will be v(y)/n.* Let us now consider the statistic defined by the ratio,

$$t = (m - M) / [\sqrt{v(y)} / \sqrt{n}] \quad (2.19)$$

The numerator is a random variable distributed as $N(0, \sigma^2/n)$ and the denominator is the square root of the *unbiased estimate of its variance*. The sampling distribution of the statistic 't' is the Student's t-distribution with (n – 1) degrees of freedom. It is a symmetric distribution. A confidence interval can now be constructed for the population mean M from the selected random sample, say with a confidence coefficient of $(1 - \alpha)\%$. The values of 't_α' for different values of $\alpha = Pr.[t > t_\alpha] + Pr.[(- t) < (- t_\alpha)] = 2\ Pr.[t > t_\alpha]$ and different degrees of freedom have been tabulated in, for instance, Rao, C.R. and Others (1966). The confidence interval with a confidence coefficient $(1 - \alpha)$ for the population mean M would be as in *2.19* below – easily computed from the sample observations.

$$[m - t_\alpha\, v(m) < M < m + t_\alpha\, v(m)] \quad (2.20)$$

We note that the rule 2.16 applies here also except that we use (i) the square root of the unbiased estimate of the sampling variance of the estimate of the population mean in the place of the standard error of the estimate of the population mean, and (ii) the relevant value of the 't' distribution instead of the normal distribution …… (2.21)

We have so far dealt with parent populations that are normally distributed. What will be the nature of the sampling distribution of the sample mean when the ***parent*** *population is not normally distributed?* We examine this question in the next sub-section C.

C) Assessment of Precision–Parent Population has a Non-Normal Distribution

The Central Limit Theorem ensures that, even if the population distribution is not normal,

- the sampling distribution of the sample mean will have a mean equal to the population mean *regardless* of the sample size, and
- as the sample size increases, the sampling distribution of the sample mean approaches the normal distribution.

Thus for large 'n' (sample size), say 30 or more, we can proceed with the steps mentioned in sub-section A above. Further, the Student's t-distribution also approaches the normal distribution as 'n' becomes large so that we can use the statistic 't' in sub-section B *as a normally distributed variable with mean 0 and unit variance for samples of size 30 or more.* We may then adopt the procedure outlined in sub-section A.

D) Determination of Sample Size

Random sampling methods also help in determining the sample size that is required to attain a desired level of precision. This is possible because the standard error and the coefficient of variation C.V. of the estimate, say, sample mean 'm', are functions of 'n', the sample size. C.V. is usually very stable over the years and its value available from past data can be used for determining the

sample size. We can specify the value of C.V. of the sample mean that we desire as, say, C(m) and calculate the sample size with the help of prior knowledge of the population C.V., namely, C. That is,

$$C(m) = C/\sqrt{n} \text{ ; so that } \sqrt{n} = C/C(m), \text{ or, } n = [C/C(m)]^2 \quad (2.22)$$

Or we can define *the desired precision in terms of the error that we can tolerate in our estimate* of 'M' (***permissible error***) and link it with the desired value of C(m). Then,

$$n = [2.576C / e]^2 \text{ , where the permissible error } e = |(m - M)| / M. \quad (2.23)$$

If the sanctioned budget is F for the survey: Let the cost function be of the form $F_0 + F_1 n$, consisting of two components – overhead cost and cost per unit to be surveyed. As this is fixed as F, $F = F_0 + F_1 n$, and the sample size becomes $\mathbf{n = (F - F_0)/F_1}$. The coefficient of variation of C(m) is not at our choice in this situation since it gets fixed once 'n' is determined. We can, however, determine the error in the estimate of 'm' from this sample (*in terms of the RSE of m*), if the population CV, C is known. If further we suppose that the loss in terms of money is proportional to the value of RSE of m, say, Rs. '*l*' per 1% of RSE of 'm', the total cost of the survey becomes, $\mathbf{L(n) = F_0 + F_1 n + l\,(C/\sqrt{n})}$. ***We can then determine the sample size that minimises this new cost (which includes the cost arising out of loss)***. Differentiating L(n) w.r.t n and equating to zero and simplifying,

$$n = [(l/2)\,(C/F_1)]^{2/3} \quad (2.24)$$

See also the sub-section below on stratified sampling.

7.5.4 Methods of Random Sampling

We have so far dealt with random samples drawn form a population. We did not specify the size of the population. We had assumed that the population is infinite in size. In practice, a population may have a size N, however, large. Let us, therefore, consider drawing random samples of size 'n' from a population of size 'N'. We shall consider the following methods of random sampling:

a) Simple Random Sampling (With Replacement) [SRSWR],

b) Simple Random Sampling (Without Replacement) [SRSWOR]

c) Interpenetrating Sub-Samples (I-PSS),

d) Systematic Sampling (sys),

e) Sampling with Probability Proportional to Size (pps)

f) Stratified Sampling (sts),

g) Cluster Sampling (cs) and

h) Multi-Stage Sampling (mss)

We shall indicate in the following sections a description of the above methods, the relevant operational procedure for drawing a sample and the expressions/ formulae for (a) the estimator of the population mean/total/ proportion, (b) the sampling variance of the sample mean/total/population and (c) unbiased estimate of the sampling variance.

Check Your Progress 2

1) Briefly describe the meaning of the terms 'sampling distribution of a statistic' and also what is meant by unbiased estimator of a parameter.

……

……

……

2) How is random sampling procedure helpful in correcting for the bias of an estimate? Illustrate this with the help of an example.

……

……

……

3) Explain the terms 'coefficient of variation' and 'relative standard error'.

……

……

……

7.5.4.1 Simple Random Sampling with Replacement (SRSWR)

The method: This method of drawing samples at random ensures that (i) each item in the population has an equal chance of being included in the sample and (ii) each possible sample has an equal probability of getting selected. Let us select a sample of 'n' units from a population of 'N' units by simple random sampling with replacement (SRSWR). We select the first unit at random, note its identity particulars for collection of data and place it back in the population. We choose at random another unit – this could turn out to be the same unit selected earlier or a different one, note its identity particulars and place it back. We repeat this process 'n' times to get an SRSWR sample of size 'n'. In such a sample one or more units *may* occur more than once. A sample of 'n' distinct units is also possible. It can be shown that the number of possible samples that can be selected by SRSWR method is N^n and that the probability of any one sample being chosen is $1/N^n$.

Operational procedure for selection of the sample by SRSWR method: Tables of Random Numbers are used for drawing random samples. These tables contain a series of four-digit (or five-digit or ten-digit) random numbers. Supposing a sample of 30 units is to be selected out of a population of 3000 units. First allot one number from the set of numbers to 0001 to 3000 as the identification number to each one of the population units. The problem of drawing the sample of size 30 then reduces to that of selecting 30 random numbers, one after another, from the random number tables. Turn to a page of the Tables *at random* and start noting down, from the first left-most column of (four or five or ten-digit) random numbers, the first four digits of the numbers from the top of the column downwards. Continue this operation on to the second column till the required sample size of 30 is selected. If any of the random numbers that comes up is more than 3000, reject it. If some numbers (< 3000) get repeated in the process, it means that the corresponding units of the population would be selected more than once, this being sampling with replacement.

Estimators from SRSWR samples (using notations set down earlier):

$m_{srswr} = (1/n) \sum_i y_i$ ($\sum_i$, i = 1 to n) is an unbiased estimator for M (2.25)

Note: If a unit gets selected in the sample more than once, the corresponding value of y_i will also have to be repeated as many times in the summation for calculating m_{srswr} .

Sampling Variance of m_{srswr} : $V(m_{srswr}) = \sigma^2/n = [1/n]\ [E(y_i^{\ 2}) - M^2]$ (2.26)

Standard Error of m_{srswr} : $SE(m_{srswr}) = \sigma / \sqrt{n}$ (2.27)

CV or RSE of m_{srswr} : $C(m_{srswr}) = (1/\sqrt{n})(\sigma / M) = C(y)/\sqrt{n}$. (2.28)

Note that the sampling variance, SE and CV(RSE) of the sample mean in SRSWR is much less than SE and CV of the variable y and these decrease as the sample size increases. The precision of the sample mean in SRSWR, as an estimator of M increases as the sample size increases. However, the extent of decrease in the standard error will not be commensurate with the size of the increase in the sample size. We would need an unbiased estimator of σ^2, as σ^2 may not be known. This is

$v(y) = [1/(n-1)][ss]^2$ (2.29);

Therefore, $v(m_{srswr}) = v(y)/n$ (2.30)

an unbiased estimate of Y, or, $Y^*_{srswr} = Nm$ (2.31)

$V(Y^*_{srswr}) = N^2 (\sigma^2 /n)$ (2.32)

$v(Y^*_{srswr}) = N^2 (1/n)[1/(n-1)] \quad (y_i - m_{srswr})^2$ (2.33)

the sample proportion 'p_{srswr}' is an unbiased estimate of P (2.34)

$V(p_{srswr})$ is PQ/n (2.35); and $v(p_{srswr}) = npq/(n-1)$ (2.36)

$C(p_{srswr})$-($= \sqrt{[(1/n)(PQ)]}/ P = [1/\sqrt{n})]\sqrt{[Q/P]}$. (2.37)

Confidence intervals for the population mean/proportion and the sample size for a given level of precision and/or permissible error can now be derived easily.

7.5.4.2 Simple Random Sampling without Replacement (SRSWOR)

The Method: This method of sampling is the same as SRSWR but for one difference. If a unit is selected, it is ***not*** placed back before the next one is selected. This means that no unit gets repeated in a sample. Operationally, we draw random numbers between 1 and N and if a random number comes up again, it is rejected and another random number is selected. This process is repeated till 'n' *distinct* units are selected. It can be shown that the number of samples of size n that may be selected from a population of 'N' units by this method is ${}_NC_n = N\ !\ /[\ (N-n)\ !\ n\ !\ ] = [N(N-1)\ (N-2)\ \ldots(N-n+1)]\ /\ [n(n-1)\ (n-2)\ldots\ldots.1]$. *The probability $P_{srswor}(S)$ of any one of the samples being chosen is, $1/{}_NC_n$.*

Estimators from SRSWOR samples:

$m_{srswor} = (1/n) \sum_i y_i$, $\sum_i$, i =1 to n, is an unbiased estimator of M (2.38)

$V(m_{srswor}) = [(N-n)/ (N-1)]\ [\sigma^2/n]$

$= [(N-n)/(N-1)][1/n][(1/N) \sum_i (Y_i - M)^2]$, $\sum_i$, i=1 to N (2.39)

$V(m)_{srswor} < V(m_{srswr})$ since $(N - n) / (N - 1)$ is less than 1 for $n > 1$, (2.40)

Both m_{srswor} and m_{srswr} are unbiased estimators of M *but m_{srswor} is a more efficient estimator of M than m_{srswr} .* The factor $[(N - n)/(N - 1)]$ in *(2.40)* is called the ***finite population correction or finite population multiplier.*** *The finite population correction required for finite population need not, however, be used when the sampling fraction (n / N) is less than 0.05.*

$v(m_{srswor}) = [(N - n)/N][1/n][1/(n - 1)][\ _i (y_i - m)^2]$, $\ _i$, $i = 1$ to n

$=[(N - n)/N][1/n][1/(n - 1)][ss]^2$, (2.41)

Unbiased estimate of population total Y, that is $Y^*_{srswor} = Nm_{srswor}$ (2.42)

$V(Y^*_{srswor}) = N^2\ V(m_{srswor})$ (2.43)

unbiased estimate of $V(Y^*_{srswor})$*, namely,* $v(Y^*_{srswor}) = N^2\ v(m_{srswor}$ (2.44)

$C(Y^*_{srswor}) = C(m_{srswor})$ (2.45)

the sample proportion 'p' is an unbiased estimate of 'P' in SRSWOR also. (2.46)

$V(p) = [(N - n)/ (N - 1)]\ [PQ/n]$, where $P + Q = 1$ (2.47)

$v(p) = [(N - n)/N]\ [pq/(n - 1)]$, where $p + q = 1$ (2.48)

$C(p) = \sqrt{[(N - n)/(N - 1)]}\ \sqrt{[(1/n)}\ \sqrt{[Q/P]}$ and (2.49)

Check Your Progress 3

1) A population has 80 units. The relevant variable has a population mean of 8.2 and a variance of 4.41. These SRSWR samples of size (i) 16, (ii) 25 and (iii) 49 are drawn from the population. What is the standard error (SE) of the sample means from the three samples? Is the extent of reduction in SE commensurate with that of the increase in sample size?

   ……

   ……

   ……

2) What are the results when the sampling method in drawing the three samples in problem 1 above is changed to SRSWOR? What is your advice regarding the choice between increasing the sample size and changing the sampling method from SRSWR to SRSWOR?

   ……

   ……

   ……

3) Indicate whether the following statements are true (T) or false (F). If false, what is the correct position?

   1) The standard error of the sample mean decreases in direct proportion to the sample size. (T/F)

   2) SRSWOR method of sampling is more advantageous than SRSWR for a sampling fraction of 0.02 (T/F)

   3) If Y* = Nm and Variance of m is V(m), the variance of Y* is NV(m). (T/F)

7.5.4.3 Interpenetrating Sub-Samples (I-PSS)

Suppose a sample is selected in the form of two or more sub-samples drawn according to the same sampling method so that each such sub-sample provides a valid estimate of the population parameter. The sub-samples drawn in this way are called ***interpenetrating sub-samples (I-PSS).*** This is operationally convenient, as the different sub-samples could be allotted to different investigators. The sub-samples need not be independently selected. There is, however, an important advantage in selecting *independent interpenetrating sub-samples. It is then possible to easily arrive at an unbiased estimate of the variance of the estimator even in cases where the sampling method/design is complex and the formula for the variance of the estimator is complicated.*

Let $\{t_i\}$, i = 1,2,…,h be unbiased estimates of a parameter θ based on 'h' independent interpenetrating sub-samples. Then,

$t = (1/h) \sum_i t_i$, ($\sum_i$, i = 1 to h) is an unbiased estimate of θ (2.50)

$v(t) = [1/h(h-1)] [\sum_i (t_i - t)^2]$, ($\sum_i$, i = 1 to h) is an unbiased estimate of V(t) (2.51)

If the unbiased estimator 't' of the parameter θ is *symmetrically distributed* (for example, normally distributed), the probability of the parameter θ lying between the maximum and the minimum of the 'h' estimates of θ obtained from the 'h' sub-samples is given by:

$$\text{Prob.}[\text{Min of } \{t_1, t_2, \text{---- } t_h\} < \theta < \text{Max of } \{t_1, t_2, \text{-----} t_h\}] = [1-(1/2)^{(h-1)}] \quad (2.52)$$

This is a confidence interval for θ from the sample. The probability increases rapidly with the number of I-P sub-samples – from 0.5 (two sub-samples) to 0.875 (four sub-samples).

7.5.4.4 Systematic Sampling

The Method: Let $\{U_i\}$, i = 1,2,……. N be the units in a population. Let 'n' be the size of the sample to be selected. Let 'k' be the integer nearest to N/n - denoted usually as [N/n] — the reciprocal of the sampling fraction. Let us choose a random number from 1 to k, say, 'r'. We then choose the r^{th} unit, that is, U_r . Thereafter, we select every k^{th} unit. In other words, we select the units U_r, U_{r+k} , U_{r+2k} ,…… This method of sampling is called *systematic sampling with a random start***.** 'r' is known as *the random start* and 'k' *the sampling interval***.** There would thus be 'k' possible systematic samples, each corresponding to one random start from 1 to k. The sample corresponding to the random start 'r' will be

$$\{U_{r+jk}\}, \quad j = 0,1,2, \; r+jk \le N.$$

The sample size of all the 'k' systematic samples will be 'n' if N = nk. All the 'k' systematic samples will not have a sample size 'n' if N ≠ nk. For example, if we have a sample of 100 units and we wish to select systematic samples of size 14, the sampling interval is k = [100/15] or 7 The samples with the random starting 1 and 2 will be of size 15 while the other 5 systematic samples (with random starts 3 to 7) will be of size 14.

In systematic sampling, units of a population could thus be selected at a uniform interval that is measured in time, order or space. We can for instance choose a sample of nails produced by a machine for five minutes at the interval

of every two hours to test whether the machine is turning out nails as per the desired specifications. Or, we could arrange the income tax returns relating to an area in the order of increasing gross income returned and select every fiftieth tax return for a detailed examination of the income of assesses of the area. Systematic samples are thus operationally easier to draw than SRSWR or SRSWOR samples. Only one random number needs to be chosen for selecting a systematic sample.

Estimators from Systematic Samples:

An unbiased estimator of the population mean M based on a systematic sample is given by a **slight variant of the sample mean**, namely,

$m_{sys}^* = (k/N) \sum_i y_i$, $\sum_i$, i =1 to n*, n* is the size of the selected sample and k the sampling interval v(2.53)

If N = nk, $m_{sys}^* = m$ the sample mean. If N ≠ nk, there is a bias in using the sample mean as the estimator for M, and

the bias in using the sample mean as an estimator of M is likely to be small in the case of systematic samples selected from a large population. (2.54)

The disadvantages, referred to above, in systematic sampling, namely, N not being a multiplier of the sample size n and the sample mean not being an unbiased estimator of the population mean can be overcome by adopting a procedure called ***Circular Systematic Sampling (CSS).*** If 'r' is the random start, and k the integer nearest to N / n, we choose the units.

$\{U_{r+jk}\}$, if r+jk ≤ N and $\{U_{r+jk-N}\}$, if r+jk > N ; j = 0,1,2,…… (n - 1).

Taking the earlier example of selecting a systematic sample of size 14 from a population of 100 units (N == 100, k = 7 and n = 15) all the samples can be made to have a size of 15 by adopting the CSS. A random start of 5 will lead to the selection of a sample of the 15 units 5,12,19,26,33,40,47,54,61, 68.75,82,89,96 and 3 (96 + 7 – 100). This procedure ensures equal probability of selection to every unit in the population.

Besides constancy of the sample size from sample to sample, the CSS procedure ensures that m_r the sample mean is an unbiased estimate of the population mean. (2.55)

Let nk = N. Then m* = m. There are k possible samples, each sample with a probability of 1/k. Let the sample mean of the r-th systematic sample be $m_r = (1/n) \sum_i y_{ir}$, where y_{ir} is the value of the characteristic under study for the i-th unit in the r-th systematic sample, summation is from i = 1 to n. As already noted m_i is an unbiased estimator of M or $E(m_r) = M$. We thus have k possible unbiased estimates of M. Denoting the sample mean in systematic sampling as m_{sys}, the sampling variance of m_{sys}, and related results of interest are:

$V(m_{sys}) = \sigma_b^2$ (the *between*-sample variance). (2.56)

$V(m_{sys}) = V(y) - \sigma_w^2$, where σ_w^2 is *within*-sample variance. (2.57)

Equation *2.57* shows that (i) $V(m_{sys})$ is less than the variance of the variable under study or the population variance, since σ_w^2 is > 0 and (ii) $V(m_{sys})$ can be reduced by increasing σ_w^2**,** or by increasing the *within*-sample variance. (ii) would happen ***if the units within each systematic sample are as heterogeneous as***

possible. Since we select a sample of 'n' units from the population of N units by selecting every k-th element from the random start 'r', the population is divided into 'n' groups and we select one unit from each of these 'n' groups of population units. Units *within* a sample would be heterogeneous if there is heterogeneity *between* the 'n' groups. This would imply that units *within* each of the n groups would have to be as homogeneous as possible. All these suggest that the sampling variance of the sample mean is related to the *arrangement of the units in the population.* This is both an advantage and disadvantage of systematic sampling. An arrangement that conforms to the conditions mentioned above would lead to a smaller sampling variance or an efficient estimate of the population mean while a 'bad' arrangement would lead estimates that are not as efficient.

7.5.4.5 Sampling with Probability Proportional to Size (PPS)

The Sampling Method: We have so far considered sampling methods in which the probability of each unit in the population getting selected in the sample was equal. There are also methods of sampling in which the probability of any unit in the population getting included in the sample varies from unit to unit. One such method is sampling with probability proportional to size (pps) in which the probability of selection of a unit is proportional to a given measure of its size. This measure may be a characteristic related to the variable under study. One example may be the employment size of a factory in the past year and the variable under study may be the current year's output. Does this method lead to a bias in our results, as units with smaller sizes would be under represented in the sample and those with larger sizes would be over represented. It is true that if the sample mean 'm' were to be used to estimate the population mean M, m would be a biased estimator of M. However, ***what is done in this method of sampling is to weight the sample observations with suitable weights at the estimation stage to obtain unbiased estimates of population parameters, the weights being the probabilities of selection of the units.***

Estimates from pps sample of size 1: Let the population units be $\{U_1, U_2, \text{----}, U_N\}$.Let the main variable Y and the related size variable X associated with these units be$\{Y_1, X_1; Y_2, X_2; \ldots\ldots\ldots\ldots Y_N, X_N\}$. The probability of selecting any unit, say, U_i in the sample will be $P_i = (X_i / X)$, where $\sum_i X_i = X$, where $\sum_i$, i = 1 to N. Let us select *one unit* by pps method. Let the unit selected thus have the values y_1 and x_1 for the variables y and x. The variables y and x are random variables assuming values Y_i and X_i respectively with probabilities P_i , i = 1,2,……N. The following results based on the sample of size 1 can be derived easily:

An unbiased estimator of population total Y is $Y^*_{(1)\,pps} = y_1 / p_1$ (2.58)

An unbiased estimator of M is $m^*_{(1)pps} = (1/N)\, Y^*_{(1)pps} = (1/N)(y_1 / p_1)$ (2.59)

$V[Y^*_{(1)\,pps}] = \sum_i (Y_i^2 / P_i) - Y^2$ (2.60)

$V[m^*_{(1)pps}] = (1/N^2)\, V[Y^*_{(1)pps}] = (1/N^2) \left[\sum_i (Y_i^2 / P_i) - Y^2 \right]$ (2.61)

These show that *the variance of the estimate will be small if the P_i are proportional to Y_i.*

Estimators from pps sample of size > 1 [pps with replacement (pps-wr)]

A sample of n (> 1) units with pps can be drawn with or without replacement. Let us consider a pps-wr sample. Let $\{y_i , p_i\}$ be respectively the sample observation on the selected unit and the initial probability of selection at the i–th draw, i = 1,2,----n. Each (y_i / p_i) , i = 1,2,----- n in the sample is an unbiased estimate $[Y_{i(pps\text{-}wr)}^*]$ of the population total Y and $V(Y_{i(pps\text{-}wr)}^*) = \sum_i (Y_r^2 /P_r) – Y^2$, $\sum_r$, r = 1 to n. (see ***2.60***). Estimates from pps-wr samples are:

$$Y^*_{pps\text{-}wr} = (1/n) \sum_i (y_i / p_i) = (1/n) \sum_i Y^*_{(i)pps\text{-}wr} ; \sum_i , i = 1 \text{ to } n. \quad (2.62)$$

$$V(Y^*_{pps\text{-}wr}) = (1/n)[\sum_r (Y_r^2 /P_r) – Y^2] ; \sum_r , r = 1 \text{ to } N. \quad (2.63)$$

$$V(m_{pps\text{-}wr}) = (1/N^2)(1/n)[\sum_r (Y_r^2 /P_r) – Y^2] ; \sum_r , r = 1 \text{ to } N. \quad (2.64)$$

$$v(Y^*_{pps\text{-}wr}) = [1/\{n(n-1)\}][\sum_r (y_r^2 / p_r^2 – nY^{*2}]; \sum_r , r = 1 \text{ to } n; \text{ (using 2.51)} \quad (2.65)$$

Operational procedure for drawing a pps-wr sample: The steps are:

1) Cumulate the sizes of the units to arrive at the cumulative totals of the unit sizes. Thus

   $T_{i-1} = X_1 + X_2 + \text{-----+ } X_{i-1}$; $T_i = X_1 + X_2 + \text{---+ } X_{i-1} + X_i = T_{i-1} + X_i$; i = 1, 2,……..., N.

2) Then choose a random number R between 1 and $T_N = X_1 + X_2 +\text{……..+ } X_N = X$.

3) Choose the unit U_i if R lies between T_{i-1} and T_i, that is, if $T_{i-1} < R \leq T_i$. The probability $P(U_i)$ of selecting the i-th unit will thus be $P(U_i) = (T_i - T_{i-1}) /T_N = X_i / X = P_i$

4) Repeat the operation ‘n’ times for selecting a sample of size n with pps-wr.

7.5.5.6 Stratified Sampling

The Method: We might sometimes find it useful to classify the universe into a number of groups and treat each of these groups as a separate universe for purposes of sampling. Each of these groups is called a ***stratum*** and the process of grouping ***stratification***. Estimates obtained from each stratum can then be combined to arrive at estimates for the entire universe. This method is very useful as (i) it gives estimates not only for the whole universe but also for the sub-universes and (ii) it affords the choice of different sampling methods for different strata as appropriate. It is particularly useful when a survey organistion has regional field offices. This method is called ***Stratified Sampling.***

Let us divide the population (universe) of N units into k strata. Let N_s be the number of units in the s-th stratum. Y_{si} be the value of the i-th unit in the s-th stratum. Let the population mean of the s-th stratum be M_s. $M_s = (1/N_s) \sum_i Y_{si}$, $\sum_i$, i = 1,2,----,N_s (that is over the units within the s-th stratum) and the population M is $=(1/N) \sum_s N_s M_s = \sum_s W_s M_s$, where $W_s = (N_s / N)$ and $\sum_s$ being over the strata s = 1,2,……,k). Suppose that we select random samples from each stratum and the sampling method for different strata are different. Let the unbiased estimate of the population mean M_s of the s-th stratum be m_s. Denoting ‘st’ for stratified sampling, **an unbiased estimator of M is given by**

$$m_{st} = \sum_s W_s m_s = (1/N) \sum_s N_s m_s, \quad \sum_s , s = 1 \text{ to } k. \quad (2.66)$$

$$V(m_{st}) = \sum_s W_s^2 V(m_s) = (1/N^2) \sum_s N_s^2 V(m_s), \quad \sum_s , s= 1 \text{ to } k \quad (2.67)$$

$Cov.(m_s,m_r) = 0$ for $s \neq r$; (samples from diff. strata are independently chosen) .. (2.68)

$Y_{st}^* = \sum_s Y_s^*$; $\sum_s$, s = 1 to k. (2.69)

$V(Y_{st}^*) = \sum_s V(Y_s^*)$, $\sum_s$, s = 1 to k. (2.70)

Thus estimators with smaller variance (efficient estimators) can be obtained in stratified sampling if we form the strata in such a way as to minimise *intra-strata* or within-strata variation, that is, variance within strata. This would mean maximising between-strata or *inter-strata* variation, since the total variation is made up of within-strata and between-strata variation. In other words, *units in a stratum should be homogeneous.*

Stratified sampling enables us to choose the sample we wish to select by drawing independent samples from each of the different strata in to which we have grouped the universe. *How do we allocate the total sample size 'n' among the different strata?* One way is to *allocate the sample size to different strata in proportion to the size of individual strata measured by the number of units in these strata, namely,* N_s, [$\sum_s N_s = N$, (s= 1 to k)]. This method is especially appropriate in situations where no information is available except the sizes of the strata. The sample sizes for the samples from the stratum, say, the s-th stratum, would then be $n_s = n(N_s/N)$ and $\sum_s n_s$ can easily seen to be equal to 'n'. There are other methods like allocation of the sample size among strata in proportion to the stratum totals of the variable under study, that is, Y_s , the stratum total of the s-th stratum, 's' = 1 to k. We shall not go into the details of other methods here except one situation, namely, ***when we have a fixed budget F sanctioned for the survey.*** Let the cost function F be of the form $F_0 + \sum_s n_s F_s$, ($\sum_s$, s = 1 to k), where F_0 , n_s and F_s are respectively the overhead cost, the sample size in stratum 's' and the per unit cost of surveying a unit in stratum 's' (s = 1,2,-----, k). We can determine the optimum stratum-wise-sample-size by minimising the sampling variance of the sample mean (***2.67***) subject to the constraint that the cost of the survey is fixed. The stratum-wise optimum sample size is given by

$$n_s = [(F - F_0)]\ [W_s\sqrt{(V_s / F_s)}] / [\ \sum_s W_s\sqrt{(V_s F_s)}]\ \ , s = 1, k. \quad (2.71)$$

The stratum sample size should, therefore, be proportional to $W_s\sqrt{(V_s / F_s)}$. The minimum variance with the n_s so determined is,

$$\text{Min. } V(m_{st}) = [\ \ W_s\sqrt{(V_s F_s)}]^2 / (F - F_0)] \quad (2.72)$$

Check Your Progress 4

1) When is PPS method adopted?

……

……

……

2) When will the sampling variance of Y_{pps} * be small?

……

……

……

3) Say True (T) and False (F):

   a) PPS and stratified sampling can be combined with other sampling methods. (T/F)

   b) $V(m_{st})$ is reduced by ensuring that units within individual strataare heterogeneous. (T/F)

   c) The size of a stratified sample can be allocated among the strate, the size being the number of population units in a stratum. (T/F)

4) A systematic sample of size 18 has to be selected from a population of 124. What problems do you face in selecting the sample? Is the sample mean the unbiased estimator of the population mean M? How do you overcome these problems?

   ……

   ……

   ……

7.5.4.7 Cluster Sampling

The method: Supposing we are interested in studying certain characteristics of individuals in an area. We would naturally select a random sample of individuals from all the individuals residing in the area and collect the required information from the selected individuals. We might also think of selecting a sample of households out of all the households in the area and collect the required details from all the individuals in the selected households. *The households in the area are **clusters** of individuals and what we have done is to select **a sample of such clusters** and to collect the information needed from **all the individuals in the selected clusters** instead of selecting a random sample of individuals from all persons in the area.* What we have done is ***cluster sampling***. ***Cluster Sampling is a process of forming suitable clusters of units and surveying all the units in a sample of clusters selected according to an appropriate sampling method. The clusters of units are formed by grouping neighbouring units or units that can be conveniently surveyed together***. Sampling methods like srswr, srswor, systematic sampling, pps and stratified sampling discussed earlier can be applied to sampling of clusters by treating clusters themselves as sampling units. The clusters can all be of equal size or varying sizes, that is, the number of units can be the same, or vary from cluster to cluster. Clusters can be mutually exclusive, that is, a unit belonging to one cluster will not belong to any other cluster. They could also be overlapping.

Estimates from cluster sampling: Let us consider a population of NK units divided into N *mutually exclusive* clusters of K units each – a case of clusters of equal size. The population mean M and the cluster means are given respectively by $M = (1/N) \sum_s m_s$, $\sum_s$ being over clusters s = 1 to N and $m_s = (1/K) \sum_i Y_{si}$, $\sum_i$ being from i = 1 to K within the s-th cluster. Let us draw a sample of one cluster by srs. ***The cluster mean $m_{c\text{-}srs}$ (the subscript c-srs denotes cluster sampling with srs) is an unbiased estimate of M.*** The sampling variance of the sample cluster mean is

$$\mathbf{V(m_{c\text{-}srs}) = (1/N) \sum_s (m_s - M)^2 = \sigma_b^2 = Variance\ \textit{between}\ clusters;\ \sum_s,\ s = 1\ to\ N;} \quad (2.73)$$

Let us compare $V(m_{c\text{-}srs})$ with the sampling variance of the sample mean *when K units are drawn from NK units by SRSWR method.* How does the "sampling efficiency" of cluster sampling compare with that of SRSWR?. ***The sampling efficiency of cluster sampling compared to that of SRSWR, $E_{c/srswr}$, is defined as the ratio of the reciprocals of the sampling variances of the unbiased estimators of the population mean obtained from the two sampling methods***. The sampling variances and sampling efficiency are

$$V(m_{srswr}) = (1/K)\ [(1/NK) \quad {}_s \ {}_i Y_{si}^2 - M^2] = \sigma^2/K \tag{2.74}$$

$$\sigma^2 = \sigma_w^2 + \sigma_b^2 = \textit{within}\text{-cluster variance} + \textit{between}\text{-cluster variance.} \tag{2.75}$$

$$E_{c/srswr} > 1 \ \text{ if } \ \sigma_w^2 > (K-1)\sigma_b^2 \tag{2.76}$$

Thus, cluster sampling is more efficient than SRSWR if the *within*-cluster variance is larger than (K –1) times the *between*-cluster variance. Is this likely? This is not likely as the *between*-cluster variance will usually be larger than the *within*-cluster variance due to *within*-cluster homogeneity. ***Cluster sampling is in general less efficient than sampling of individual units from the point of view of sampling variance***. Sampling of individuals could provide a better cross section of the population than a sample of clusters since units in a cluster tend to be similar.

7.5.4.8 Multi-Stage Sampling

We noted in the sub-section on cluster sampling that random sampling of units directly is more efficient than random sampling of clusters of units. But cluster sampling is operationally convenient. How to get over this dilemma? We may first select a random sample of clusters of units and thereafter select a random sample of individual units from the selected clusters. We are thus selecting a sample of units, but from selected clusters of units. What we are attempting is a ***two-stage sampling***. This can thus be a compromise between the efficiency of direct sampling of units and the relatively less efficient sampling of clusters of units. This type of sampling would be more efficient than cluster sampling but less efficient than direct sampling of individual units. In the sampling procedure now proposed, the clusters of units are ***the first stage units (fsu)*** or ***the primary stage units (psu).*** The individual units constitute ***the second stage units (ssu)*** or ***the ultimate stage units (usu)***.

This procedure of sampling can also be generalised to multi-stage sampling. Take for instance a rural household survey. The fsu's in such a survey may consist of districts, the ssu's may be the *tehsils or taluks* chosen from the districts selected in the first stage, the third stage units could be the villages selected from the *tehsils or taluks* selected in the second stage and the fourth and the ultimate stage units (usu's) may be the households selected from the villages selected in the third stage. Such multi-stage sampling procedures help in utilising such information related to the variable under study as may be available in choosing the sampling method appropriate at different stages of sampling. In a multi-stage sampling, estimates of parameters are built up stage by stage. For instance, in two-stage sampling, estimates of the *sample aggregates relating to the fsu's* are built up from the ssu's using the sampling method adopted for selecting the ssu's. These estimates are then used with the sample probabilities of selection of fsu's to build up estimates of the relevant population parameters.

7.6 THE CHOICE OF AN APPROPRIATE SAMPLING METHOD

We have considered a number of random sampling methods in the foregoing sub-sections. A natural question that arises now is – ***which method is to be adopted in a given situation?*** Let us consider this question, although the answer to it lies scattered across the foregoing sub-sections. The choice of a sampling design depends on considerations like *a priori* information available about the population, the precision of the estimates that a sampling design can give, operational convenience and cost considerations.

1) When we do not have any *a priori* information about the nature of the population variable under study, SRSWR and SRSWOR would be appropriate. Both are operationally simple. However, ***SRSWOR is to be preferred,*** since $V(m_{srswor}) < V(m_{srswr})$**.** This advantage holds *only when the sampling fraction is not small, or N and n are not large.*

2) Systematic sampling is operationally even simpler than SRSWR and SRSWOR, but it should not be used for sampling from populations where periodic or cyclic trends/variations exist, though this difficulty can be overcome if the period of the cycle is known. $V(m_{sys})$ can be reduced if the units chosen in the sample are *as heterogeneous as possible.* But this will call for a rearrangement of the population units before sampling.

3) When additional information is available about the variable 'y' under study, say, on a variable (size variable) 'x' related to 'y', the pps method should be preferred. The sampling variance of Y* (or m) gets reduced when the probability of selection of units $P_i = (X_i /N)$ are proportional to Y_i , that is, the size X_i is proportional to Y_i or the variables x and y are linearly related to each other and the regression line passes through the origin. In such cases pps is more efficient than SRSWR. Further, this method can be utilised along with other sampling methods and their relative efficiencies. pps *is operationally simple.* Pps-wor combines the efficiency of SRSWOR and the efficiency-enhancing capacity of pps. However, most of the procedures of selection available, estimators and their variance for pps-wor are complicated and are not commonly used in practice. This is particularly so in large-scale sample surveys with a small sampling fraction, as in such cases sampling without replacement does *not* result in much gain in efficiency. Hence unless the sample size is small, we should prefer pps-wr.

4) Stratified sampling comes in handy when we wish to get estimates at the level of sub-populations or regions or groups. This method also gives us the freedom to choose different sampling methods/designs in different strata as appropriate to the group (stratum) of the population and the opportunity to utilise available additional information relating to the stratum. The sampling variance of estimators can also be brought down by forming the strata in such a way as to ensure *homogeneity of units within individual strata*. In fact, the stratum sizes can be so chosen as to minimise the variance of estimators, when there is a ceiling on the cost of the survey. ***Stratified sampling with SRS, SRSWOR or pps-wr presents a set of efficient sampling designs***.

5) Sometimes, sampling of groups of individual units than direct sampling of units might be found to be operationally convenient. Supposing it is easier to get a complete frame of clusters of individual units than that of units or,

only such a frame, and not that of the units, is available. (e.g. households are clusters of individuals). In such circumstances, cluster sampling is adopted. ***This is in general less efficient than direct sampling of individual units, as clusters usually consist of homogeneous units. A compromise between operational convenience and efficiency could be made by adopting a two-stage sampling design***, by selecting a sample of individual units (second stage units) from sampled clusters (the first stage units). A multi-stage design would be useful in cases where clusters have to be selected at more than one stage of sampling.

6) Finally, we can use the technique of independent I-PSS in conjunction with the chosen sampling design to get at (i) an unbiased estimate of V(m) for any sampling design or estimator of V(m), however complicated, (ii) a confidence interval for 'M' based only on the I-PSS estimates (when the population distribution is symmetrical) and (iii) a tool for monitoring the quality of work of the field staff and agencies.

7) ***SRSWOR, stratified sampling with SRSWOR and, when available information permits, pps-wr and stratified sampling with pps-wr, turn out to be a set of the more efficient and operationally easy designs to choose from. I-PSS can also be used in these designs where possible and necessary***.

Check Your Progress 5

1) We wish to study the wage levels of factory labour. What type of sampling method would you adopt for the study and why if (a) just a list of factories is available with the Chief Inspector of Factories of different State Governments, (b) if the list in (a) above also gives the total number of employees in the individual factories at the end of last year and (c) the list also indicates both the kind of product manufactured in the factory along with the information specified in (b) above.

……

……

……

7.7 LET US SUM UP

There are broadly three methods of collecting data. The array of tools used for data collection by such methods has expanded over time with the advent of modern technology. Confining data collection efforts to a sample from the population of interest to the study inevitably leads to questions like the use of random and non-random samples. Judgment sampling, convenience sampling, purposive sampling, quota sampling and snowball sampling all belong to the latter group. The absence of a clear relationship between a non-random sample and the parent universe and the presence of the researcher's bias in the selection of the sample render such samples useless for drawing ***valid*** conclusions about the parent population. But these methods are inexpensive and quick ways of getting a preliminary idea of the universe for use in designing a detailed enquiry and in exploratory research. Random samples, on the other hand, are free from such drawbacks and have properties that help in arriving at valid conclusions about the parent population.

The simplest of the sampling methods – **SRSWR** - ensures equal chance of selection to every unit of the population and yields a sample in which one or more units may occur more than once. 'm_{srswr}' is an unbiased estimator of M. Its precision as an estimator of M increases as the sample size increases. **SRSWOR** yields a sample of *distinct* units. 'm_{srswor}' is also unbiased for 'M'. ***SRSWOR is a more efficient than SRSWR as*** $V(m_{srswor}) < V(m_{srswr})$. *But this advantage disappears when the sampling fraction is small (< 0.05).* Both provide an unbiased estimator of V(m). An operationally convenient procedure - ***interpenetrating sub-samples (I-PSS)*** – also provides an unbiased estimator of V(m) for any sampling design and estimator for V(m), however complicated.

Systematic sampling is a simple and operationally convenient method used in large-scale surveys that requires only a random start and the sampling interval k = [N/n] for drawing the sample. A slight variant of 'm' is unbiased for 'M'. **C*ircular systematic sampling*** takes care of problems that arise when *N/n is not an integer.*. An unbiased estimate of V(m) is not possible but this problem can be tackled easily. ***Systematic sampling is not recommended when there is a periodic or cyclic variation in the population.*** *This problem too can be overcome if the period of the cycle is known.*

An example of methods where the probability of selection varies from unit to unit is **pps**. The "size" could be the value of a variable related to the study variable. In pps, each y_i / p_i , where y_i is the value of the study variable associated with the selected unit and p_i the probability of selection of the unit, is an unbiased estimate (Y*) of the population total Y and [(1/N) Y*] an unbiased estimator of M. As V(Y*) is small if the probabilities P_i are roughly proportional to Y_i , ***pps sampling is more efficient than SRS if the size variable x is proportional to y, that is, x and y are linearly related and the regression line passes through the origin.*** pps sampling can be done with SRSWR, SRSWOR or systematic sampling. In pps-srswr, $[(1/n) \sum_i (y_i / p_i)]$, ($\sum_i$ i = 1 to n), is an unbiased estimator of Y. This being the mean of n independent unbiased estimates with the same variance V(Y*), v(Y*) can be derived using the I-PSS technique.

Stratified Sampling is used when (i) estimates are needed for subgroups of a universe or (ii) the subgroups could be treated as sub-universes. It gives us the freedom to choose the sampling method as appropriate to each stratum. Estimates of parameters are available for the sub-universes (strata) and these can then be combined over the strata to get estimates for the entire universe. ***SE of estimates based on stratified sampling can be small if we form the strata in such a way as to minimise intra-strata variance. Each stratum should thus consist of homogeneous units, as far as possible. Stratum-wise sample sizes can also so chosen as to minimise the variance of estimators***.

Another operationally convenient sampling method, **cluster sampling**, is to sample groups of units or clusters of units at random and collect data from *all* the units of the selected clusters. For example, the household is a cluster of individuals. SRSWR, SRSWOR, pps or systematic sampling can be used for sampling clusters. ***Cluster sampling is, in general, less efficient than direct sampling of units from the point of view of sampling variance. The question here is one of striking a balance between operational convenience and cost reduction on the one hand and efficiency of the sampling design on the other***.

We could improve the efficiency of cluster sampling by selecting a random sample of units from each of the selected clusters - introduce another stage of

sampling. This is **two-stage sampling**. ***This would be more efficient than cluster sampling but less efficient than direct sampling of units.*** **Multi-stage sampling** can also be done. Such designs are commonly used in large-scale surveys as these facilitate the utilisation of information available and the choice of appropriate sampling designs at different stages.

Thus while non-random sampling methods are useful in exploratory research and preliminary work on planning of enquiries, random sampling techniques lead to *valid* judgments regarding the universe. Among random sampling methods, ***SRSWOR, stratified sampling with SRSWOR and, when available information permits, pps-wr and stratified sampling with pps-wr, turn out to be a set of the more efficient and practically useful designs to choose from. I-PSS can also be used in these designs where possible and necessary***.

7.8 SOME USEFUL BOOKS & REFERENCES

Burgess, R.G.(ed) (1982) : Field Research: A Sourcebook and Field Manual. (Contemporary Social Research 4), George Allen and Unwin, London.

Des Raj & Chandok P(1998) : Sampling Theory, Narosa Publishing House, New Delhi.

Levin, Richard I. & Rubin, David S. (1991) : Statistics for Mangement, Fifth Edition, Prentice-Hall of India (Private) Limited, M-97 Connaught Circus, New Delhi – 110001.

Krishniah, P.R. & Rao, C.R. (Eds.) (1988) : Handbook of Statistics – Vol. 6 : Sampling, North Holland, Amsterdam.

Mukhopadhyay, Parimal (1998) : Theory & Methods of Survey Sampling, Prentice-Hall of India Pvt. Ltd., New Delhi.

Murthy, M.N. (1967) : Sampling Theory and Methods, Statistical Publishing Society, 204, Barrackpore Trunk Road, Kolkota-700108.

Rao, C.R., Mitra, S.K. and Matthai, A. (1966) : Formulae and Tables for Statistical Work, Statistical Publishing Society, Kolkota – 700108.

Sampath, S. (2005) : Sampling Theory & Methods, Second Edition, Narosa Publishing House, New Delhi, Chennai, Mumbai, Kolkata.

Singh, Kultar (2007) : Quantitative Social Research Methods, Sage Publications (Pvt.) Limited, New Delhi

Singleton Jr., Royce A. & Straits, Bruce C. (2005) : Approaches to Social Research, 4th Edition, Oxford University Press, New York – Oxford.

Viswanathan, P.K. (2007) : Business Statistics – An Applied Orientation, Darling Kindersely (India) Pvt. Ltd. – licensees of Pearson Education in South Asia.

7.9 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Three methods of data collection are: the census and survey method, the observation method, and the experimental method.

2) A function of population values is a parameter; a function of sample observations is a statistic.

$$M = \frac{1}{N}\sum Y_i,\ m = \frac{1}{N}\sum y_i,\ \sigma^2 = \frac{1}{N}\sum Y_i^2 - M^2,\ \text{and } s^2 = \frac{1}{N}\sum y_i^2 - m^2,$$

where M/m is the mean of the population/sample values are the expressions of population/sample means and population/sample variance.

3) The purpose of drawing/studying a sample is to arrive at inferences about the characteristics of the unknown population. Thus, the sample mean is an estimator of the population mean (of the variable under study). The actual value of the estimator when computed from a sample is the estimate.

4) A representative sample is one that possesses the characteristics of the population in the same proportion as in the population. A conscious or subjective selection of units introduces a bias that erodes the sample's capacity to be a true representative of the characteristics of the population. It is for this reason that a random sampling procedure is preferred. Although non-random sampling is drawn without any regard to the rigors of a random sampling procedure, it gives a rough idea of the unknown features of the population under focus. Implemented judiciously, it can serve the objectives of providing preliminary inputs for planning a detailed enquiry later.

Check Your Progress 2

1) Sample observations in a random sample are random variables. A statistic is a function of sample observations and is, therefore, also a random variable. This random variable assumes different values in repeated random samples. These values form a frequency distribution of the statistic. This frequency distribution will, in the limit, tend to a probability distribution that gives the probabilities of the statistic assuming different values. This probability distribution is called the sampling distribution of the statistic. Let a statistic 't' be an estimator of a population parameter, say, T. The estimator 't' is an unbiased estimator of the parameter 'T', if the mean of the sampling distribution of 't' is 'T'. Since the mean of the sampling distribution of 't' can also be denoted by E(t), we can express this condition as E(t) = T.

2) Let 't' be an estimator of 'T'. The sampling distribution of the statistic 't' can be derived *only* if the statistic is based on a random sample selected from the population with a parameter 'T'. The mean of the sampling distribution can then be derived from the sampling distribution of 't'. That is, E(t) can be derived. If E(t) is $\neq$ T and E(t) = KT or T+K, we can correct the estimator 't' as (t/K) or (t – K) to arrive at an unbiased estimator of T. An example is the unbiased estimator of σ^2.

3) See Sub-section 7.5.3 A

Check Your Progress 3

1) The answer for (i) would be $\sqrt{\frac{4.41}{16}} = \frac{2.1}{4} = 0.525$. The answer for (ii) is $\sqrt{\frac{4.41}{25}} = \frac{2.1}{5} = 0.42$. The percentage decrease in SE works out to [(0.525 – 0.42)x 100] / [0.525] = 20%. The percentage increase in the sample size is $= \frac{9}{16} * 100 = 56.25\%$. The extent of increase is larger than the extent of decrease in the standard error. Try to work out (iii) and the compare the results of (ii) & (iii).

2) Use the formula 2.39 for *variance* of the sample mean under SRSWOR. Note that, *for the same sample size,*

$$V(m_{srswor}) - \frac{N-n}{N-1} V(m_{srswr}) = 1 - \frac{n-1}{N-1} V(m_{srswr}).$$

For n =16 and using SRSWOR sampling, SE works out to $\frac{64}{79} * 0.525 = 0.425$. Compare it with the SE for n = 25 in problem 1. SE of the sample mean was reduced to more or less this level by *increasing the size of the SRSWR sample from 16 to 25*. Work out (ii) & (iii)

3) a) F. SE(m) decreases in inverse proportion to the square root of the sample size.

b) F. SRSWOR is more advantageous if the sampling fraction is > 0.5.

c) F. $V(Y^*) = N^2 V(m)$.

Check Your Progress 4

1) When information on an auxiliary variable ‘x’ related to the variable ‘y’ under study is available for each unit of the population.

2) When P_i are proportional toY_i . That is, when the variable ‘x’ and ‘y’ are linearly related to each other and the regression line passes through the origin.

3) (a) T; (b) F: they should be homogeneous; (c) T.

4) The integer ‘k’ nearest to 124/18 is 7. Choose a random start between 1 and 7 for selecting the sample. Since N = 124 is not a multiple of n = 18, two problems arise. (i) Samples with random start 1 to 5 are of size 18 and others of size 17. The statistic $\frac{k}{N}\sum y_i = \frac{7}{124}\sum y_i$ and *not* the sample mean is an unbiased estimate of M. Both the problems at (i) and (ii) are overcome by adopting the circular systematic sampling procedure.

Check Your Progress 5

1) a) State-wise lists are available. Stratified sampling with each State (or part of it depending on the number of factories) being a stratum is an obvious choice. We need to collect details relating to *workers* in the factories. Within each stratum a cluster of factories can be selected and wage data relating to *all the workers* in each selected factory may be collected. Alternatively, we could adopt a two-stage sampling – select a sample of factories (fsu’s) and a sample of workers (ssu’s) from the selected factories, to enhance the efficiency of sampling. SRSWOR would be appropriate for each stage of selection.

UNIT 8 MEASUREMENT AND SCALING TECHNIQUES

Structure

8.0 Objectives

8.1 Introduction

8.2 Concept of Measurement

8.3 Measurement Issues in Research

8.4 Scales of Measurement

8.5 Criteria for Good Measurement

- 8.5.1 Reliability
- 8.5.2 Validity
- 8.5.3 Practicality

8.6 Errors in Measurements

8.7 Scaling Techniques

- 8.7.1 Comparative Scaling Techniques
- 8.7.2 Non-Comparative Scaling Techniques

8.8 Let Us Sum Up

8.9 Key Words

8.10 Some Useful Books

8.11 Answers or Hints to Check Your Progress Exercises

8.0 OBJECTIVES

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

- state the concept of measurement and its need;
- explain the various scales of measurement;
- discuss the various scaling techniques; and
- identify the criteria for good measurement.

8.1 INTRODUCTION

Measurement is the foundation of scientific enquiry. As we know that research begins with a 'problem' or topic. Thinking about a problem results in identifying concepts that capture the phenomenon being examined. Concepts are mental ideas representing the phenomenon. The process of conceptualization involves defining the concepts abstractly in theoretical terms. Operationalisation of concepts for research purpose involves moving from the abstract to the empirical level. This underlines the need to measure the various attributes of the people, the characteristics of objects or phenomenon. Further issues like decent work, human wellbeing, happiness, quality of education etc. which have several qualitative and quantitative dimensions are emerging important issues of research in economics. Such issues are being probed

through development of indicators and composite indexes which necessarily involve measurement of such indicators. Hence, a student of economics is expected to be well versed in the measurement scales, criteria of good measurement, and important scaling techniques. This unit throws light on these issues. Let us begin to discuss the concept of measurement.

8.2 CONCEPT OF MEASUREMENT

Simply speaking the process of assigning numbers to various attributes of people, objects or concepts is known as measurement. Technically, measurement is a process of mapping aspects of a domain to other aspects of a range according to some rule of correspondence.**Tyler(1963)** defines measurement as, "assignment of numerals according to rules". **Nunally (1970)**viewed that, "Measurement consists of rules, for assigning numbers to objects in such a way to represent quantities of attributes". Thus, Nunally focuses on both rules and manner in which numbers are assigned to an object.

According to **Campbell**, "Measurement is defined as the assignment of numerals to objects or events according to rules". The fact that numerals can be assigned under different rules leads to different kinds of scales and different kinds of measurement.

Thus, measurement is a process to assign numbers or other symbols to characteristics of objects according to certain rules. Assigning numbers permit statistical analysis of the resulting data and facilitate the communication of measurement rules and results. The rules for assigning numbers hence need to be standardized and uniformly applicable. In the assignment process, there must be one to one correspondence between the numbers and characteristics being measured. In this process, variables can be divided into two basic types – quantitative andqualitative variables.

8.3 MEASUREMENT ISSUES IN RESEARCH

A number of concerns arise in the process of measurement which need to be addressed by the researcher. Some of the important issues include the following:

1) Whether the underlying characteristics of the concept allows ordering (ordinal level) or categorizing (nominal level)?
2) Whether the features of the concept are discrete or continuous with fine gradation?

   The answer of these two questions will enable to determine the level of measurement inherent to the concept. As a general rule, efforts should be made that measurement represent the highest scale of measurement for a concept because it allows the use of more powerful statistical techniques for analysis.
3) Another important issue pertains to number of indicators for measurement of a concept. Some simple concepts can be measured by one indicator whereas abstract concepts are measured with more than one indicators. How many indicators are appropriate to measure a concept is to be decided by the researcher.
4) Another measurement issue concerns the source of valid and reliable measure. Hence, considerable attention should be given to identify valid and reliable measures for the study.

5) Measurement should be free from measurement errors like invalidity error and unreliability.

6) Measurement validity is distinct from internal validity and external validity because these two are separate research design issues.

7) A final measurement issue concerns the proper use of available data. For that, a research should be well aware of all available data sources and the types of data available with the various data compilation agencies.

8.4 SCALES OF MEASUREMENT

Before discussing different levels of measurement, let us remember that there are three postulates of measurement: (i) Equalities or identities, (ii) Rank order, (iii) Additivity.

You were introduced the elementary concept and types of scales of measurement in Unit 1. Even at the cost of repetition let us recall that Stanley S.Stevens (1946) in his seminal paper, "On the theory of scales of measurement" published in "Science" classified types of scales into four categories: (1) Nominal, (2) Ordinal, (3) Interval, (4) Ratio.

1) **Nominal Scale**

A qualitative scale without order is called nominal scale. In nominal scale numbers are used to name identity or classify persons, objects, groups, gender, industry type. Even if we assign unique numbers to each value, the numbers do not really mean anything. This scale neither has any specific order nor it has any value. In case of nominal measurement, statistical analysis is attempted in terms of counting or frequency, percentage, proportion, mode or coefficient of contingency. Addition, subtraction, multiplication and division are not possible under this scale.

2) **Ordinal scale**

A qualitative scale with order is called ordinal scale. In ordinal scale numbers denote the rank order of the objects or the individuals. Numbers are arranged from highest to lowest or lowest to highest. For example, students may be ranked I^{st}, 2^{nd} and 3^{rd} in terms of their academic achievement. The statistical operations which can be applied in ordinal measurement are median, percentiles and rank correlation coefficient. Ordinal scales do not provide information about the relative strength of ranking. This scale does not convey that the distance between the different rank values is equal. Ordinal scales are not equal interval measurement. Further this scale does not incorporate absolute zero point.

3) **Interval scale**

Interval scale includes all the characteristics of the nominal and ordinal scale of measurement. In interval scale numerically equal distance on the scale indicate equal distances in the attributes of the object being measured. For example, a scale represent marks of students using the attributes range 0 to 10, 10 to 20, 20 to 30, 30 to 40 and 40 to 50, and so forth. The mid point of each range (ie. 5, 15, 25, 35 and 45 etc.) are equidistance from each other. The data obtained from an interval scale is known as interval data. The appropriate measures used in this scale are: arithmetic mean, standard deviation, Karl Pearson's coefficient of correlation and tests like t-test and f-test. We cannot apply coefficient of variation in the interval scale.

4) **Ratio Scale**

Ratio scale has all the characteristics of nominal, ordinal and interval scale. It also possesses a conceptually meaningful zero point in which there is a total absence of the characteristic being measured. Ratio scales are common among physical sciences rather than among social sciences. The examples of ratio scales are the measures of height, weight, distance and so on. All measures of central tendencies that can be used in this scale include geometric and harmonic means.

The properties of these four scales can be summarized in following tabular form:

Properties / Scale	Category	Ranking	Equal interval	Zero Point
Nominal	✓			
Ordinal	✓	✓		
Interval	✓	✓	✓	
Ratio	✓	✓	✓	✓

Properties / Scale	Indicates Difference	Indicates Direction of Difference	Indicates Amount of Difference	Absolute Zero
Nominal	×			
Ordinal	×	×		
Interval	×	×	×	
Ratio	×	×	×	×

You will notice in the above tables that only the ratio scale meets the criteria for all four properties of scales of measurement. Interval and Ratio scale data are sometimes referred to as parametric and Nominal and Ordinal data are referred to as non-parametric.

Examples of Scales of Measurement –

Scale	Example	Statistics
Nominal	Gender Yes-no Students roll number Objects/Groups	Frequency, percentage, proportion Mode, Coefficient of contingency, chi-square (x^2)
Ordinal	Class Rank, Socio Economic status Academic Achievement	Median, percentile, rank, correlation, Range
Interval	Student grade point, Temperature, Calendar dates, Rating Scale	Mean, correlation, Range, Standard deviation, rank order variance, Karl pearson's correlation, t-test, f-test etc.
Ratio	Weight, Height, Salary, Frequency of buying a product	Mean, Median, Mode, Range, Variance, Standard deviation, coefficient of variation, rank order variance, Karl pearson's correlation, t-test, f-test

8.5 CRITERION FOR GOOD MEASUREMENT

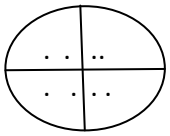
There are three major criteria for evaluating measurement:

1) Reliability
2) Validity
3) Practicality

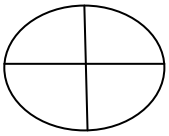
8.5.1 Reliability

Reliability refers to the degree to which the measurement or scale is consistent or dependable. If we use same construct again and again for measurement, it would lead to same conclusion. Reliability is consistency in drawing conclusion.

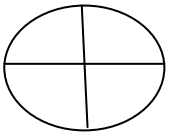
Reliability and Validity



Reliable and Valid



Valid but not Reliable



Reliable but not valid

8.5.2 Validity

Validity refers to the extent an instrument or scale tests or measures what it intends to measure. This means validity is the extent to which differences found with a measuring instrument reflect true differences among those being tested. This can be done by seeking other relevant evidences. There are three types of validity: content validity; criterion validity and construct validity. Content validity indicates the extent to which it provides coverage of the issues under study. Criterion validity examines how well a given measure relates to one or more external criterion based on empirical observations. And, construct validity explains the variation observed in tests conducted on several individuals. Construct validity is closely related with factor analysis.

8.5.3 Practicality

From practical view point, measure should be economical, convenient and interpretable. The measure should not be lengthy and difficult. The measure can be done by highly specialized persons.

8.6 ERRORS IN MEASUREMENTS

Measurement need to be precise and unambiguous for validity and reliability of any research study. Hence, measurement should be either free from errors or have minimum error. Any measurement usually involves two types of errors:

- Measurement invalidity
- Unreliability

Measurement invalidity refers to the degree to which the measure incorrectly captures the concept.

Unreliability refers to inconsistency in what the measure produces under repeated uses. If any measure, on average give some score for a case on variable and gives other score, when it is used again, it is said to be unreliable.

The possible sources of error are:

a) **Respondent** by not expressing strong negative feelings.

b) **Situation** wherein interviewer and respondent are not in good rapport.

c) **Measurer** – errors may also creep because of incorrect coding, faulty tabulations and statistical calculations or behavior or attitude of the interviewer.

d) **Instrument** errors may also arise because of the defective measuring instrument.

Check Your Progress 1

1) What do you mean by measurement?

……

……

……

……

……

2) Give examples of interval scale and ratio scale.

……

……

……

……

……

3) How will you examine the validity of a measurement?

……

……

……

……

……

4) Identify three measure concerns that need to be addressed in the process of measurement.

……

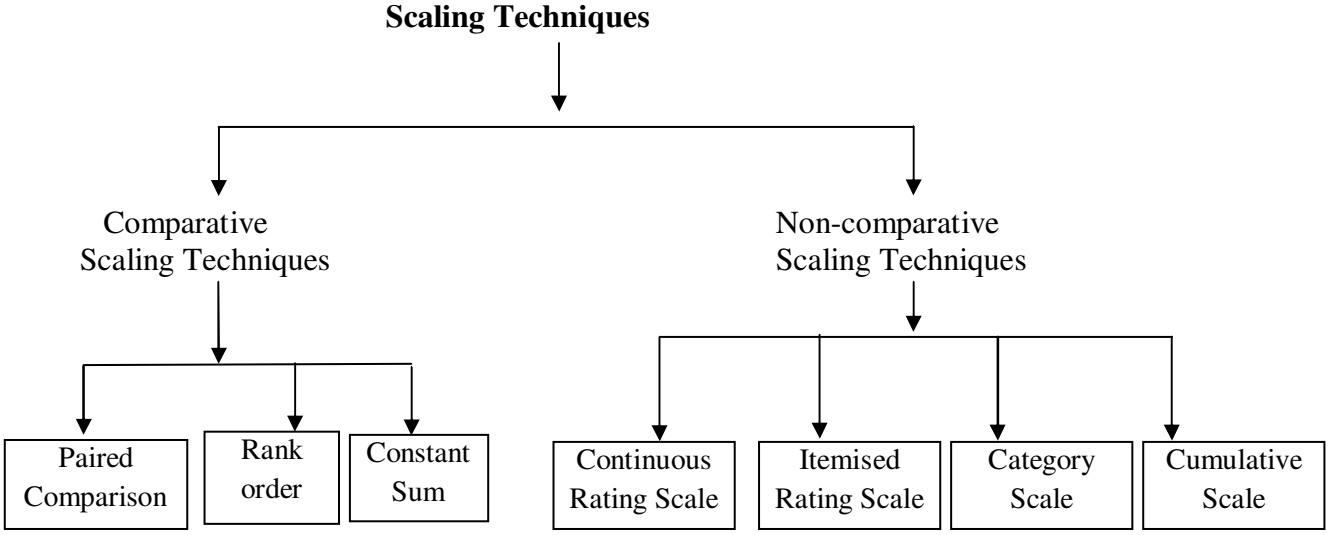
……

……

……

8.7 SCALING TECHNIQUES

Scaling describes the way of generation of a continuum upon which measured objects are located. Broadly, scaling techniques can be classified into two categories: comparative, and non comparative scaling.



8.7.1 Comparative Scaling Techniques

Comparative scales involve the comparison of objects directly with one another. For example, in a study of consumer preferences for different telecom services, a respondent may be asked to rank choice according to his preferences. In this technique, data must be interpreted in relative terms and have only ordinal or rank order properties. In comparative scales small differences between stimulus objects can be detected.

Example

Rank the following factors in order of your importance in choosing mobile services (Assign 1 to most important and 5 to least important factor. Please do not repeat the ranks.)

Factor: Price, connectivity, call drop, Internet, Download speed.

Rank :…… …… …… ……. …….

There are various comparative scaling techniques. Three most commonly used comparative scaling techniques are discussed here.

a) Paired comparison

b) Rank order

c) Constant sum

a) **Paired Comparison**

The paired comparison is most commonly used scaling technique. This method is simply a binary choice. In this method respondents choose the stimulus or items in each pair that has the greater magnitude on the choice dimension they are instructed to use. This method is used when the study requires to distinguish between the two objects. The data obtained in this method is ordinal in nature.

Example

In a study of consumer preferences for the two brands of milk product i.e. Amul and Mother Dairy, a consumer is asked to indicate which one of the two brands he would prefer for personal use:

1) Which milk product of the following you prefer on the basis of taste. Please tick mark (√).

Amul	Mother Dairy

2) Which milk product will you prefer on the basis of packaging? Please tick mark (√).

Amul	Mother Dairy

3) Which product will you prefer on the basis of price? Please tick mark (√).

Amul	Mother Dairy

This technique is useful when the researcher wants to compare two or more than two objects. If there are more than two objects (for example n objects) to compare, the total comparison will be;

$$\text{Number of comparison} = \frac{n[n-1]}{2};$$

n= number of objects.

b) **Rank Order**

This method is very popular among researchers and provides ordinal data. In this method, respondents are provided various objects and asked to rank the objects in the list. Rank order method is less time consuming. In this method, if there are n objects, only (n-1) decisions need to be made. Respondent can easily understand the instructions for ranking. This technique produces ordinal data.

Example

Rank the various brands of mobile phone in order of preference. The most preferred can be ranked 1, the next as 2 and so on. The least preferred will have the last rank. No two brands should receive the same rank number.

Sl. No.	Brand	Rank
1	Nokia	
2	Motorola	
3	Samsung	
4	HTC	
5	Karbonn	
6	Spice	
7	Xolo	
8	Gionee	
9	LG	
10	Sony	

c) **Constant Sum**

In constant sum scaling, respondents are asked to allocate a constant sum of units to a specific set of objects on the basis of pre-defined criterion. If an object is not important, the respondent can allocate zeropoint and if an object is most important may allocate maximum points out of the fixed points. The total fixed points are 100. The total may be taken as some other value depending on the study.

Example

Allocate preferences regarding a movie based on various predefined criteria. Respondents were asked to rate each criteria in such a way that the total becomes 100.

Criteria	Respondent Preference
Content Quality	20
Music	20
Sound	15
Story	30
Breadth of Coverage (local, regional Supplements, national & global)	15
Total	**100**

The data obtained in this method is considered an ordinal scale. The advantage of this method is that it provides fine discrimination among objects without consuming too much time. However, allocation over large number of objects may create confusion for the respondent. This method cannot be used as a response strategy with illiterate people and children.

8.7.2 Non Comparative Scaling Techniques

Non-comparative scaling techniques involve scaling of objects independent to some specific standard. Respondent evaluate only one object at a time. For example, in a study of consumer preferences for different telecom services, a consumer may be asked to rate a list of factors that he/she would consider while choosing a particular telecom company.

Example

Please rate the following factors you think important in choosing mobile service. Rate 1 to least important and 5 to most important (rating could be on any scale. In this example, we have used a 5 point scale).

Factor	Price	Connectivity	Call Drop	Download Speed	Internet
Rating					

In this scaling technique, data is usually in interval scale. It can be continuous, metric/numeric also. Some of the commonly used non-comparative scaling techniques are –

i) Continuous Rating Scale

ii) Itemised Rating Scale

iii) Category Scale

iv) Cumulative Scale

i) **Continuous Rating Scale**

This rating scale is also known as graphic rating scale.In continuous rating scale, respondents indicate their rating by marking at appropriate position on a line. The line is labeled at both ends from one extreme criterion to the other. The line may contain points 0,10,20,…..,100.

Example

How would you rate a magazine with regard to its quality.

Quality Indicators	Scale Measurement
Content Coverage	Most_____x_____Least Above 80 60 40 20 0
Language	Most_____x_____Least
Presentation Style	Most_____x_____Least

A very large number of ratings are possible if the respondents are literate to understand and accurately differentiate the objects. The data generated from continuous rating scale can be treated as numeric and interval data. The researcher can divide the line into as many categories as desired and assign scores based on the categories under which the ratings fall.

ii) **Itemized Rating Scale**

In this scale, respondents are provided with a scale having numbers/ descriptions associated with each category. The respondents are asked to select the best fitting category with the object.The commonly used itemised rating scale are:

a) Likert Type Scale (Summated Scale)

b) Semantic Differential Scale

c) Stapel Scale

a) **Likert Type Scale (Summated Scale)**

Likert (1932) proposed a simple and straight forward method for scaling attitudes that is most frequently used today. This scale is also known as summated rating scale. Summated scales consist of a number of statements which express either a favourable or unfavourable attitude towards the given object to which the respondant is asked to respond to each of the statements in terms of several degrees of agreement or disagreement. Let us understand this method with an example.

Against the following statements relating to psychological well-being, please tick mark any one of the following options shown against the statement

Sl. No.	Statements	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
		5	4	3	2	1
1	I always live in present and do not worry about future.					
2	Many people think without purpose in life but I am not such type of persons.					
3	I generally feel sense of perfection during the moments I live in.					
4	I am very well in managing and discharging most of my responsibilities.					
5	I give importance to the innovative ideas and experiences that challenge my thought about me and the world.					

Depending on the wording of an individual item, an extreme answer of strongly agree or strongly disagree will indicate the most favorable response on the underlying attitude measured by the questionnaire. This scale is relatively easy and quick to compute. The data in this scale is of interval scale.

Some of the important advantages of this scale are: (i) it is relatively easy to construct as it can be performed without a panel of judges, (ii) it is considered more reliable because respondents answer each statement included in the

instrument, (iii) it can easily be used in respondent centred and stimulus centred studies, (iv) this takes less time to construct. However, there are several limitations of this type of scale i.e. (i) with this scale, we can simply examine whether respondents are more or less favourable to a topic but it is difficult to tell how much more or less they are, (ii) This does not rise more than to a ordinal scale structure, (iii) the total score of an individual respondent has little clear meaning since a given total score can be secured by a variety of answer patterns.

b) **Semantic Differential Scale**

This scale was developed by Osgood, suit &Tannenbaum (1957). Semantic differential scale includes a seven point scale in comparison to Likert's five point scale. The semantic differential scale is based on the proposition that an object can have several implied or suggestive meanings to an expressed opinion. The semantic differential scale can be scored on either -3 to +3 or 1 to 7. This scale can also provide interesting comparison between products organizations and so on. The results of this scale are further analysed by the factor analysis.

Example: A researcher has developed an item seeking your perceptions regarding magazine. Please mark√on each line that best indicates your perception. Please be sure to mark every scale and do not omit any scale.

India Today Magazine is

	+3 +2 +1 0 -1 -2 -3	
Easy to read	____ : ____ : ____:____:_√__:____:____	Hard to read
Modern	____ : ____ : ____:____:_√__:____:____	Old fashion
Rational	____ : ____ : ____:____:____:____:____	Emotional
Unreliable	____ : ____ : ____:____:____:____:____	Reliable
Worthless	____ : ____ : ____:____:____:____:____	Valuable
Unorganized	____ : ____ : ____:____:____:____:____	Organized
Orthodox	____ : ____ : ____:____:____:____:____	Liberal
Vain	____ : ____ : ____:____:____:____:____	Modest

Semantic differential scale provides a very convenient and quick way of gathering impressions on one or more than one concept. The data generated from this scale can be considered as numeric in some cases, and can be summed to arrive total scores adjectives. It must define a single dimension and each pair must be bipolar opposites labeling the extremes.

Like Likert Type Scale, this scale has also several advantages: (i) it is an efficient and easy way to secure attitude from a large sample, (ii) the total set of responses (both direction) gives a comprehensive picture of the meaning of an object as well as a measure of the subject of the rating, (iii) it is a standarised technique that can be easily repeated.

However, this technique escapes many of the problems of response distortion found with more direct methods.

c) **Stapel Scale**

Stapel scales are named after John stapel who developed these scales. Stapel scale consists of a single criterion in the centre with 10 categories numbered from -5 to +5 without a neutral point (zero). The scale is usually presented vertically. Negative rating indicates that the respondent inaccurately describes the object and positive rating indicates that the respondent describes the object accurately.

Example

Rate the Departmental store by marking (√) on the following factors. +5 indicate that the factor is most accurate for you and -5 indicate that the factor is most inaccurate for you.

+5		-5	
+4		-4	
+3		-3	
+2		-2	
+1		-1	
Services		**Products**	
-1		+1	
-2		+2	
-3		+3	
-4		+4	
-5		+5	

In this scale, data generated is interval data. In this method, data can be collected through telephonic interview. The data obtained from Stapel scale can be analysed in the same way as semantic differential scale.

iii) **Category Scale**

In Category scaling method, objects are grouped into a predetermined number of categories on the basis of their perceived strength along certain dimension. This scale is a dichotomous scale. This method is useful for socio demographic questions. The response is this category we get, typically 'Yes' or 'No' type. Data in this category is either nominal or ordinal. Under this category scale, instructions and response tasks are quick and simple, and many options can be included.

Example

A) **Single Category Scale**

1) Do you own a house?

Yes	No

2) Do you own a car?

Yes	No

3) Do you own a Mobile phone?

Yes	No

b) **Multiple Category**

1) Please indicate your income by marking (√) in the income group you fall

5000 - 10000	
10000 - 15000	
15000 - 20000	
20000 - 25000	

iv) **Cumulative Scale (Guttman Scale)**

Like other scales, it consists of series of statements to which a respondent expresses his agreement or disagreement. The statements are in a form of cumulative series i.e. in a way, an individual who replies favourably to say item no. 3 also replies favourably to item no. 2 and 1 and so on. The individual's score is worked out by counting the number of points concerning the number of statements he answers favourably. Knowing the total score, we can estimate as to how a respondent has answered individual statements constituting cumulative scales.

Let us understand with the illustration of an **example**.

Here is an example of a Guttman scale - the Bogardus Social Distance Scale:

(Least extreme)

1) Are you willing to permit immigrants to live in your country?
2) Are you willing to permit immigrants to live in your community?
3) Are you willing to permit immigrants to live in your neighbourhood?
4) Are you willing to permit immigrants to live next door to you?
5) Would you permit your child to marry to an immigrant?

(Most extreme)

Cumulative scale (scalogram analysis) like any other scaling technique has several advantages as well as limitations. The advantages are – It assures that only a single dimension of attitude is measured, (ii) Researcher's subjective judgement is not allowed to creep in the development of scale since the scale is determined by the replies of respondents.

Its main limitation is that (i) In practice perfect cumulative or unidimesional scale are very rarely found and approximation is used, (ii) Its developmental procedure is cumbersome in comparison to other scaling methods.

Check Your Progress 2

1) What is the distinction between comparative measuring scale and non comparative measuring scale?

   …….

   …….

   …….

   …….

2) What do you mean byLikert Scale? Give its two examples.

   …….

   …….

   …….

   …….

3) What are the sources of measurement errors?

   …….

   …….

   …….

8.8 LET US SUM UP

Operationalisation of concepts for research purpose need measurement of various attributes of the people, the characteristics of objects or phenomenon. Measurement is a process to assign numbers to the characteristics according to certain rules. Measurement scales are of four types – nominal, ordinal, interval and ratio. A number of concerns arise in the process of measurement. Important among them are: the scale measurement i.e. the validity, reliability and practicality, number of indicators for measurement etc. Scaling techniques are broadly of two types – comparative, and non-comparative. Within comparative category, paired, rank order and constant sum are important techniques. Within non-comparative techniques, four sub-categories i.e. continuous rating scale, itemized rating scale, category scale and cumulative scale have been covered in this unit. All these scaling techniques have their relative advantages and limitations and are used depending upon the goal and the nature of characteristics of objects/phenomenon to be measured.

8.9 KEY WORDS

Measurement : The assignment of numerals to objects or events according to rules.

Nominal : Numbers are used to name identity or classify persons, objects, groups or gender.

Ordinal Scale	**:**	Scale with order is called ordinal scale.
Interval Scale	**:**	In this scale numerically equal distance on the scale indicate equal distances in the attributes of the object being measured.
Ratio Scale	**:**	A conceptually meaningful zero point in which there is a total absence of the characteristics being measured.
Validity	**:**	The extent to which an instrument or scale tests or measures what the measurement or scale i.e. consistent or dependent.
Reliability	**:**	The degree to which the measurement or scale i.e. consistent or dependent.
Practicality	**:**	Measure should be economical convenient and interpretable.
Errors in Measurement	**:**	Discrepancies between the obtained scores and the corresponding scores.
Scaling	**:**	The way of generation of a continuum upon which measured objects are located.
Comparative Scaling		The comparison of objects directly with one another.
Paired Comparison	**:**	A binary choice, respondents choose the objects in each pair that has the greater magnitude on the choice dimension.
Rank Order	**:**	Respondents are provided with various objects and asked to rank the list of objects.
Constant Sum	**:**	Respondents are asked to allocate a constant sum of units to a specific set of objects with respect to predefined criterion.
Non Comparative Scaling	**:**	Scaling of object independently to some specify standard.
Continuous Rating Scale		Respondents indicate their rating by marking at appropriate position on a line.
Itemised Rating Scale		Respondents are asked to select the best fitting category with the object.
Likert Scale	**:**	Respondents are asked to select the best fitting category with the object.
Semantic Differential Scale		Seven point scale based on the proposition that an object can have several implied or suggestive meanings to an expressed opinion.
Stapel Scale	**:**	A single criterion in the centre with 10 categories numbered from -5 to +5 without a neutral point.
Category Scale	**:**	Objects are grouped in to a predetermined number of categories on the basis of their perceived strength along certain dimension.

8.10 SOME USEFUL BOOKS/PAPERS

Stevens S.S. (1946): *On Theory of Scales of Measurement, Science New Series,* Vol. 103, No. 2684, pp 77-680.

Kothari C.R. (2008): *Research Methodology Methods and Techniques,* New Age International Publishers. Chapter 5 page 69 to 82.

Babbie Earl, (2010): *The Practice of Social Research*, 12th Edition, WodsworthCengage Learning, CA, USA.

Malhotra, N.K. and S. Dash (2009): *Marketing Research,* An Applied Orientation, Pearson Education, New Delhi.

Singh, A.K.: *Tests Measurements and Research Method in Behavioural Sciences,* Bharti Bhawan Publishers and Distributors, Patna.

Gregory R.J. (2006): *Psychological Testing History, Principles and Applications,* Pearson Education, New Delhi, India.

Michael S. Lewis-Beck (ed) (2004): *The Sage Encyclopedia of Social Science Research Methods,* Vol. 2, Page no. 161 to 165, Sage Publications, New Delhi, India.

8.11 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 8.2
2) See Section 8.4
3) See Section 8.5
4) See Section 8.3

Check Your Progress 2

1) See Sub-section 8.7.1 and 8.7.2
2) See Sub-section 8.7.2 under the subhead Itemised Rating Scale
3) See Section 8.6.

Indira Gandhi
National Open University
School of Social Sciences

MEC-109
Research Methods in Economics

Block

3

QUANTITATIVE METHODS-I

UNIT 9
Two Variable Regression Models **5**

UNIT 10
Multivariable Regression Models **37**

UNIT 11
Measures of Inequality **59**

UNIT 12
Construction of Composite Index in Social Sciences **93**

Expert Committee

Prof. D.N. Reddy
Rtd. Professor of Economics
University of Hyderabad, Hyderabad

Prof. Romar Korea
Professor of Economics
University of Mumbai, Mumbai

Prof. Harishankar Asthana
Professor of Psychology
Banaras Hindu University
Varanasi

Dr. Manish Gupta
Sr. Economist
National Institute of Public Finance and Policy
New Delhi

Prof. Chandan Mukherjee
Professor and Dean
School of Development Studies
Ambedkar University, New Delhi

Prof. Anjila Gupta
Professor of Economics
IGNOU, New Delhi

Prof. V.R. Panchmukhi
Rtd. Professor of Economics
Bombay University and Former
Chairman ICSSR, New Delhi

Prof. Narayan Prasad **(Convenor)**
Professor of Economics
IGNOU
New Delhi

Prof. Achal Kumar Gaur
Professor of Economics
Faculty of Social Sciences
Banaras Hindu University, Varanasi

Prof. K. Barik
Professor of Economics
IGNOU
New Delhi

Prof. P.K. Chaubey
Professor, Indian Institute of
Public Administration, New Delhi

Dr. B.S. Prakash
Associate Professor in Economics
IGNOU, New Delhi

Shri S.S. Suryanarayana
Former Joint Advisor
Planning Commission, New Delhi

Shri Saugato Sen
Associate Professor in Economics
IGNOU, New Delhi

Course Coordinator and Editor: Prof. Narayan Prasad

Block Preparation Team

Unit	Resource Person	IGNOU Faculty (Format, Language and Content Editing)
9	Dr. Anoop Chatterjee Associate Professor in Economics ARSD College (University of Delhi), Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi
10	Shri B.S. Bagla Associate Professor in Economics PGDAV College (University of Delhi), Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi
11	Prof P.K. Chaubey IIPA, New Delhi	Prof. Narayan Prasad Professor of Economics, IGNOU, New Delhi
12	Dr. Sunil K Mishra Fellow, Institute for Human Development New Delhi	Prof. Narayan Prasad Professor of Economics, IGNOU, New Delhi

Print Production

Mr. Manjit Singh
Section Officer (Pub.)
SOSS, IGNOU, New Delhi

October, 2015

© Indira Gandhi National Open University, 2015

ISBN-978-81-266-

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi
Printed at :

BLOCK 3 QUANTITATIVE METHODS-I

Economic theory is basically concerned about economic laws. These laws (or hypotheses) are qualitative statements on the economic behaviour of various agents both at the micro and macro levels. The validity of such qualitative statements, however, depends upon their rigorous empirical verification by means of the available data.

The implicit argument between the economic theory and its empirical verification is that if the stated laws holds good in the real world situation, it should be reflected in the displayed pattern of the relevant data. The identification and the examination of such patterns are conducted by employing various statistical techniques. The primary statistical tool employed for such a purpose is the *Regression Modeling*. Hence, a thorough knowledge of Regression Models is absolutely essential for conducting any serious research in Applied Economics.

An economic law or theory essentially postulates some relationship that exist among the economic variables. Suppose the relationship between two variables Q and P (i.e. Quantity and Price) is expressed in the form (say linear form) $Q = \alpha + \beta P$. Such an expression amount to an exact linear relationship which is hard to find in reality. Usually, various non-deterministic factors (called random factors) affect such a relationship. Therefore, in order to allow for the effect of such random factors, a random component is incorporated into the model i.e. $Q = \alpha + \beta P + U$, where U is a random variable. With this, the strict mathematical relationship expressed before becomes statistical in character. Regression technique can now be applied to estimate the two parameters α and β by using the actual data. This is the crux of a *Two Variable Regression Model* explained in **Unit 9**.

Unit 10 explains the multiple regression models by including more than two independent random variables. The same procedure as done in the case of two variable regression models is applied to examine whether the estimated parameters correspond to the postulated relationship. The treatment accorded to the Two Variable Regression Models and the Multiple Regression Models in this course is first limited to the ordinary least square (OLS) method subsequently, another method viz., the maximum likelihood method (MLM) is also explained. The two methods taken together equip a researcher with the basic tools for undertaking empirical research.

Improvement in well being of the poor has been one of the important goals of economic policy and to a significant extent it is determined by the growth and distribution of its income. Distribution patterns have an important bearing on the relationship between average income and poverty levels. Extreme inequalities are economically wasteful. Further, income inequalities also interact with other life-chance inequalities. Hence reducing inequalities has become priority of public policy. **Unit 11** deals with the conceptual and measurement aspects of income inequality.

The complex social and economic issues like child deprivation, food security, human development, human wellbeing etc. are difficult to measure in terms of single variable. They are expression of several indicators. Composite Index is an important statistical techniques to express the single value of several interdependent or independent variables. Hence, **Unit 12** throws light on various methods to construct Composite Index.

UNIT 9 TWO VARIABLE REGRESSION MODELS

Structure

9.0 Objectives

9.1 Introduction

9.2 The Issue of Linearity

9.3 The Non-deterministic Nature of Regression Model

9.4 Population Regression Function

9.5 Sample Regression Function

9.6 Estimation of Sample Regression Function

9.7 Goodness of Fit

9.8 Functional Forms of Regression Model

9.9 Classical Normal Regression Model

9.10 Hypothesis Testing

9.11 Let Us Sum Up

9.12 Exercises

9.13 Key Words

9.14 Some Useful Books

9.15 Answers or Hints to Check Your Progress Exercises

9.16 Answers or Hints to Exercises

9.0 OBJECTIVES

After reading this unit, you will be able to:

- know the issue of linearity in regression model;
- appreciate the probabilistic nature of the regression model;
- distinguish between the population regression model and the sample regression model;
- state the assumptions of the classical regression model;
- estimate the unknown population regression parameters with the help of the sample information;
- explain the concept of goodness of fit;
- use various functional forms in the estimation of the regression model;
- state the role of classical normal regression model; and
- conduct some tests of hypotheses regarding the unknown population regression parameters.

9.1 INTRODUCTION

The empirical research in economics is concerned with the statistical analysis of economic relations. Often these relations are expressed in the form of regression equations involving the dependent variable and the independent variables. The formulation of an economic relation in the form of a regression equation is called a regression model in econometrics. In such kind of a regression model, the variable under study is assumed to be a function of certain explanatory variables. The major purpose of a regression model is the estimation of its parameters. In addition, it is also useful for testing hypotheses and making certain forecasts. However, our concern will be mainly with the estimation of parameters and testing of some hypotheses. At this stage, we shall refrain from discussing the issue of forecasts from a regression model. A regression model that contains one explanatory variable is called a two variable regression model. In this unit, we shall focus on a two variable regression model. It may be mentioned here that we are essentially focusing on what is known as the two variable classical regression model.

9.2 THE ISSUE OF LINEARITY

When we talk about a regression model; at our level, what we have in our mind is a linear regression model. It is important to have a clear idea about the concept of linearity in the context of a regression model.

Suppose, we have a regression equation like,

$$Y = \alpha + \beta X$$

Generally, we call this a linear regression equation. By linearity we often mean a relationship, in which, the dependent variable is a linear function of the independent variable. In this case, the graphical representation of the regression equation is a straight line. However, there can be another interpretation of linearity. Consider the following regression equation,

$$Y = \alpha + \beta X + \gamma X^2$$

Now, this regression equation is linear in parameter, in the sense that the highest power of any parameter (constant) is one. But this equation is non-linear in variable because the highest power of the independent variable is two. In fact, conventionally speaking, it is a quadratic regression equation. In this way, we can have any number of polynomial regression equations with the highest power of the independent variable going up to any positive integer. And all these equations may be linear in parameter.

But consider this regression equation,

$$Y = \alpha + \beta^2 X$$

This regression equation is linear in variable, but non-linear in parameter because the power of β is two.

We should note now that from the point of view of regression analysis, we shall consider only those models which are linear in parameter, no matter whether they are linear in variable or not. After all, the main purpose of a regression model is the estimation of its parameters. These parameters have to be linear for the purpose of their straightforward estimation, say,

by least square method. For example, the parameters of the second regression equation presented above, can be easily estimated by a simple extension of the least square method that we have studied in Unit 8. It should be mentioned here that there are some regression equations that are non-linear in parameters. By applying suitable transformations they can be reduced to linear in parameter regression equations. We may consider the following regression equation

$$Y = \alpha X^{\beta}$$

This regression equation in its present form is non-linear in parameter. However, by applying log transformation, we obtain the following equation

$$\log Y = \log \alpha + \log \beta X$$

It can be clearly seen that the equation has now been transformed into one that is linear in the log of parameters. We can easily estimate log values of these parameters by using the usual least square method and then obtain the estimated values of the original parameters by applying the antilog procedure. However, some times, the regression equations can be of non-linear in parameter type that no transformation can render them to a linear in parameter form. Such equations should be the basis of non-linear or intrinsically non-regression models. There is no direct method available for the estimation of the parameters of this kind of regression models and they are generally estimated by following some standard iteration procedure. Any discussion on such iteration procedure, however, is beyond the scope of the present unit.

9.3 THE NON-DETERMINISTIC NATURE OF THE REGRESSION MODEL

We have already referred to the statistical nature of the regression equation in the last unit. By the very nature of social science, we cannot expect the relationships that may exist among different variables to be exact or deterministic. There is always some random element involved in them. As a result, for a particular value of the independent variable X, the value of the dependent variable Y cannot be exactly determined from such a relationship. Let us consider the example of a consumption function. Here, for a given level of income, when we try to identify the corresponding level of consumption, in all probability, we shall not be able to obtain a definite value of consumption; instead, a host of values will be available. And this will be the case with all possible levels of income. The diagram below shows this.

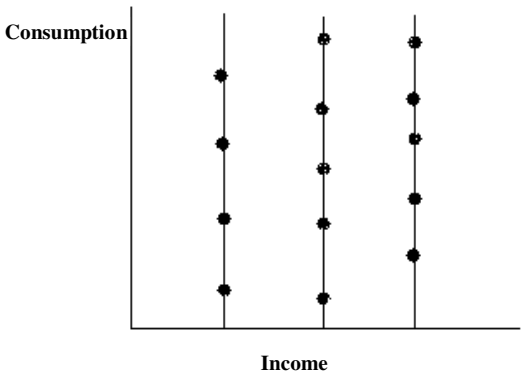


Fig. 9.1: Bi-Variate Population

Thus, the dependent variable *Y* tends to be probabilistic or stochastic in nature. In fact, for each value of the independent variable *X*, there will be a distribution of the values of the dependent variable *Y*. Accordingly; we specify the regression model by incorporating a random or stochastic variable. As mentioned in the previous unit, it should be clear that in our formulation the dependent variable is stochastic but the independent or explanatory variables are non-stochastic in nature. We should point out here that at an advanced treatment of the regression model, even the explanatory variables are considered to be stochastic in nature. It is the element of randomness either in the dependent variable alone or both in the dependent and independent variables, that make the regression model non-deterministic in nature.

9.4 POPULATION REGRESSION FUNCTION

The first step in the regression model is the conceptualization of a population regression function. Let us assume that there is a bi-variate population of (*X, Y*) of size *N*. Suppose, in this population, *Y* is the dependent variable and *X* is the independent variable. Let *X* and *Y* be related to each in a linear fashion in the following way:

$$Y = \alpha + \beta X + U$$

The above equation is called the population regression function. Here, α and β are the unknown parameters. It is clear from our discussion on linearity that this function is linear in variable as well as in parameter.

In this population regression function, the variable *U* deserves some attention. We have already mentioned that in social Science, there cannot be an exact relationship and it is at the most statistical in nature. In our formulation, in fact, *Y* is stochastic or random, but *X* is non-stochastic or deterministic in nature. Consequently, a random variable like *U* is introduced in the population regression function to incorporate this element of randomness of a statistical relationship. The random variable *U* is also called the disturbance term. It is a sort of catch – all variables that represent all kinds of indeterminacies of an inexact relationship. Thus it may represent, for example,

a) *Inherent randomness in human behaviour*: The unpredictability of human psyche is also reflected in the human behaviour. As a result, even after taking into account of all possible factors, a regression equation cannot fully explain the value of the dependent variable for the given values of all possible explanatory variables.

b) *Effect of omitted variables*: Sometimes for the sake of parsimony, all the explanatory variables are not included in a regression model. The disturbance term can be taken as a representative variable of all such omitted variables.

c) *Effect of measurement error*: Sometimes, it is not possible to measure the values of the dependent variable accurately. Consequently, the disturbance term is introduced in the regression model to represent the combined effect of all such possible sources of measurement error.

d) *Error in the formulation*: Often, the functional form of the regression model proves to be far from a correct depiction of the underlying relationship between the dependent variable and the independent variables.

We incorporate the disturbance term to correct the distortions that may arise from a wrong specification of the regression model.

We should note that apart from the above-mentioned factors, there might be other unidentifiable factors that might also be influencing the dependent variable in an unknown manner. We introduce the disturbance term for incorporating the effect of all such unknown random factors. It should be clear that for some of the reasons mentioned above the disturbance term may assume a positive value and for some other factor it may tend to be negative. Consequently, for all practical purposes, the net or the mean effect of the random disturbance can be taken to be zero.

Difference between the Disturbance Term and the Intercept

There is a difference between the interpretation of the disturbance term and that of the intercept term α in our regression equation. As we have noted, that the disturbance represents the random effects of the known and unknown variables that have not been included in the regression model. The intercept term, on the other hand, stands for all such variables, that are known to have some definite non-random effects on the dependent variable. For example, in the demand function, if we just formulate a two variable regression equation between the quantity demanded and the price instead of a multiple regression equation, then we are consciously not including variables like prices of other commodities and the income of the consumer. And all these variables affect the quantity demanded in a known fashion. So, the intercept term here represents the average effect of all such known factors that are not explicitly included in the model.

It should be clear now that it is the disturbance term that makes the relationship statistical in nature and renders the entire procedure so rich in content.

Population Regression Line

The main objective of the regression analysis is to obtain the value of the dependent variable for a given value of the independent variable. This can be attempted by running a regression on the population regression function $Y = \alpha + \beta X + U$, i.e., by fitting a regression line through the population cluster shown in Figure 9.1. Obviously, the value of Y that we shall obtain from such a population regression line will be an average value. In other words, we shall get the equation for the population regression line by applying the expectation to the population regression function, $Y = \alpha + \beta X + U$. Thus, the population regression equation will be given by

$$E(Y / X) = \alpha + \beta X$$

It may be mentioned here that while applying the expectation operator, E (U) can be taken as zero because, as we have discussed above, the mean effect of U tends to be zero.

Sometimes we call the above expression, the population regression line and loosely write it as

$$Y = \alpha + \beta X$$

In this expression, however, we should be clear that Y stands for the conditional mean of Y for a given X.

Assumptions of the Classical Regression Model

As we have mentioned earlier, the population regression model $Y = \alpha + \beta X + U$ is unknown in the sense that that the numerical values of its parameters are not known. As a result, we are not in a position to find the value taken by *Y* on an average for a given value of *X*. This means that we have to estimate these parameters from sample information. We already know that a very simple and popular method of estimation is the least square procedure. However, we should ensure that these least square estimates faithfully represent the unknown population parameters, otherwise, the whole purpose is defeated. This is possible only if the disturbance term *U* satisfies some assumptions. These assumptions along with two other that we have already mentioned about, are known as the assumptions of the classical regression model. The assumptions are:

1) The disturbance term *U* has a zero mean for all the values of *X*, i.e., $E(U) = 0$.
2) Variance of *U* is constant for all the values of X, i.e., $V(X) = \sigma^2$. It may be mentioned here that this assumption is known as the assumption of homoscedasticity in the econometric literature.
3) The disturbance terms for two different values of *X* are independent i.e., $\text{cov}(U_i, U_j) = 0$, for $i \neq j$.
4) *X* is non-stochastic.
5) The model is linear in parameter.

The assumption of non-stochasticity of *X* leads to a corollary. It is, *X* and *U* are independent i.e., $\text{cov}(X, U) = 0$.

9.5 SAMPLE REGRESSION FUNCTION

To put the whole discussion in the proper perspective; we have an unknown bi-variate population. The only information that we have is some randomly selected values of *Y* from this population that correspond to some fixed values of *X*. Suppose, in this way we have *n* pairs of values of *X* and *Y*. So, we can say that we have a random sample of size *n*. Our purpose is to estimate the parameters of this unknown population from our sample information so that we can have a reasonable estimate of an average value of *Y* for a given value of *X* that is valid for the entire population. In other words, our objective is to estimate the population regression function from the sample observations. We can proceed in that direction by discussing the concept of a sample regression function. The form of the sample regression function is quite similar to that of the population regression function. It can be presented as

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X + \hat{U}$$

In the above expression, $\hat{Y}, \hat{\alpha}, \hat{\beta}$ and $\hat{U}$ should be interpreted as the sample estimators for their respective unknown population counterparts. Thus, we are hypothesizing that corresponding to the linear population regression function $Y = \alpha + \beta X + U$, there is a linear sample regression function $\hat{Y} = \hat{\alpha} + \hat{\beta} X + \hat{U}$. Our purpose is now to estimate the population regression function from the sample regression function. It means that we have to estimate $\hat{\alpha}$ and $\hat{\beta}$ from the given sample and take them as the estimates for the unknown population

parameters. It should be clear that due to sampling fluctuations, we might obtain different values of $\hat{\alpha}$ and $\hat{\beta}$ from different samples. What we have to ensure is that on the average, they represent the population parameters. Thus, the entire issue now boils down to the estimation of $\hat{\alpha}$ and $\hat{\beta}$. This is what is known as the estimation of the sample regression function and we are going to discuss it in the next section.

9.6 ESTIMATION OF SAMPLE REGRESSION FUNCTION

The commonly used procedure is the least square method that we discussed in the previous unit. To recapitulate, we have a sample of observations represented by a cluster of points in the sample space and we have to pass a regression line through this cluster in such a fashion that the sum of the squares of the deviations of the observed values of *Y* from their estimated values from the regression line for different values of *X* is minimum. We are reproducing the two relevant diagrams from Unit 8 for the sake of convenience.

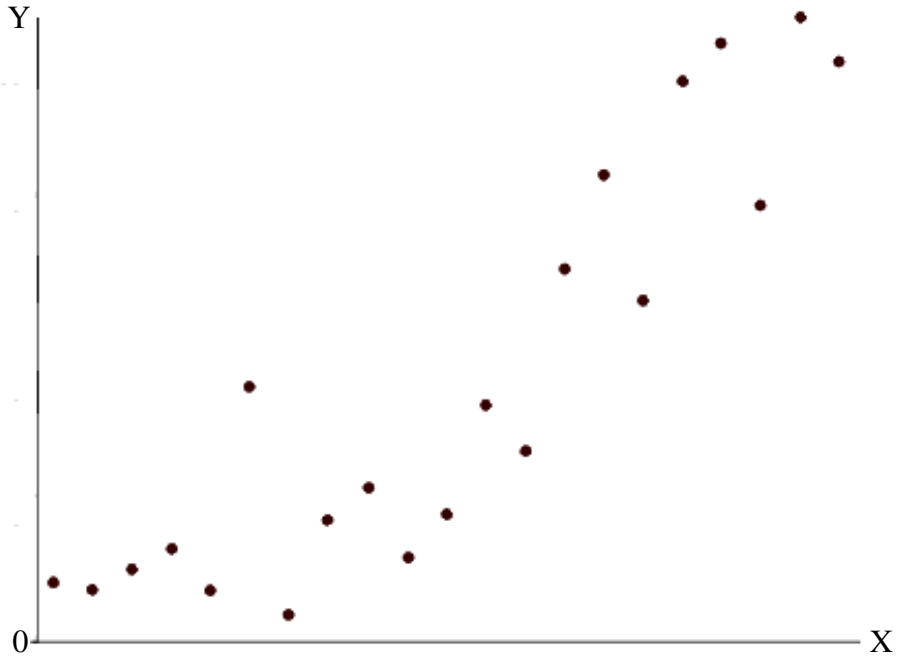


Fig. 9.2: Sample Scatter-Plot

Thus, we have a scatter of sample values as shown in the above diagram and we have to pass a regression line through it. We can summarise the least square procedure for obtaining such a line now. Putting mathematically the procedure amounts to Minimize

$$\sum u^2 = \sum \left(Y - \hat{Y}\right)^2 = \sum \left(Y - \hat{\alpha} - \hat{\beta}X\right)^2 \text{ with respect to } \hat{\alpha} \text{ and } \hat{\beta}.$$

Following the usual minimization procedure, we obtain two normal equations given by

$$\sum Y = n\hat{\alpha} + \hat{\beta}\sum X$$

and

$$\sum XY = \hat{\alpha}\sum X + \hat{\beta}\sum X^2 .$$

Solving the two normal equations simultaneously, we obtain

$$\hat{\beta} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

and

$$\hat{\alpha} = \frac{\sum X^2 \sum Y - \sum X \sum XY}{n\sum X^2 - \left(\sum X\right)^2} = \bar{Y} - \hat{\beta}\bar{X} .$$

Let us now derive an important result for the estimated slope coefficient $\hat{\beta}$.

Writing lower case letters for deviations

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{\sum x(Y - \bar{Y})}{\sum x^2} = \frac{\sum xY - \bar{Y}\sum (X - \bar{X})}{\sum x^2}$$

But $\sum (X - \bar{X}) = 0$

Thus $\hat{\beta} = \dfrac{\sum xY}{\sum x^2}$

Now putting $k = \dfrac{x}{\sum x^2}$, as x's are given.

We have

$$\hat{\beta} = \sum kY$$

Thus, $\hat{\beta}$ is a linear function of the observed values of *Y*. This is an important result, which we shall use later in the issue of hypothesis testing.

The following diagram shows a regression line, with the values of $\hat{\alpha}$ and $\hat{\beta}$ given by the above expressions fitted into our sample scatter.

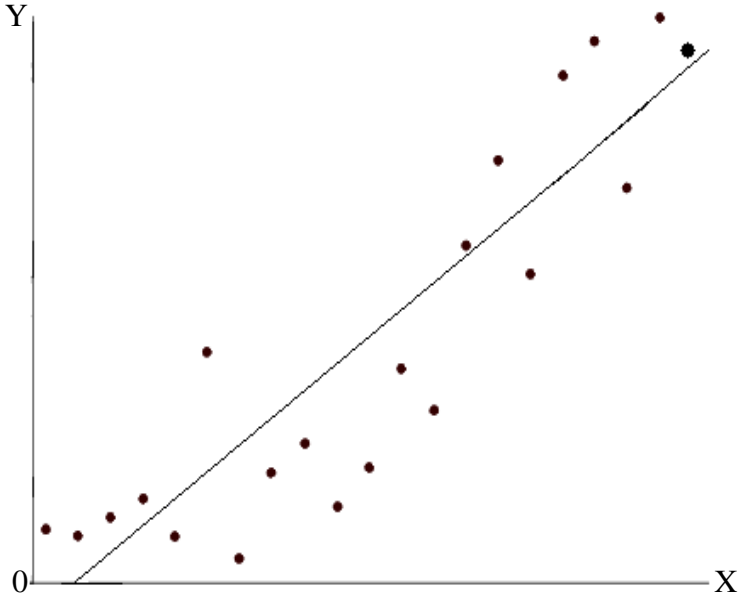


Fig. 9.3: Sample Scatter-Plot with the Sample Regression Line

It may be noted here that the sample regression line has the equation, $Y = \hat{\alpha} + \hat{\beta}X$ with the values of $\hat{\alpha}$ and $\hat{\beta}$ obtained from the above-mentioned least square method. It can be clearly seen that the values of $\hat{\alpha}$ and $\hat{\beta}$ can be calculated in terms of sample observations only and no other information that is not known is required for this purpose. As we can see, both $\hat{\alpha}$ and $\hat{\beta}$ are linear functions of *X* and *Y*. Consequently, they are called linear estimators.

Gauss-Markov Theorem

The least square estimators $\hat{\alpha}$ and $\hat{\beta}$ can be taken as the estimators of the unknown population parameters α and β because they satisfy certain desirable properties. We can state them without proofs below.

Under the assumptions of the classical regression model as mentioned earlier,

1) Least square estimators are linear. As we have already seen from the expressions of $\hat{\alpha}$ and $\hat{\beta}$, they are linear functions of the variables.

2) Least square estimators are unbiased i.e., $E(\hat{\alpha}) = \alpha$ and $E(\hat{\beta}) = \beta$. This means, if we consider different values of the least square estimates obtained from a number of random samples for a given population on the average, they will be equal to the unknown population parameters. They will not be systematically overestimating or underestimating the population parameters.

3) Among all the linear unbiased estimators, least square estimators have the minimum variance. In this sense, they are termed as the efficient estimators.

All these properties of the least square estimators lead to what is known as the Gauss-Markov Theorem. The theorem states:

> **Under the assumptions of the classical linear regression model, among all the linear unbiased estimators, the least square estimators have the minimum variance. In other words, the least square estimators are the best linear unbiased estimators or in short, BLUE.**

Standard Error of the Regression Estimate

We have already stated that due to sampling fluctuations, we should expect to obtain different values of the least square estimates $\hat{\alpha}$ and $\hat{\beta}$ from sample to sample. Consequently, if we measure the standard deviations of these values of the two least square estimates from their respective expectation or mean, we shall get an idea about the extent to which these estimates are affected by the sampling fluctuations. In other words, the standard deviations of the least square estimates can be taken as a measure of the precision of these estimates. **The standard deviations of the least square estimates are known as the standard errors of the estimates.** It should be clear that the standard errors are the standard deviations of the sampling distributions of the least square estimates. The standard deviations or standard errors are obtained by taking the positive square root of the variances of $\hat{\alpha}$ and $\hat{\beta}$. The expressions for both the variance and standard of the least square estimators are given below.

$$\text{var}(\hat{\alpha}) = \frac{\sum X^2}{n\sum\left(X - \bar{X}^2\right)}\sigma^2$$

$$se(\hat{\alpha}) = \sqrt{\frac{\sum X^2}{n\sum\left(X - \bar{X}^2\right)}\sigma^2}$$

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{\sum\left(X - \bar{X}\right)^2}$$

$$se(\hat{\beta}) = \frac{\sigma}{\sqrt{\sum\left(X - \bar{X}\right)^2}}$$

It should be recollected that σ is the standard deviation of the error term *U* in our population regression model and we assume that it is a constant. However, since the population regression model is unknown, it follows that σ is also unknown. As a result, for the calculation of variance and standard error of the least square estimates $\hat{\alpha}$ and $\hat{\beta}$, we need to estimate σ. It can be proved that an unbiased estimator of σ is $\sqrt{\frac{\sum \hat{U}^2}{n-2}}$. Here, *n – 2* is what is known as the degrees of freedom in statistics. This concept you must have studied in your compulsory course in statistics. Now, $\hat{U} = Y - \hat{Y}$ is the sample regression error term and accordingly we can calculate it. Thus, by replacing σ by its unbiased estimator we can compute the standard errors of both $\hat{\alpha}$ and $\hat{\beta}$. It should be noted here that we can write the unbiased estimator of σ as

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{U}^2}{n-2}} = \sqrt{\frac{\sum\left(\hat{U} - \bar{\hat{U}}\right)^2}{n-2}}$$

This is, because, $\bar{\hat{U}}$ is the mean of the sample regression errors. Now, while calculating it, the positive errors will tend to cancel the negative errors and as a result, it will be reduced to zero. From above expression of the estimator of σ, it can easily be interpreted as the standard deviation of the sample observations about the estimated sample regression line. The sample estimator $\hat{\sigma} = \sqrt{\frac{\sum \hat{U}^2}{n-2}}$ is known as the standard error of estimate or the standard error of the regression.

9.7 GOODNESS OF FIT

In the earlier section, we have discussed the procedure for fitting the sample regression line and its purpose. Once such a regression line is fitted, we may be interested in examining how good has been the fit in the sense how faithfully it can describe the unknown population regression line. This is known as the issue of goodness of fit. In this matter, the regression error term or residual $\hat{U}$

plays an important role. Small quantities of residuals imply that a large proportion of variation in the dependent variable has been explained by the regression equation and consequently, the fit is good. Similarly, large quantities of residuals obviously point to a poor fit. At this stage, what we are interested in is to obtain a quantitative measure of the goodness of fit that is free of any unit for the purpose of comparability. In the previous unit, we have referred to the coefficient of determination, which is the square of the correlation coefficient. It can be shown that this coefficient of determination acts as a measure of goodness of fit. We may consider it now. The variation in the dependent variable Y about its mean can be conceptualized as

$$\text{var}(Y) = \sum \left(Y - \bar{Y}\right)^2$$

We can decompose it into two components. The first is the variation explained by the regression. The second is the portion that remains unexplained by the regression. Thus we can write

$$\left(Y - \bar{Y}\right) = \left(Y - \hat{Y}\right) + \left(\hat{Y} - \bar{Y}\right)$$

Squaring and applying summation on both the sides

$$\sum \left(Y - \bar{Y}\right)^2 = \sum \left(\hat{Y} - \bar{Y}\right)^2 + \sum \left(Y - \hat{Y}\right)^2 + 2\sum \left(\hat{Y} - \bar{Y}\right)\left(Y - \hat{Y}\right)$$

Now let us try to find the value of $\sum \left(\hat{Y} - \bar{Y}\right)\left(Y - \hat{Y}\right)$. We have the regression equation

$$Y = \hat{\alpha} + \hat{\beta}X + \hat{U}$$

$$\text{or } \bar{Y} = \hat{\alpha} + \hat{\beta}\bar{X} + \bar{\hat{U}}$$

Subtracting the second equation from the first equation

$$Y - \bar{Y} = \hat{\beta}\left(X - \bar{X}\right) + \hat{U}, \text{ because } \bar{\hat{U}} = \frac{\sum \hat{U}}{n} = 0, \text{ as } \sum \hat{U} = 0.$$

Writing lower case letters for the deviations of the variables from their mean, we have the sample regression equation in the form

$$y = \hat{\beta}x + \hat{u}$$

Now

$$\begin{aligned}
&\sum \left(\hat{Y} - \bar{Y}\right)\left(Y - \hat{Y}\right) \\
&= \sum \left(\hat{Y} - \bar{Y}\right)\hat{U} \\
&= \sum \hat{Y}\hat{U} - \bar{Y}\sum \hat{U} \\
&= \sum \hat{Y}\hat{U}, \text{ because } \sum \hat{U} = 0 \\
&= \sum \left(\hat{\alpha} + \hat{\beta}\text{X}\right)\hat{U} \\
&= \hat{\alpha}\sum \hat{U} + \hat{\beta}\sum X\hat{U} \quad \because \sum \hat{U} = 0 \\
&= \hat{\beta}\sum X\hat{U}
\end{aligned}$$

We have

$x = X - \bar{X}$

$\Rightarrow \sum x = \sum \left(X - \bar{X} \right)$

or $\sum x\hat{U} = \sum \left(X - \bar{X} \right)\hat{U} = \sum X\hat{U} - \bar{X}\sum \hat{U} = \sum X\hat{U}$

Now

$\sum x\hat{U} = \sum x\left(y - \hat{\beta}x \right) = \sum xy - \hat{\beta}\sum x^2$

We know, $\hat{\beta} = \dfrac{\sum \left(X - \bar{X} \right)\left(Y - \bar{Y} \right)}{\sum \left(X - \bar{X} \right)} = \dfrac{\sum xy}{\sum x^2}$

Putting the value of $\hat{\beta}$ in the above expression for $\sum x\hat{U}$,

$\sum x\hat{U} = \sum xy - \dfrac{\sum xy}{\sum x^2}\sum x^2 = \sum xy - \sum xy = 0$

Hence,

$\sum \left(\hat{Y} - \bar{Y} \right)\left(Y - \hat{Y} \right) = 0$

Thus,

$\sum \left(Y - \bar{Y} \right)^2 = \sum \left(\hat{Y} - \bar{Y} \right)^2 + \sum \left(Y - \hat{Y} \right)^2$

If $\sum \left(Y - \bar{Y} \right)^2$ is defined as the total sum of squares (TSS), $\sum \left(\hat{Y} - \bar{Y} \right)^2$ as the explained sum of squares (ESS) and $\sum \left(Y - \hat{Y} \right)^2$ as the residual sum of squares (RSS),

We have

TSS = ESS + RSS

Dividing both the sides by TSS,

$\dfrac{\text{TSS}}{\text{TSS}} = \dfrac{\text{ESS}}{\text{TSS}} + \dfrac{\text{RSS}}{\text{TSS}}$

Defining the ratio of ESS to TSS as R^2, we have

$\text{R}^2 = 1 - \dfrac{\text{RSS}}{\text{TSS}} = \dfrac{\text{ESS}}{\text{TSS}}$

It should be clear now that R^2 or coefficient determination can be interpreted as the proportion of total variation in *Y* explained by the regression of *Y* on *X*.

Let us now consider a numerical example to fix the concepts and ideas that we have discussed so far.

Example 9.1

The hypothetical figures for the total labour force and the number of employed out of that for the period 1991-2000 are given below in Table 9.1. Run a regression of the number of employed on the total labour force.

Table 9.1: Actual Employment and Labour Force

Year	Employed (million)	Labour Force (million)
1991	100	120
1992	125	140
1993	140	165
1994	160	185
1995	175	200
1996	195	210
1997	230	250
1998	245	255
1999	270	305
2000	295	320

Let Y be the dependent variable employed and X be the independent variable labour force. The detailed calculation for working out the regression is shown below:

Year	Y	X	XY	X^2	$y = Y - \bar{Y}$	$x = X - \bar{X}$	x^2	y^2	xy
1991	100	120	12000	14400	-93.5	-95	9025	8742.25	8882.5
1992	125	140	17500	19600	-68.5	-75	5625	4692.25	5137.5
1993	140	165	23100	27225	-53.5	-50	2500	2862.25	2675
1994	160	185	29600	34225	-33.5	-30	900	1122.25	1005
1995	175	200	35000	40000	-18.5	-15	225	342.25	277.5
1996	195	210	40950	44100	1.5	-5	25	2.25	-7.5
1997	230	250	57500	62500	36.5	35	1225	1332.25	1277.5
1998	245	255	62475	65025	51.5	40	1600	2652.25	2060
1999	270	305	82350	93025	76.5	90	8100	5852.25	6885
2000	295	320	94400	102400	101.5	105	11025	10302.25	10657.5
Sum	1935	2150	454875	502500			40250	37902.5	38850
Mean	193.5	215							

$$\hat{\beta} = \frac{\sum xy}{\sum x^2} = \frac{38850}{40250} = 0.965217 \text{ and } \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 193.5 - (0.965217 \times 215) = -14.217$$

Year	$\hat{Y} = \hat{\alpha} - \hat{\beta}X$	$\hat{U} = Y - \hat{Y}$	$\hat{U}^2$	$\hat{Y} - \bar{Y}$	$(\hat{Y} - \bar{Y})^2$
1991	101.8043	-1.80434	3.255643	-91.6957	8408.094
1992	121.1087	3.89132	15.14237	-72.3913	5240.503
1993	145.2391	-5.23911	27.44822	-48.2609	2329.114
1994	164.5434	-4.54344	20.64289	-28.9566	838.4821
1995	179.0217	-4.0217	16.17407	-14.4783	209.6212
1996	188.6739	6.32613	40.01992	-4.82613	23.29153
1997	227.2826	2.71745	7.384535	33.78255	1141.261
1998	232.1086	12.89137	166.1873	38.60864	1490.627
1999	280.3695	-10.3695	107.5262	86.86949	7546.307
2000	294.8477	0.15226	0.023183	101.3477	10271.36
Sum		0.00045	403.8043		37498.67

In the next stage we compute the following:

$$V(\hat{\beta}) = \frac{\sigma^2}{\sum x^2},\ V(\hat{\alpha}) = \frac{\sum X^2}{n\sum x^2}\sigma^2 \text{ and } \mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}.$$

But the population variance σ^2 is unknown. So we have to use an unbiased sample estimator σ^2 for the same. Such an unbiased estimator is given by $\hat{\sigma}^2 = \frac{\sum \hat{U}^2}{n-2}$, where $n - 2$ is the degrees of feedom. Here, $n - 2 = 10 - 2 = 8$.

Therefore, $\hat{\sigma}^2 = \frac{\sum \hat{U}^2}{n-2} = \frac{403.8043}{8} = 50.475537.$

Thus, $V(\hat{\beta}) = \frac{\hat{\sigma}^2}{\sum x^2} = \frac{50.475537}{40250} = 0.001254$

$$s.e.(\hat{\beta}) = \sqrt{V(\hat{\beta})} = \sqrt{0.001254} = 0.0354118$$

$$V(\hat{\alpha}) = \frac{\sum X^2}{n\sum x^2}\hat{\sigma}^2 = \frac{502500}{10 \times 40250} \times 50.475537 = 63.016042$$

$$s.e.(\hat{\alpha}) = \sqrt{V(\hat{\alpha})} = \sqrt{63.016042} = 7.9382644$$

$$\mathrm{R}^2 = \frac{\mathrm{ESS}}{\mathrm{TSS}} = \frac{37498.67}{37902.5} = 0.989345$$

Finally, we present a summary of all the important results concerning the above-mentioned question.

$\hat{\beta} = 0.965217$ $V(\hat{\beta}) = 0.001254$ and $s.e.(\hat{\beta}) = 0.0354118$

$\hat{\alpha} = -14.0217$ $V(\hat{\alpha}) = 63.016042$ and $s.e.(\hat{\alpha}) = 7.9382644$

$\hat{\sigma}^2 = 50.475537$

$R^2 = 0.989345$ Degrees of Freedom $(d.f.) = 8$

The estimated regression line is given by

$\hat{Y} = -14.0217 + 0.965217X$

It may be mentioned here that it is this kind of a summary of the results that is generally used for the further analysis of the regression model. Let us now interpret some of the results from the above summary. The regression line estimates the average employment (Y) for a given level of labour force (X). The slope coefficient $\hat{\beta} = 0.965217$ estimates the rate of change of employment with respect to labour force. For example, if 100 more persons start looking for jobs, about 97 of them actually get employed. The intercept $\hat{\alpha} = -14.0217$ can be interpreted as the average combined effect of all those variables that might also affect employment but have been omitted for the purpose of the above-mentioned regression. The coefficient of regression $R^2 = 0.989345$ indicates that about 99 per cent of the variation in employment can be explained by a variation in the labour force, which is indeed high and indicates a good fit to the given sample.

Check Your Progress 1

1) Discuss the meaning of linearity in the regression model.

……

……

……

……

……

2) How is the regression model non-deterministic in nature?

……

……

……

……

……

……

……

3) Discuss the assumptions of the classical regression model.

……

……

……

……

……

4) State the Gauss-Markov Theorem.

……

……

……

……

……

5) Explain the concept of goodness of fit.

……

……

……

……

……

9.8 FUNCTIONAL FORMS OF REGRESSION MODEL

We have already mentioned that our focus is on a linear regression model. The issue of linearity has also been considered in Section 9.2. It is in fact linear in parameter regression model that is relevant for us. However, linear in parameter regression models may also have different functional forms. We shall now briefly discuss four such regression models.

1) **Linear Model**

This functional form is the most common one and we have already discussed it. Its equation is given by

$$Y = \alpha + \beta X + U$$

As we know, it is both linear in parameter and linear in variable. This model can be estimated by the ordinary least square (OLS) method. The least square

estimators $\hat{\alpha}$ and $\hat{\beta}$ are the unbiased estimators of the unknown population parameters α and β.

Here the regression coefficient β measures the rate of change of *Y* per unit change of *X*.

Example 9.2

In this example we shall consider the issue of fitting a Keynesian consumption function to the Indian data. Table 9.2 presents the GDP at factor cost and final consumption expenditure (FCE) figures for the Indian economy during the period 1980-2001 at 1993-94 prices.

Table 9.2: GDP and Final Consumption Expenditure (Rs. Crore)

Year	GDP	FCE	Year	GDP	FCE
1980	401128	390671	1991	701863	620806
1981	425073	407728	1992	737792	637307
1982	438079	415924	1993	781345	666950
1983	471742	446396	1994	838031	695825
1984	492077	461675	1995	899563	740109
1985	513990	485090	1996	970082	794093
1986	536257	504854	1997	1016595	826834
1987	556778	525558	1998	1082747	888513
1988	615098	557527	1999	1148367	952489
1989	656331	584970	2000	1198592	972932
1990	692871	610169	2001	1267945	1027254

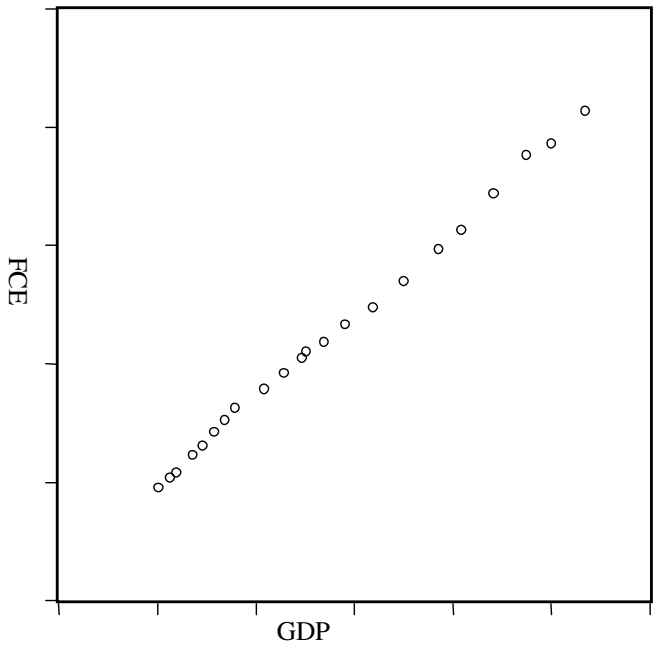


Fig. 9.3: Scatter of Final Consumption Expenditure against GDP

The above scatter of final consumption expenditure against GDP been obtained from Table 9.2. The scatter makes it clear that there seems to be a linear relationship between the two variables. Consequently, we have run a linear regression between the variables by using the data of Table 9.2. The results are presented below:

$$FCE = 108206.4 + 0.719674\ GDP$$

$$s.e.(\hat{\alpha}) = 6233.203 \qquad s.e.(\hat{\beta}) = 0.007865$$

$$t\ (\hat{\alpha}) = 17.35968 \qquad t\ (\hat{\beta}) = 91.50314$$

$$R^2 = 0.997617 \qquad d.f. = 20$$

A high R^2 of more than 0.99 implies a tight fit. The estimated $\hat{\beta} = 0.719674$ indicates a marginal propensity to consume of about 72 percent and this can be quite expected in the Indian economy. We shall explain the significance of degrees of freedom and *t* values later in the issue of tests of hypotheses. However, we shall continue to present these two statistics in our subsequent examples also.

2) **Log-linear Model**

This model is also known as log-log, double-log or constant elasticity model. Its original form is given by

$Y = \alpha X^{\beta} e^{U}$, Where, *e* is the base of natural log (*e* is approximately equal to 2.718). In the original form, the model is non linear in nature. However, by applying the log transformation, the model is transformed to

$$\ln Y = \ln \alpha + \beta \ln X + U$$

If we put $\ln Y = y$, $\ln \alpha = a$, $\beta = b$ and $\ln X = x$, the model reduces to

$$y = a + bx + U$$

From the above transformation, it is very clear that the model now becomes linear in parameter and linear in log of the variables; and thus, can be estimated by the ordinary least square method. Here, *a* will be estimating $\ln \alpha$ and then by taking the antilog of *a*, we shall obtain the estimate for α. The regression coefficient $b = \beta$ deserves our special attention. It measures a change in log *Y* per unit of a change in log *X*. In the language of calculus, $\beta = \dfrac{d \ln Y}{d \ln X}$. Now, a change in the log of some variable implies a proportional change in it. Thus, β is the ratio of a proportional change in *Y* to a proportional change in *X*. in other words, β measures the elasticity of *Y* with respect to *X*. This is the rationale for the log linear regression model being termed as the constant elasticity model.

Example 9.3

This is an example from Gujrati (2003). The table below presents the quarterly data on total personal consumption expenditure (PCExp) and expenditure on durables (ExpDur) for the U.S. economy measured in 1992 billions of dollar for the period 1993:1-1998:3.

Table 9.3: Personal Consumption Expenditure and Expenditure on Durables in U.S.A.

Year	PCExp	ExpDur
1993:1	4286.8	504
1993:2	4322.8	519.3
1993:3	4366.6	529.9
1993:4	4398	542.1
1994:1	4439.4	550.7
1994:2	4472.2	558.8
1994:3	4498.2	561.7
1994:4	4534.1	576.6
1995:1	4555.3	575.2
1995:2	4593.6	583.5
1995:3	4623.4	595.3
1995:4	4650	602.4
1996:1	4692.1	611
1996:2	4746.6	629.5
1996:3	4768.3	626.5
1996:4	4802.6	637.5
1997:1	4853.4	656.3
1997:2	4872.7	653.8
1997:3	4947	679.6
1997:4	4981	648.8
1998:1	5055.1	710.3
1998:2	5130.2	729.4
1998:3	5181.8	733.7

Source: *Damodar Gujrati (2003), Basic Econometrics*

We have plotted the log of expenditure on durables against the log of total personal consumption expenditure in Figure 9.4 below.

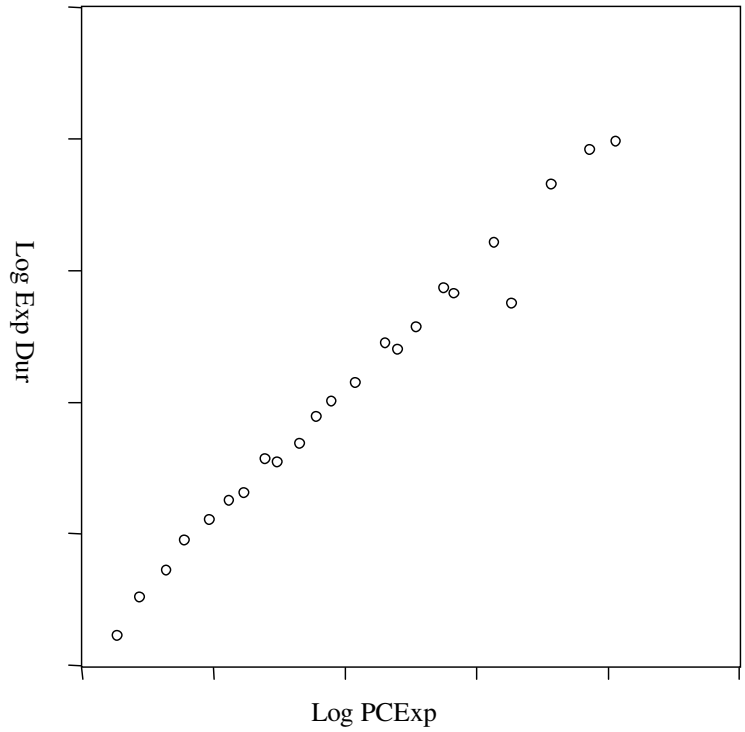


Fig. 9.4: Scatter of Log of Exp on Durables against Log of Personal Con Exp

The scatter above is clearly indicative of a linear relationship between the log of the two variables. As a result, we fit in a log linear model to the data on total personal consumption expenditure and the expenditure on consumer durables. We present some of the results from the fit below.

$$\ln ExpDur = -9.697098 + 1.905633 \ln PCExp$$

$$s.e.(\ln \hat{\alpha}) = 0.434127 \qquad s.e.(\hat{\beta}) = 0.051370$$

$$t\ (\ln \hat{\alpha}) = -22.33702 \qquad t\ (\hat{\beta}) = 37.09622$$

$$R^2 = 0.984969 \qquad d.f. = 21$$

We have an R^2 of about 0.99. This is indicative of a good fit. In this fit, the slope coefficient is an estimate of the elasticity of expenditure on durables with respect to personal consumption expenditure. Thus, $\hat{\beta} = 1.905633$ indicates that the expenditure on durables is highly elastic to total personal consumption expenditure.

3) **Semi-Log Model**

One of the common forms of semi-log regression models is the so-called growth rate model. The model is expressed as

$$\log Y = \alpha + \beta t + U$$

In this model the dependent variable is expressed in log, whereas, the independent variable is the time period, and it is measured in absolute term. Since, log operator is applied only to the left hand side of the equation, the model is called the semi-log model. In this model, the slope coefficient β is a measure of the proportional change in the dependent variable for a unit change in the time period. Thus it measures the growth rate of the dependent variable.

Example 9.4

In this example, we are going to verify the so-called 'Hindu Rate of Growth' of about 3.5 per cent that Indian economy consistently witnessed in 1960s and 1970s.

In Table 9.4, we present India's GDP at 1993-94 prices for the period 1960-1980.

Table 9.4: India's GDP during 1970-1980

Year	GDP	Year	GDP
1960	206103	1971	299269
1961	212499	1972	298316
1962	216994	1973	311894
1963	227980	1974	315514
1964	245270	1975	343924
1965	236306	1976	348223
1966	238710	1977	374235
1967	258137	1978	394828
1968	264873	1979	374291
1969	282134	1980	401128
1970	296278		

Figure 9.5 presents the scatter of the log of GDP against the time period denoted by the letter *t*.

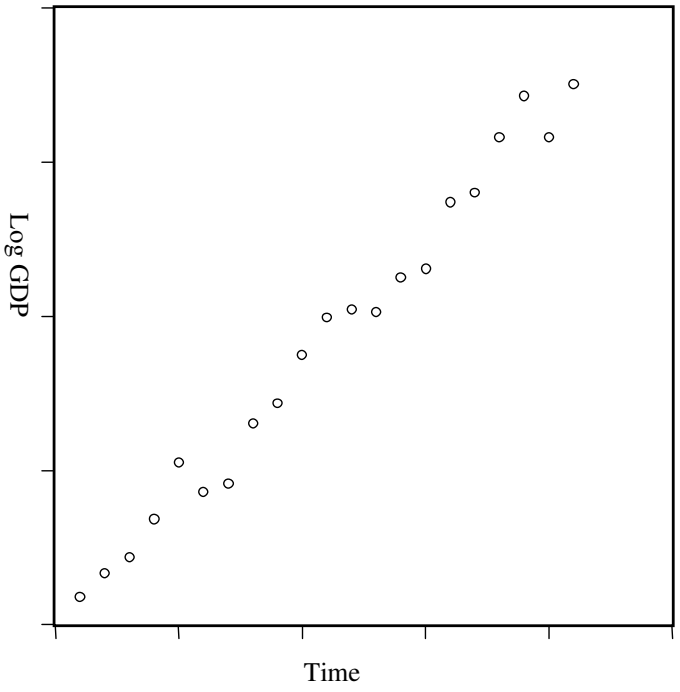


Fig. 9.5: Scatter of India's GDP against Time

The scatter above suggests a linear relationship between the log of GDP and the time period. Therefore, we run a regression of log GDP on time period *t*. we present the result below.

$$\log GDP = 12.19473 + 0.033729\,t$$
$$s.e.(\hat{\alpha}) = 0.012517 \qquad s.e.(\hat{\beta}) = 0.000997$$
$$t(\hat{\alpha}) = 974.2658 \qquad t(\hat{\beta}) = 33.83607$$
$$R^2 = 0.983675 \qquad d.f. = 19$$

The value of R^2 is quite high. The estimated slope coefficient of 0.034 indicates a rate of growth of 3.4 per cent during the period 1960-1980. This adequately supports the phenomenon of 'Hindu rate of Growth' in the 60s and the 70s.

4) **Reciprocal Model**

In the reciprocal model, the dependent variable is regressed on the reciprocal of the independent variable. Its functional from is

$$Y = \alpha + \beta\frac{1}{X} + U$$

Although the model is non-linear in variable (because the power of X is –1), it is linear in parameter. Hence, it can be estimated by the OLS method. The model has an important characteristic. As the independent variable X increases indefinitely, the term $\beta\frac{1}{X}$ approaches zero, β being a constant. Consequently, the dependent variable Y approaches a limiting value or an asymptotic value equal to the intercept α. An application of this model can be in the relationship between per capita GNP and the child mortality. As per capita GNP increases, the infant mortality rate is expected to fall but we cannot expect it to fall independently if per capita GNP continues to increase indefinitely. In such a situation, in all probability, the mortality rate will tend to a limiting value. Gujrati (1993) presents an example of applying the reciprocal model to study the relationship between the per capita GNP and the infant mortality rate by using the cross-section data for 64 countries. His results are presented below.

$$Y = 81.79436 + 27237.17\frac{1}{X}$$
$$s.e.(\hat{\alpha}) = 10.8321 \qquad s.e.(\hat{\beta}) = 3759.999$$
$$t(\hat{\alpha}) = 7.5511 \qquad t(\hat{\beta}) = 7.2535$$
$$R^2 = 0.4590$$

where, Y is the child mortality rate (the number of deaths of children under the age of 5 per year per 1000 live births) and X is the per capita GNP at constant prices. The above-mentioned results show that if per capita GNP increases indefinitely, the infant mortality approaches a limiting rate of about 82 deaths per 1000 children born.

Another application of the reciprocal model can be in the estimation of the Phillips curve. This curve proposes an inverse relationship between the change in unemployment rate and that in inflation rate. With a growth in unemployment rate, the growth in inflation rate is expected to fall. In fact Phillips curve postulates that if the unemployment rate increases beyond its so-called natural rate, the growth in the inflation rate should become negative. However, with an indefinite increase in the unemployment rate, the inflation

growth cannot be expected to fall indefinitely. It should stabilize at some negative limiting value. Thus, Phillips Curve can be an appropriate case for the use of the reciprocal model.

9.9 CLASSICAL NORMAL REGRESSION MODEL

We have already seen that a two variable classical linear regression model can be presented as

$$Y = \alpha + \beta X + U$$

In this model, U is the population disturbance term. We can estimate the unknown parameters of this regression model from the sample information by using the least square or, as it is sometimes called, ordinary least square method. The least square estimates possess some desirable properties if the population disturbance U satisfies some five assumptions that we have discussed in Section 9.3. These five assumptions are adequate for the estimation purposes. But the scope of a regression model is just not restricted to the estimation of the parameters. An important purpose of the sample estimates is to test some hypotheses about the unknown population regression parameters with their help. And we can do this if we make some assumption about the distribution of U. This we do by making the assumption of normality for U. The assumption essentially means that the population regression disturbance term follows normal distribution with mean zero, a constant variance equal to σ^2 and a zero covariance. In fact U has a conditional distribution, in the sense, that, for each of the given values of the non-stochastic independent variable X, we might have a distribution of different values of U. This will be clear if we reproduce the diagram of Figure 9.1 with the unknown population regression line fitted into it.

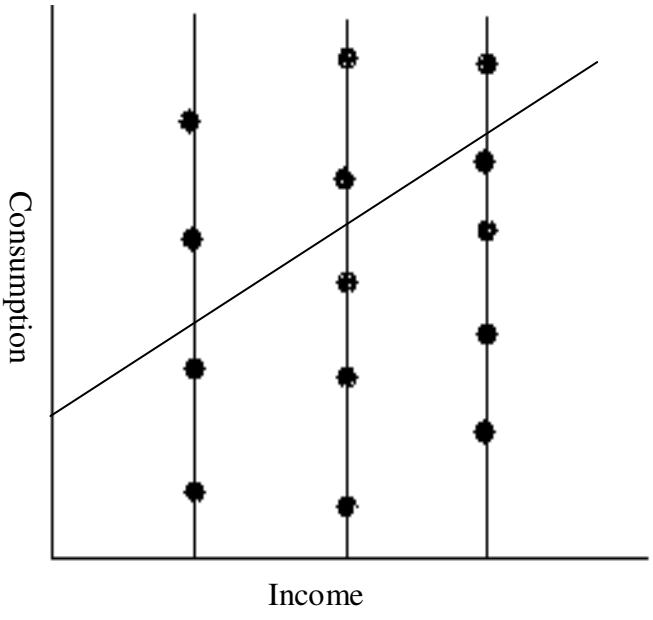


Fig. 9.6: Bi-Variate Population with the Unknown Regression Line

In Figure 9.6, we can see that for a given observed value of income, we can have different observed values of consumption represented by the bold dots on each of the vertical lines. The vertical distance between an observed value of consumption and the corresponding estimated value from the regression line for a given income level measures the value of the disturbance term. Thus there

are many possible values of *U* for each of the given income levels. As a result, there are conditional distributions of *U* for different levels of income.

The significance of the normality assumption is that these conditional distributions of *U* should all be independently normally distributed with the same mean equal to zero and, the same variance equal to σ^2. Mathematically speaking,

$$E(U_i) = 0 \qquad \text{for all } i\text{s}$$
$$V(U_i) = \sigma^2 \qquad \text{for all } i\text{s}$$
$$Cov(U_i U_j) = 0 \quad \text{for } i \neq j$$

In other words, the population disturbance variable should have independent and identical normal distribution.

If we make this additional normality assumption along with the earlier-mentioned five assumptions about the disturbance term U, then the regression model is called the classical normal regression model.

It may be mentioned here that the normality assumption is important not only for hypothesis testing but also for the fact that under this assumption we can use an alternative procedure for estimating the population regression parameters. This procedure is known as the method of maximum likelihood estimation. Although for regression parameters the method gives the same estimates as the least square estimates, the approach followed is quite different. In fact this procedure has some stronger theoretical characteristics than the least square method. However, we are not discussing this procedure here because it is beyond our scope.

In the next section, we shall discuss the issue of hypothesis testing and see how the normality assumption plays a crucial role there.

9.10 HYPOTHESIS TESTING

Considering the two variable regression model

$$Y = \alpha + \beta X + U$$

We might be interested in examining whether the unknown parameter α or β assumes a particular value or not. This is what is known as hypothesis testing in statistics. We can conduct such a test from the sample estimate of α or β as the case might be. Although we may test some hypothesis about the intercept α, our main concern in regression model is the slope coefficient β. We have the estimated sample regression function

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X + \hat{U}$$

Here, $\hat{\beta}$ is the sample estimate of β and we shall use it for estimating the unknown β. As mentioned earlier that $\hat{\beta}$ might vary from sample to sample collected from the same population. As a result, $\hat{\beta}$ is a random variable with some probability distribution. Now, for the purpose of hypothesis testing, we need to know the form of this distribution of $\hat{\beta}$.

It is here that the assumption of normality helps us. Let us consider it. From Section 9.6, we know that $\hat{\beta} = \sum kY$ and thus, it is a linear function of the observed values of Y. But, $Y = \alpha + \beta X + U$. Consequently, we can see, $\hat{\beta} = \sum k(\alpha + \beta X + U)$. Now, we already know that the ks, the Xs and the parameters are all given. As a result, finally, $\hat{\beta}$ becomes a linear function of the random disturbance variable U. Thus, $\hat{\beta}$ has the same distribution as U. Therefore, the assumption of normality U, implies that $\hat{\beta}$ is distributed as normal.

We have already seen that $E(\hat{\beta}) = \beta$ and $se(\hat{\beta}) = \frac{c}{\sqrt{\sum (X - \bar{X})^2}}$. It follows then that

$$\hat{\beta} \sim N\left(\beta, \frac{\sigma}{\sqrt{\sum (X - \bar{X})^2}}\right)$$

You must have learnt from your compulsory course in quantitative methods, that now a standard normal variable can be formed with mean equal to zero and standard deviation equal to one and it can be used for testing of hypotheses regarding β. But the problem here is that the population standard deviation c is unknown and we have to use an estimate for it. However, we know that an unbiased estimate of c is given by $\hat{\sigma}^2 = \frac{\sum \hat{U}^2}{n-2}$, where, $\hat{U}$ is the sample regression error term and is therefore, computable, and n is the size of the sample. Again you must be knowing that when we standardize $\hat{\beta}$ by subtracting its mean from it and divide it by its estimated standard deviation, it no longer follows normal distribution and it in fact follows a *student-t* distribution with *n-2* degrees of freedom. This *student-t* distribution has a mean equal to zero and a standard deviation equal to one. Thus, it is this standard *student-t* distribution that is used for testing of hypothesis about β. We have a table for such a standard *student-t* distribution for different degrees of freedom. And like the standard normal table, this table is put into use for such tests of hypotheses.

Let us see now how we can conduct some test of hypothesis regarding the unknown population regression coefficient by considering some examples that we have already discussed in Section 9.7. Consider Example 9.2. In this example, we have considered the issue of the consumption function in the Indian economy by regressing final consumption expenditure on GDP. To begin with, we might be interested in examining whether the relationship is significant or not. Thus, our focus is on the regression coefficient β and by the nature of the inquiry, we have to conduct a two-tailed test. We can proceed by setting our null hypothesis and the alternative hypothesis in the following manner:

Null hypothesis	$H_0 : \beta = 0$
Alternative hypothesis	$H_1 : \beta \neq 0$

The acceptance of the null hypothesis implies the rejection of the alternative hypothesis and this in turn implies that on the basis of the sample information there does not seem to be any significant relationship between GDP and the final consumption expenditure. On the other hand, the rejection of the null hypothesis implies the acceptance of the alternative hypothesis and this in turn implies that on the basis of the sample information there seems to exist significant relationship between GDP and the final consumption expenditure. The next step is to construct our test statistic, which is conventionally denoted by *t*. Thus

$$t = \left| \frac{\hat{\beta} - E\left(\hat{\beta}\right)}{s.e\left(\hat{\beta}\right)} \right|$$

Under the null hypothesis,

$$t = \left| \frac{\hat{\beta} - 0}{\sqrt{\sum \frac{\hat{U}^2}{n-2}}} \right| = \left| \frac{\hat{\beta}}{\sqrt{\sum \frac{\hat{U}^2}{n-2}}} \right|$$

And this is distributed as a *student-t* with *n-2* degrees of freedom. From the results of the Example 9.2 we can easily compute the value of this *t* statistic since, both the values of the estimated regression coefficient (0.719676) and the standard error of the estimated regression coefficient (0.007865) are mentioned there. In fact, even the computed value of the *t* statistic (91.50314) is given. In addition, the degree of freedom is also given (20). Normally, any test of hypothesis is conducted either at 1 per cent level of significance or at 5% level of significance. From the student-t distribution table we can find out the critical values of the test statistic t for both 1 per cent level of significance and at 5% level of significance at a degree of freedom of 20 for two-tailed test. The values are 2.845 and 2.086 respectively. Thus, the computed value of *t* far exceeds both the critical values. Thus, on the basis of the sample information we cannot accept the null hypothesis. This amounts to accepting the alternative hypothesis. Consequently, personal consumption expenditure does seem to be dependent upon GNP in India during the sample period 1980-2001.

Once the relationship between GDP and personal consumption expenditure has been established some further test can be conducted about the likely value of the marginal propensity to consume out of GNP. For example we can conduct a test regarding whether the marginal propensity to consume is 80 per cent or not. Thus we have

Null hypothesis	$H_0 : \beta = 0.80$
Alternative hypothesis	$H_1 : \beta \neq 0.80$

Thus our *t* statistic in this case is given by

$$t = \left| \frac{\hat{\beta} - E\left(\hat{\beta}\right)}{s.e\left(\hat{\beta}\right)} \right| = \left| \frac{0.719676 - 0.80}{0.007865} \right| = \left| \frac{-0.080324}{0.007865} \right| = 10.212841.$$

Now this value of the test statistic also exceeds the critical values of 2.845 and 2.086 for a degree of freedom of 20 at 1 per cent and 5 per cent levels of significance respectively. Thus, on the basis of the sample information the

difference between estimated value of β and its mean is so much that even in 1 out 100 cases or 5 out of 100 cases we do not expect to obtain such a difference. Thus on the basis of the sample information, we are not in a position to accept the null hypothesis. Hence, in all probability during the sample period of 1980-2001, India's marginal propensity to consume out GDP has not been as high as 80 per cent.

In the above example, we considered two-tailed tests of hypotheses. This however, does not rule out the possibility of conducting one-tiled tests. It all depends upon the type of inquiry that we intend conducting.

Thus, a discussion on the test of hypotheses finally marks the end of our discussion on two variable regression model.

Check Your Progress 2

1) When will you use a log linear regression model?

……

……

……

……

……

2) How do you interpret the estimated slope coefficient of a log linear regression model?

……

……

……

……

……

3) If you want to estimate India's rate of growth of per capita income during the period 1980-2000, what should be the functional form of your regression model?

……

……

……

……

……

……

4) Explain the concept of a classical normal regression model. What is its use?

…….

…….

…….

…….

…….

…….

5) Explain briefly the steps for testing the significance of the regression coefficient.

…….

…….

…….

…….

…….

…….

9.11 LET US SUM UP

Our focus has been exclusively on the two variable regression model. The purpose of this unit has been to specify what is known as the classical regression model in the literature for the population. It is essentially a linear in parameter regression model. This regression model is stochastic in nature in the sense that here the dependent variable is taken as a random variable as against the independent variable being considered as non-random in nature. The element of stochasticity is introduced by including a random error or disturbance term in the model. Our major concern has been to estimate the unknown parameters of the population regression model from the known sample information and tests some hypotheses about these parameters. In this regression model we usually make five assumptions about the disturbance term in order to effectively estimate the parameters by implementing the ordinary least square procedure. The estimates thus obtained satisfy some desirable properties as enunciated in the Gauss-Markov Theorem. In this connection, it may be mentioned here that if some of these assumptions are not fulfilled, some problems regarding the sample estimates of these unknown parameters arise. These problems and their solutions have not been discussed in this unit. We have also considered different functional forms of the regression model that might be relevant to use in different situations. For the purpose of hypothesis testing, we have to extend the assumptions of the classical regression model to include one more assumption about the probability distribution of the disturbance term. The resultant regression model is then known as the classical normal regression model. We have considered the use of this model for the tests of hypotheses.

9.12 EXERCISES

1) The following table presents the data for broad money supply M_3 (Rs. Crore) for India during 1980-2002. Fit in an appropriate regression model and estimate the rate of growth of money supply in India during this period.

Table 9.5: India's Broad Money Supply 1980-2002

Year	M3
1980	50966
1981	59793
1982	68515
1983	80577
1984	95295
1985	111096
1986	130653
1987	153207
1988	179687
1989	213856
1990	249493
1991	292403
1992	344238
1993	399048
1994	478196
1995	552953
1996	642631
1997	752028
1998	901294
1999	1056025
2000	1224092
2001	1420025
2002	1647976

2) Consider the following hypothetical data. Fit in a log linear regression model and interpret the estimated regression coefficient. Show all the calculations.

Dependent Variable (Y)	Independent Variable (X)
189.8	173.3
172.1	165.4
159.1	158.2
135.6	141.7
132.0	141.6
141.8	148.0
153.9	154.4
171.5	163.5
183.0	172.0
173.2	161.5
188.5	168.6
205.5	176.5
236.0	192.4
257.8	205.1
277.5	210.1
291.1	208.8
284.5	202.1
274.0	213.4
279.9	223.6
297.6	228.2
297.7	221.3
328.9	228.8
351.4	239.0
360.4	241.7
378.9	245.2

Adapted from Maddala (2001)

3) In Example 9.4, the sample estimate of India's rate of growth during the period 1960-80 has been obtained to be about 3.4 per cent per year. Test the hypothesis that during the same period the rate of growth has in fact been 4 per cent per year.

9.13 KEY WORDS

Assumptions of the Classical Regression Model:

1) The disturbance term U has a zero mean for all the values of X, i.e., $E(U) = 0$.

2) Variance of U is constant for all the values of X, i.e., $V(X) = \sigma^2$.

3) The disturbance terms for two different values of X are independent i.e., $\text{cov}(U_i, U_j) = 0$, for $i \neq j$.

4) X is non-stochastic.

5) The model is linear.

Classical Normal Regression Model : A linear regression model, in which, in addition to the five assumptions of the classical regression model, one more assumption of the error term being normally distributed is made.

Classical Regression Model : The conventional regression model whose parameters are estimated by the ordinary least square procedure.

Functional Forms of Linear Regression Model : Different kinds of linear regression models.

Gauss-Markov Theorem : Under the assumptions of the classical regression model, among all the linear unbiased estimators, the least square estimators have the minimum variance. In other words, the least square estimators are the best linear unbiased estimators or BLUE.

Goodness of Fit : The ratio of the explained sum of squares to the total sum of squares. Also known as the coefficient of determination.

Hypothesis Testing : Conducting tests regarding hypotheses made about the unknown parameters of the population regression model with the help of the estimated sample regression function.

Linear Regression Model : A regression model that is essentially linear in parameter.

Non-Deterministic Regression Model : A regression model in which the dependent variable is random or probabilistic but the independent variable is non-random.

Parameters : Unknown intercept and the slope coefficient of a population regression model.

Population Regression Function : A linear regression model with a stochastic error term for a given population.

Sample Regression Function : The sample regression function that is fitted into the sample data for estimating the population regression model.

Standard Errors of Estimate : Standard deviations of the estimated sample regression intercept and slope coefficient.

Two Variable Regression Model : A regression model with one explanatory variable.

9.14 SOME USEFUL BOOKS

Gujrati, Damodar N. (2003); *Basic Econometrics*, McGraw-Hill, New York, U.S.A.

Maddala, G.S. (2002): *Introduction to Econometrics*, John Wiley & Sons (Asia), Singapore.

Pidyck, Robert S. & Rubinfeld, Daniel L. (1991); *Econometric Models & Economic Forecasts*, McGraw-Hill, New York, U.S.A.

9.15 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 9.2
2) See Section 9.3
3) See Section 9.4
4) See Section 9.6
5) See Section 9.7

Check Your Progress 2

1) See Section 9.8
2) Measures the elasticity of the dependent variable with respect to the independent variable.
3) See Section 9.8
4) See Section 9.9
5) See Section 9.10

9.16 ANSWERS OR HINTS TO EXERCISES

1) 15.9 per cent
2) Do Yourself
3) Do yourself

UNIT 10 MULTIVARIABLE REGRESSION MODELS

Structure

10.0 Objectives

10.1 Introduction

10.2 Regression Model with Two Explanatory Variables

- 10.2.1 Estimation of Parameters
- 10.2.2 Variance and Standard Errors

10.3 Interpretation of Regression Coefficients

- 10.3.1 Goodness of Fit: Multiple Coefficient of Determination: R^2
- 10.3.2 Analysis of Variance (ANOVA)

10.4 Inclusion and Exclusion of Variables

10.5 Generalisation to n-explainatory Variables

10.6 Problem of Multi-co-linearity

10.7 Problem of Hetero-scedasticity

10.8 Problem of Autocorrelation

10.9 Maximum Likelihood Estimations

10.10 Let Us Sum Up

10.11 Key Words

10.12 Answers or Hints for Check Your Progress Exercises

10.0 OBJECTIVES

This Unit aims at equipping the learners with techniques of multiple regressions which come handy in analysing a number of situations where one is required to study impact of a number of independent variables (factors) on some particular dependent variable. After going through this unit, you will be able to:

- have a fairly good comprehension of technique of multiple regression analysis;
- appreciate the need for extending the basic idea of regression as discussed in Unit 8;
- apply those techniques for more realistic model formulation of explaination of economic phenomena; and
- spot and avoid pit falls which may arise (in quantitative analysis).

on accounts of auto-correlation, hetero-scedasticity and multi-co-linearity on the one hand and mis-specification of the model, (in the sense of including irrelevant as well as excluding relevant variables from the model) on the other.

10.1 INTRODUCTION

In Unit 8, you have studied basics of the Classical Linear Regression Model. You regressed dependent variable Y on independent variable X. In other

words, you tried to explain changes in Y in terms of the changes in X. You had hypothesised a linear relationship between Y and X of this type:

$$Y = \alpha + \beta X$$

Then, you went on to 'estimate' the two constants $\hat{\alpha}$ and $\hat{\beta}$. Once the estimates were ready, you had the estimated model or 'relationship' with you given by

$$\hat{Y} = \hat{\alpha} + \hat{\beta} x$$

In the present unit we are going to extend this type of analysis further to make it more 'realistic' and 'comprehensive'. We will do it in a number of steps: first of all, we shall introduce one more explainatory variable and re-examine' the model. Next step will be to generalise the model to n-explainatory variables. At the next stage, we try to interpret the partial regression co-efficients. After that, we will try to examine how large should the number 'n' be – that is how many explainatory variables must be included in the model and what should be the statistical touch stone for arriving at such a decision. Then, we examine the conditions, or 'assumptions', which make these extensions and generalisations possible. We also examine the possible effects of violation of one or more assumptions. We shall, in particular, pay attention to problems of **multi-co-linearity**, **Hetero-scedasticity** and **auto-correlation**. Then, we take a look at another class of estimates, the Maximum Likelihood Estimators. Finally, we give a brief exposition to some uses of multiple regression analysis in economics.

10.2 REGRESSION MODEL WITH TWO EXPLANATORY VARIABLES

At the outset, please note one small change in notation: in section 10.1 – the introduction to the present unit, we have used the specification of model as

$$Y = \alpha + \beta X$$

However, now onwards, we will write the model in slightly different form:

$Y = \beta_0 + \beta_1 X_1$ This form helps us in presentation in the sense that as and when, we add more explanatory variables X_2, X_3, etc. we simply indicate their coefficients with respective β_1, β_2, β_3, etc.

So the model that we consider in the present section has two explanatory variables X_1, and X_2. The non-stochastic specification will be:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (10.1)$$

The same model in stochastic form will have 'error terms' included:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \mu \qquad (10.2)$$

$$= E(Y) + \mu \qquad (10.3)$$

We can use subscripts't' with Y_0, X_1, X_2 and μ to denote 't^{th}' observation. Thus, the above equations can be written as

$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t}$

$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \mu_t$ and

$Y_t = E(Y_t) + \mu_t$ respectively

In the relationships above, the component ($\beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t}$) is systematic or deterministic component, which is equal to mean value E(Y_t) — or a point on the regression line.

The component μ_t is non-systematic random component determined by factors other than X_1 and X_2.

The model specified by equations 10.1 and 10.2 is a linear regression model which is linear in parameters.

10.2.1 Estimation of Parameters: The Ordinary Least Squares Approach

The Ordinary Least Squares Approach

We collect a sample of observations on Y_1, X1 and X_2 and write down sample regression function

$$Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + e_t \quad (10.4)$$

Note that b_0, b_1, b_2 have replaced the corresponding population parameters β_1, β_2, β_3 here. The random component μ_t replaced by e_t or the sample error term.

So, simply speaking:

b_0 = the estimator for β_0

b_1 = the estimator for β_1

b_2 = the estimator for β_2

We know from the Unit 8 that the ***principle of ordinary least squares*** selects those values for unknowns (b_0, b_1, b_2) such that residual sum of squares (RSS) = $\sum e_t^2$ is minimum. We can develop this idea by rewriting equation 10.4 as

$$e_t = Y_t - b_0 - b_1 X_{1t} - b_2 X_{2t} \quad (10.5)$$

Square on both sides and sum up to get RSS:

$$\Sigma e_t^2 = \Sigma \left[Y_t - b_0 - b_1 X_{1t} - b_2 X_{2t} \right]^2 \quad (10.6)$$

Differentiating the (10.6) w.r.t b_0, b_1, b_2 and equating to zero gives us the three normal equations:

$$\overline{Y}_t = b_0 + b_1 \overline{X}_{1t} + b_2 \overline{X}_{2t} \quad (10.7)$$

$$\Sigma Y_t X_{1t} = b_0 \Sigma X_{1t} + b_1 \Sigma X_{1t}^2 + b_2 \Sigma X_{1t} X_{2t} \quad (10.8)$$

$$\Sigma Y_t X_{2t} = b_0 \Sigma X_{2t} + b_1 \Sigma X_{1t} X_{2t} + b_2 \Sigma X_{2t}^2 \quad (10.9)$$

These three equations give us the following expressions for b_0, b_1, and b_2 respectively:

$$b_0 = \overline{Y} - b_1 \overline{X} \ - b_2 \overline{X_2} \tag{10.10}$$

$$b_1 = \frac{\left(\Sigma y_t x_{1t}\right)\left(\Sigma x_{2t}^2\right) - \left(\Sigma y_t x_{2t}\right)\left(\Sigma x_{1t} x_{2t}\right)}{\left(\Sigma x_{1t}^2\right)\left(\Sigma x_{2t}^2\right) - \left(\Sigma x_{1t} x_{2t}\right)^2} \tag{10.11}$$

and

$$b_2 = \frac{\left(\Sigma y_t x_{2t}\right)\left(\Sigma x_{1t}^2\right) - \left(\Sigma y_t x_{1t}\right)\left(\Sigma x_{1t} x_{2t}\right)}{\left(\Sigma x_{1t}^2\right)\left(\Sigma x_{2t}^2\right) - \left(\Sigma x_{1t} x_{2t}\right)^2} \tag{10.12}$$

The lower case letters denote, as usual, the deviations from the respective means. Thus $y_t = (y_t - \overline{Y}), x_{1t} = (X_{1t} - \overline{X})$ and $x_{2t} = (X_{2t} - \overline{X}_2)$..

10.2.2 Variance and Standard Errors

The variances of OLS estimators given in 10.10, 10.11 and 10.12 above are given in terms of means and deviations of X_1 and X_2 and the variance of population error terms μ_t. Thus,

$$Var(b_0) = \left(\frac{1}{n} + \frac{\overline{X}_1^2 \Sigma x_{2t}^2 + \overline{X}_2^2 \Sigma x_{1t}^2 - 2\overline{X}_1 \overline{X}_2 \Sigma x_{1t} x_{2t}}{\Sigma x_{1t}^2 \Sigma x_{2t}^2 - (\Sigma x_{1t} x_{2t})^2} \right) \sigma^2 \tag{10.13}$$

Standard Error of $b_0 = \sqrt{Var(b_0)}$ (10.14)

$$Var(b_1) = \frac{\Sigma x_{2t}^2}{\Sigma x_{1t}^2 \Sigma x_{2t}^2 - (\Sigma x_{1t} x_{2t})^2} . \sigma^2 \tag{10.15}$$

$$SE(b_1) = \sqrt{Var(b_1)} \tag{10.16}$$

and

$$Var(b_2) = \frac{\Sigma x_{1t}^2}{\Sigma x_{1t}^2 \Sigma x_{2t}^2 - (\Sigma x_{1t} x_{2t})^2} . \sigma^2 \tag{10.17}$$

$$SE(b_2) = \sqrt{Var(b_2)} \tag{10.18}$$

When σ^2 is unknown and its t unbiased OLS estimator is obtained, we find that

$$\hat{\sigma}^2 = \frac{\Sigma e_i^2}{n-3} \tag{10.19}$$

The denominator of 10.19 shows the degrees of freedom. In a sample of size 'n' we exhaust 3 degrees of freedom in estimating b_0, b_1, and b_2. Note that this reasoning is quiet general, in the sense that if out of a sample of size 'n', we

estimate 'k' parameters, remaining degrees of freedom will be n — k. We call $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$, the standard error of estimate/regression.

The residual sum of squares in 10.19 is $\sum e_t^2 = \sum (Y_t - \hat{Y}_t)^2$. This expression is equivalent to the following.

$$\Sigma e_t^2 = \Sigma y_t^2 - b_1 \Sigma y_1 x_{1t} - b_2 \Sigma y_t x_{2t} \qquad (10.20$$

10.3 INTERPRETATION OF REGRESSION COEFFICIENTS

Mathematically, b_1 and b_2 represent the partial slopes of regression plan with respect to X_1 and X_2 respectively. In other words, b_1 shows the rate of change in Y as X_1 alone undergoes a unit change, keeping all other thing, constant. Similarly, b_2 represents rate of change of Y as X_2 alone changes by a unit while other things are held constant.

Note: *This interpretation is quite similar to what you have been doing in economic theory. Recall the law of demand: quantity demanded of a commodity varied inversely with its price, 'ceteris paribus' that is when other things were kept 'constant'. What were those other things? Those were prices of complements and substitutes, income of the consumer, and tastes/likings/ disliking etc.*

10.3.1 Goodness of Fit: Multiple Coefficient of Determination:R^2

Recall that is case of a single independent variable, r^2 measured goodness of fit of the fitted sample regression line. It gave you the percentage total variation in dependent variable Y which has been explained by the single explanatory variable X. Correspondingly, when we have two explanatory variables X_1 and X_2, we would like to know the proportion of total variation in $Y = \sum y_t^2$ explained by X_1 and X_2 jointly. This information is conveyed by R^2 — the multiple coefficient of determination. Thus,

$$R^2 = \frac{ESS}{TSS}$$

When ESS is explained sum of squares and TSS is total sum of squares. You are familiar with the relationship between TSS, ESS and RSS from Unit 8.

You know that $ESS = b_1 \sum y_t x_{1t} + b_2 \sum y_t x_{2t}$

and $RSS = \sum y_t^2 - b_1 \sum y_t x_{1t} - b_2 \sum y_t x_{2t}$

Therefore R^2 can be computed as

$$R^2 = \frac{b_1 \sum y_t x_{1t} + b_2 \sum y_t x_{2t}}{\sum y_t^2} \qquad (10.21)$$

You can get the quantities in the above formula from your computation sheet for the normal equations. The R^2 lies between 0 and 1. Closer it is to one;

better is the fit – implying estimated regression line is capable of explaining greater proportions of variation in Y. The positive square root of R^2 is called coefficient of multiple correlation.

10.3.2 Analysis of Variance (ANOVA)

We know the relationship:

TSS = ESS +RSS

This is equivalent to saying:

$$\underset{(TSS)}{\Sigma y_t^2} = \underset{(ESS)}{b_1 \Sigma y_t x_{1t} + b_2 \Sigma y_t x_{2t}} + \underset{(RSS)}{\Sigma e_t^2}$$

A study of the components of TSS is called analysis of variance in the context of regression. You know that every sum of squares has some degrees of freedom (d*f*) associated with it. In our above 2-explainatory variable case, the degrees of freedom will be

TSS = n–1

RSS = n–3

ESS = 2

We can put this information in form of a table:

Table 10.1: ANOVA Table for 2-Explainatory Variable Regression

Source of Variation	Sum of Squares (SS) *df*	Mean S.S = $\frac{\delta S}{df}$
Due to regression (ESS)	$b_1 \Sigma y_t x_t + b_2 \Sigma y_t x_{2t}$ 2	$\frac{ESS}{2}$
Due to residuals	Σe_t^2 n-3	$\frac{\Sigma e_t^2}{n-3}$
TSS	Σy_1^2 n-1	

One may be interested in testing a null hypothesis H_0: $B_1 = B_2=0$. In such a case we find that

$\frac{ESS/df}{RSS/df}$ that is ratio of the variance explained by X_1 and X_2 to unexplained variance follows F distribution with 2 and n-3 degrees of freedom. In general, if regression equation estimates 'k' parameters including intercept, than F ratio has (k-1) d*f* in numerator and (n-k) d*f* in the denominator.

***Interpretation*:** Larger the variance explained by fitted regression line, larger the numerator will be in relation to denominator. Thus a larger F value is evidence **against** the 'truthfulness' of H_0: $B_1 = B_2=0$. That is, F values larger than one will indicate that hypothesis of both the variables X_1 and X_2 having no effect on Y cannot sustain.

We can also express F values in terms of R^2.

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Check Your Progress 1

1) What is multiple regression? Explain with an example.

……

……

……

……

……

2) State the assumptions of classical linear multiple regression model.

……

……

……

……

……

3) What do you mean by linearity of regression model? Explain?

……

……

……

……

……

4) How do you interpret coefficients of multiple regression model?

……

……

……

……

……

……

10.4 INCLUSION AND EXCLUSION OF EXPLANATORY VARIABLES

The Adjusted $R^2 : \overline{R^2}$

It has been noted that as we add more and more explanatory variables X_s , the explained sum of squares keeps on rising. Thus, R^2 goes on increasing as we increase X_s, the explained sum of squares keeps on rising. Thus, R^2 goes on increasing as we increase X_s. But, notice that adding each additional variable 'eats up' one degree of freedom and our definition of R^2 makes no allowance for loss of degrees of freedom. Hence, the thinking that you can improve the goodness of fit by suitably increasing the number of variables may not be justified — We know that TSS always has (n-1) degrees of freedom. Therefore, comparing two regression models with same dependent variable but differencing number of independent variables will not be justified. Therefore, we must adjust our measure of goodness of fit for degrees of freedom. This measure is called adjusted R^2, denoted by $\overline{R^2}$. It can be derived from R^2 as follows:

$$\overline{R}^2 = 1-(1-R^2)\frac{(n-1)}{(n-k)} \qquad (10.22)$$

Therefore, it is recommended that, one must include new variables only if (upon inclusion) $\overline{R^2}$ increases and not otherwise. A general guide is provided by 't' statistic, if absolute value of the co-efficient of added variable is greater than one, retain it (Note that 't' value is calculated under the hypothesis that population value of that co-efficient is zero).

10.5 GENERALISATION TO N-EXPLANATORY VARIABLES

In general, our regression model may have a large number of independent variables. Each of those variables can, on priority grounds, be expected to have some influence over the 'dependent' or 'explained' variable. Consider a very simple example. What can be determinants of demand for potatoes in a vegetables market? One obvious choice will be the price of potatoes. What else can affect the quantity demanded? Could it be availability of vegetables which can be paired off with potatoes? In that case, prior of a large number of vegetables which are cooked along with potatoes will become 'relevant explanatory variables'. You cannot ignore income of the community that patronizes the particular market. Needless to say, the dietary preferences of members of the households can also affect the demand and so on. In the next Unit, we shall discuss techniques which help us restrict the analysis to a selected 'few variables, though theoretic considerations may find a huge number of them to be 'useful' and 'powerful' determinants. In fact, in economic theory, we usually append the phrase **Ceteris paribus**, with many a statements. This phrase means keeping all other things constant. That means, we may focus on impact of only a few selected variables on the dependent variable while assuming that all other variables remain 'unchanged' during the period of analysis. This assumption may not hold so, we have to juggle the need to include more and more variables in our model with the 'gains' in

explainatory power of the model. We have developed, in previous section (10.4) a working touchstone for inclusion of more variables in terms of improvement in $\overline{R^2}$ and have tried to give it 'practical' shape in form of magnitude of 't' values of the relevant slope parameters.

With these considerations in mind we can generalise the linear regression model as follows:

We hypothesize that in population, the dependent variable Y depends upon n explainatory variables, X_1, X_2, ……..X_n. We also assume that the relationship is linear in parameters. Three more assumptions are made and they have very significant bearing on the analysis. These are:

a) Absence of Multi-co-linearity;

b) Absence of Hetero-scedasticity; and

c) Absence of Autocorrelation

We will discuss the complications, which arise because of violations of these assumptions in section 10.6, 10.7 and 10.8 respectively. So our Classical Linear General Regression Model is :

$$Y_t = \beta_0 + \beta_1 X_{1t} + \cdots\cdots\cdots\cdots + \beta_n X_{nt}$$

or

where it is understood that Y and each X will have a large numbers of values t=1…….N, forming (n+1) 'tuples' (Y_{11} X_{11}, X_{21}, ……..X_{n1}).

We can simply write

$$Y = \beta_1 X_1 + \beta_2 X_2 \ldots\ldots\ldots\ldots\ldots\ldots \beta_n X_n + \mu \qquad (10.23)$$

in Matrix equation form, we get

$$Y = X\beta + U \qquad (10.24)$$

Where $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$ and $U = \begin{bmatrix} U_1 \\ \vdots \\ U_n \end{bmatrix}$

also $X = \begin{bmatrix} X_{11}, X_{21}, X_{31}, \cdots\cdots\cdots\cdots X_{k1} \\ X_{12}, X_{22}, X_{32}, \cdots\cdots\cdots\cdots X_{k2} \\ \vdots \\ \vdots \\ X_{1n}, X_{2n}, X_{3n}, \cdots\cdots\cdots\cdots X_{kn} \end{bmatrix}$

We assume that (1) expected values of error terms are equal to zero; that is $E(u_i) = 0$ for all '*i*'. In matrix notation

$$E(u) = \begin{bmatrix} \mathrm{E}(u_i) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

2) The error terms are not correlated with one another and they all have same variance for σ^2 all sets values of the variables X. That is,

$E(u_i u_j) = 0; \quad \forall i \neq j$ and

$E(u_1^2) = \sigma^2 \forall i$

in matrix notation:

$$E\left[UU'\right] = \begin{bmatrix} E(u_1^2), E(u_1u_2) \cdots\cdots\cdots\cdots E(u_1u_n) \\ E(u_1u_2), E(u_2^2) \cdots\cdots\cdots\cdots E(u_2u_n) \\ \vdots \\ E(u_1u_n), E(u_2u_n) \cdots\cdots\cdots\cdots E(u_n^2) \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & \sigma & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

3) Number X_{1i}……..X_k are all real numbers and are free of random errors or disturbances.

4) The matrix X has been linearly independent columns. It implies that number of observations exceeds number of co-efficient to be estimated. It also implies there exists no exact linear relationships between any of the X variables.

Note: Assumptions that $E(U_iU_j) = 0$ means that error terms are not correlated. The implication of **diagonal** terms in matrix E(UU′) being all equal to c^2 is that all error terms have same variance, c^2. This is also called assumption of homo-scedasticity. The last assumption implies absence of multi-co-linearity. We can write the regression relation for the sample as:

e = Y – Xb

where *e, Y, X* and *b* are appropriate matrices.

Sum of squared residuals will be

$$\phi = \sum e_1^2 = \sum (Y_i - b_1 X_{1i} \cdots\cdots\cdots + b_k X_{ki})^2 \qquad (10.25)$$

$$= e'e = \left[Y - Xb\right]' \left[Y - Xb\right]$$

$$= Y^1 Y - 2b'X'Y + b'X'Xb$$

Note: b' X'Y is scalar and is therefore equal to its transpose Y′ Xb.

by equating 1st order partials of ϕ w.r.t each b_i, to zero, we get *k* normal equations. This set of equations in matrix form is:

$$\frac{\partial \phi}{\partial b} = -2X'Y + 2X'Xb = 0 \quad (10.26)$$

$$X'Xb = X'Y \quad (10.27)$$

when X has rank equal to *k,* the normal equation 10.27 will have a unique solution and least squares estimator *b* is equal to:

$$b = [X'X]^{-1}[X'Y] \quad (10.28)$$

We have assumed b to be estimator for β and thus E(b) = β , therefore we can rewrite 10.28 as

$$b = [X'X]^{-1} X''[X\beta + u]$$
$$= [X'X]^{-1} X'X\beta + [X'X]^{-1} X'U$$
$$= \beta + [X^1 X]^1 X^1 U$$
$$\therefore E(b) = E(\beta) + E\left[[X^1 X]^{-1} X^1 E(\mu)\right] = E(B) + [X'X]X'E(\mu)$$
$$= \beta$$

Variance of $b = \sigma^2 (X'X)^{-1}$

Notes

1) In this course our objective is simply to introduce the concepts. Those who plan to pursue the concepts at much more rigorous level can study our course on Econometric Methods (MEC) – included as optional course of M.A.(Economics) Programme.

2) The other ideas regarding coefficient of determination R^2 and adjusted R^2 remain the same as they were developed for two explainatory variable case.

Now we can safely turn to discussions about non-satisfaction or violation of assumptions.

10.6 THE PROBLEM OF MULTI-CO-LINEARITY

Many a times our X variables may be found to have some other linear relationships among themselves. This vitiates our classical regression model.

Let us illustrate it with help of our 2 explainatory variable model.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + U_1$$

Let us give specific names to variables, say, X_1 is price of commodity Y and X_2 is family income. We expect β_1 to be negative and β_2 to be positive. Now we go one step further. Let Y be demand for milk, X_1 be price of milk and let the family wise demand for milk is being estimated for a family, which produces and sells milk! Clearly, larger the value of X_1 higher the magnitude of X_2 will be.

In such situations, the estimation of co-efficient will not be possible. Recall, we wanted variables X in our matrix equations to be linearly independent. If that conditions is not satisfied X matrix becomes singular, that is its determinant tends to vanish. Thus, there will be **no solution** to the normal equation **10.26 (or 10.27).**

However, if co-linearity not perfect, we can still get OLS estimates and they remain best linear unbiased estimates (BLUE) – though one or more partial regression co-efficient may turn out to be individually insignificant.

Not only this, the OLS estimates still retain property of minimum variance. Further, it is found that multi co-linearity is essentially a sample regression problem. The X variables may not be linearly related in population but some of our suppositions while drawing a sample may create a situation of multiple linear relations.

The practical consequences of multi-co-linearity. Gujarali, (D.N.) has listed the following consequences of multiplicity of linear relationships:

1) Large variances /SEs of OLS estimates
2) Wider confidence intervals
3) Insignificant 't' ratios for β parameters
4) A high R^2 despite few significant t values
5) Instability of OLS estimators: The estimators and their standard errors (SEs) become very sensitive to small changes in data.
6) Sometimes, even signs of some of the regressions may turn out to be theoretically unacceptable like a rise in income having negative impact on demand for milk.
7) When many regressions have insignificant coefficients, their individual contributions to the explained sum of squares cannot be assessed properly.

The multi-co-linearity can be detected by:

1) high R^2 but few significant 't' ratios,
2) high pair wise correlation between explainatory variables. One can try partial correlations, subsidiary or auxiliary regressions as well. But each such technique increases burden of calculations.

Check Your Progress 2

1) When do you decide to drop a variable from the regression equation? Why?

……

……

……

……

2) When do you include more variable(s) into your model? Why?

……

……

……

……

3) "Inclusion of more variables always increases R^2 the goodness of fit. So to make a regression model 'good' what we need to do is simply increase the number of explanatory variables". Do you agree/disagree with this statement? Give reasons.

……

……

……

……

……

4) What is multi-co-linearity? What are its consequences?

……

……

……

……

10.7 PROBLEM OF HETERO-SCEDASTICITY

The Classical Linear Regression Model has a significant underlying assumption in homo scedasticity, that is, all the error terms are identically distributed with mean equal to zero and standard deviation equal to σ (or variance equal to σ2). σ2 What happens when this second part of assumption regarding distribution of variance does not hold? As a result, in symbolic terms $E(u_i)^2 = \sigma_i^2$, that is, if the expectation of squared errors is no longer equal to σ^2 — each error term has its own σ^2, or variance varies from observation to observation.

It has been observed that usually time series data does not suffer from this problem of hetero scedasticity but in cross-section data, the problem may assume serious dimensions. **The consequences of hetero scedasticity:** If the assumption of homoscedasticity does not hold, we observe the following impact on OLS estimators.

1) They are still linear
2) They are still unbiased
3) But they no longer have minimum variance – that is we cannot call them BLUE: the Best Linear Unbiased Estimators. In fact, this point is relevant both for small as well as large samples.
4) Reason for this problem hinted at in (3) above is that generally, OLS estimators have some bias built into their formulae. We try to rectify that making use of degrees of freedom.

   For instance $\hat{c}^2$, (the estimator for true population σ^2) given by $\sum e_1^2 / df$ no longer remains unbiased. And this very $\hat{c}^2$ enters into calculation of standard errors of OLS estimates.

5) Since, estimates of standard errors are themselves no longer reliable, we may end up drawing wrong conclusions using conventional reasoning based on procedures for testing the hypothesis.

How to detect Hetero scedasticity

In applied regression analysis, plotting the residual terms can give us important clues about whether or not one or more assumptions underlying our regression model hold. The pattern exhibited by e_i^2 plotted against the concerned variable can provide important clue. If no pattern is detected – homoscedasticity holds or hetero scedasticity is absent. On the other hand, if errors form a pattern with variable — expanding, increasing linear or changing is some non-linear manner thus hetero scedasticity is certainly present.

Some tests have been designed to detect presence of Hetero scedasticity, using various statistical techniques. Prominent ones are: Park Test, Glejser Test, Whites General Test, Spearman's Rank correlation Test, Goldfield - Quadnt Test etc. But in this unit, the limitation of space does not permit us to go into their details. We are forced to refer the learners again the course on Econometric Method for details in this regard.

How to tackle the hetero scedasticity?

Our ability to tackle the problem will depend upon the assumptions. We can really make about error variance. Thus, the following situations may emerge

i) When σ_1^2 is known

Here the CLRM

$Y_i = \beta_0 + \beta_1 X_1 + u_i$ can be transformed, dividing each value by corresponding σ_1 thus,

$$\frac{Y_i}{\sigma_1} = \beta_0\left(\frac{1}{\sigma_1}\right) + \beta_1\left(\frac{X_i}{\sigma_i}\right) + \frac{U_i}{\sigma_i}$$

This effectively transforms error terms to $U_i = \frac{U_i}{\sigma_1}$ which is homo-scedastic and therefore, the OLS estimators will be free of disability caused by hetero scedasticity. The estimates of β_0 and β_1 in this situation are called **Weighted Least Squares Estimators (WLSEs)**.

ii) **When σ^2 is unknown:** we make some further assumptions about error variance:

iia) Error variance proportional to the X_i s. Here, the Square Root transformation is enough. We divide on both sides by $\sqrt{X_i}$. Thus, our regression line looks like:

$$\frac{Y_i}{\sqrt{X_i}} = \frac{\beta_0}{\sqrt{X_i}} + \beta_1\frac{X_i}{\sqrt{X_i}} + \frac{U_i}{\sqrt{X_i}}$$

$$= \beta_0 \frac{1}{\sqrt{X_i}} + \beta_1 \sqrt{X_i} + U_i$$

Here $U_i = \frac{U_i}{\sqrt{X_i}}$ and this is sufficient to address the problem.

iib) Error Variance proportional to X_1^2. Here, instead of division by $\sqrt{X_i}$, we divide by X_i on both the sides and estimate

$$\frac{Y_i}{X_i} = \beta_0 \frac{1}{X_i} + \beta_1 + \frac{U_i}{X_i}$$

$$= \beta_0 \frac{1}{X_i} + \beta_1 + U_i$$

The error term will be $U_i = \frac{U_i}{X_i}$ and this will be free of hetero scedasticity, facilitating use of CLS techniques.

iii) ***Respecification of Model*:** Assigning a different functional form to the model, in place of speculating about the nature of variance may be found expedient. We can estimate this model:

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + U_i$$

This loglinear model is usually adequate to address our concerns.

10.8 PROBLEM OF AUTOCORRELATION

The classical regression model also assumes that disturbance terms U_is do not have any serial correlation. But, in many situations this assumption may not hold. The consequences of the presence of serial or auto correlation are similar to those of hetero scedasticity: the OLS are no longer BLUE.

Symbolically **no** autocorrelation means $E(U_i\ U_j) = 0$ when $i \neq j$.
Autocorrelation can arise in economic data on account of many factors:

i) there may be various reasons which cause cyclical up and down swings in economic time series. These tendencies continue till something happens which reverse them. This is called **inertia**.

ii) Misspecification of model in the form of under specification can be a cause of autocorrelation: you have fewer X variables in the model, leaving out rather large systematic components to be clubbed with errors.

iii) Cob-web phenomenon is another factor which may create this problem in certain types of economic time series (especially agricultural output etc.).

iv) ***Polishing of data*** – like adding monthly data to make quarterly or quarterly to **make** half yearly series etc. can also be responsible for the autocorrelation – as the averaging involved dampens the fluctuations of the original data.

The consequences of autocorrelation are not different from those of hetero scedasticity listed in 10.7 above. Here too OLS estimators are biased or are not BLUE, *t* & F tests are no longer reliable. Therefore, computed value of R^2 is not reliable estimate of true goodness of fit.

There are many tests for detecting autocorrelation – varying from visual inspection of error plots, the Runs Test, Swed-Eisenhart critical runs test. But most commonly used in Durbin-Watson *d* test defined as

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$$

However, again, we are holding back information on practical detections and avoidance of problem of autocorrelation for the reasons of limitation of space here.

10.9 THE MAXIMUM LIKELIHOOD ESTIMATIONS

Sometimes another class of estimators is used in place of OLS. This class of estimators is called ***maximum likelihood estimators*** (MLEs). The MLEs possess some stronger theoretical properties. But it also requires a stronger assumption about distribution of error terms. Moreover, when errors follow normal distribution, we find that OLS and MLE methods give identical estimates of β parameters, both in simple and in multiple regressions. However, MLE estimate of σ^2 is biased. Hence, if one uses assumption of normal distribution of U_is and persists with OLS, one does not miss out on any thing that may be advantageous in MLE.

The Method: Simple Regression Illustration

In the model $Y_1 = \beta_0 + \beta_1 X_1 + U_i$, the Y_1 is normally and independently distributed with means $\beta_0 + \beta_1 X_i$ and variance σ^2. Therefore, the joint probability density function of

$Y_1, Y_2, Y_3 \ldots\ldots\ldots\ldots\ldots\ldots Y_n$ with above mean and variance will be

$$f\left(Y_1 \ldots\ldots\ldots\ldots Y_n / \beta_0 + \beta_1 X_i, \sigma^2\right) \qquad \text{ML - 1}$$

But given the independence of Y_s, this function can be written as product of '*n*' individual density functions, or

$$f\left(Y_1 \ldots\ldots\ldots\ldots Y_n / \beta_0 + \beta_1 X_i + U_i, \sigma^2\right) \qquad \text{ML-2}$$

$$= \pi\left[Y_i / \beta_0 + \beta_1 X_i + U_i, \sigma^2\right]$$

$$\text{where} \quad f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} . e^{\left[-Y_2 \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}\right]} \qquad \text{ML-3}$$

Putting this value for each Y_i, in ML-2. We get the likelihood function (LF):

$$LF(\beta_0, \beta_1, \sigma^2) = \frac{1}{\sigma^n \left(\sqrt{2\pi}\right)^n} . e^{\left[-\frac{1}{2}\sum \frac{(Y_i - \beta_0 - \beta_1 - U_i)^2}{\sigma^2}\right]} \qquad \text{ML - 4}$$

The method of maximum likelihood estimation is nothing but maximization of the (LF) given in ML-4 above. We can differentiate logarithms of (LF) with respect to β_0, β_1 and σ^2 and equate the respective partials to zero to get the requisite estimation relations. So as

$$\ln(LF) = n \Big/ n\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum \frac{(Y_i - \beta_0 - \beta_1 X_1 - U_i)^2}{\sigma^2} \qquad \text{ML - 5}$$

therefore

$$\frac{\partial \ln(\text{LF})}{\partial \beta_0} = \frac{-1}{\sigma^2}\sum (Y_1 - \beta_0 - \beta_1 X_i)(-1) = 0 \qquad \text{ML - 6}$$

$$\frac{\partial \ln(\text{LF})}{\partial \beta_1} = \frac{-1}{\sigma^2}\sum (Y_1 - \beta_0 - \beta_1 X_1)(-X_i) = 0 \qquad \text{ML - 7}$$

$$\frac{\partial \ln(\text{LF})}{\partial c^2} = \frac{-n}{2c^2} + \frac{1}{2c^4}\sum (Y_1 - \beta_0 - \beta_1 X_i)^2 = 0 \qquad \text{ML - 8}$$

Simplification of ML-6 and ML-7 give us

$$\Sigma Y_i = n\beta_0 + \beta_1 \Sigma X_i \qquad \text{ML-8a}$$

$$\Sigma X_i Y_i = \beta_0 \Sigma X_i + \beta_1 \Sigma X_i^2 \qquad \text{ML-9}$$

Which are same as OLS normal equations.

Substituting values of ML (=OLS) estimates obtained by simultaneously solving ML-8 and ML-9 into ML-7 gives us

$$\sigma^2 = \frac{1}{n}\sum (Y_i - \beta_0 - \beta_1 X_i)^2$$

$$= \frac{1}{n}\sum U_i^2 \qquad \text{ML - 10}$$

Expectation of $\tilde{\sigma}^2, E(\tilde{\sigma}^2) = \frac{1}{n} E(\sum U_i^2)$

$$= \left(\frac{n-2}{n}\right)\sigma^2$$

$$\sigma^2 - \frac{2}{n}\sigma^2$$

or $\tilde{c}^2$ is biased downwards.

Multiple Regression: An example: The following regression were run on SPSS. Edited summary of results is as follows:

India's Imports, Exports and Foreign Investment Inflows

All figures are in Millions of US dollars

Year	var00001	var00002	var00003
1991-1992	19411.00	17865.00	133.00
1992-1993	21882.00	18537.00	559.00
1993-1994	23306.00	22238.00	4153.00
1994-1995	28654.00	26330.00	5138.00
1995-1996	36678.00	31797.00	4892.00
1996-1997	39133.00	33470.00	6133.00
1997-1998	41484.00	35006.00	5385.00
1998-1999	42389.00	33218.00	2401.00
1999-2000	49671.00	36822.00	5181.00
2000-2001	50536.00	44560.00	5862.00
2001-2002	51413.00	43827.00	6693.00
2002-2003	61412.00	52719.00	4555.00

var00001=Imports

var00002=Exports

var00003=Foreign Investment Inflow

Regression of Imports on Exports and Foreign Investment Inflow

R	R^2	$\overline{R}^2$	Standard error of Estimate
0.983	0.966	0.958	2726.228

Coefficients

	B	Standard Error	t
Constant	-1655.737	2658.578	-0.623
var00002	1.263	0.104	12.192
var00003	-0.288	0.522	-0.552

Coefficients Correlation

Correlation		Var00003	Var00002
	Var00003	1.0000	-0.670
	Var00002	-0.670	1.0000
Covariance	Var00003	0.272	-3.625E-02
	Var00002	-3.625E-02	1.073E-02

ANOVA

Sum of Sqrs.	*df*	Mean sqrs.	F
Regression1.9E+09	2	9.4E+08	127.097
Residual 6.7E+07	9	7432320	
Total 2.0E+09	11		

Regression of Foreign Investment Inflow on Exports and Imports

R	R^2	$\overline{R}^{-2}$	Standard error of Estimate
0.684	0.468	0.349	1712.334

Coefficients

	B	Standard Error	t
Constant	-321.545	1702.075	-0.189
var00001	-0.114	0.206	-0.552
var00002	0.272	0.257	1.060

Coefficients Correlation

Correlation		**Var00002**	**Var00001**
	Var00002	1.0000	-0.982
	Var00001	-0.982	1.0000
Covariance	Var00002	6.590E-02	-5.92E-02
	Var00001	-5.192E-02	4.24E-02

ANOVA

	Sum of Sqrs.	*df*	Mean sqrs.	F
Regression	2.3E+07	2	1.2E+07	3.952
Residual	2.6E+07	9	2932089	
Total	5.0E+07	11		

The above regressions are based on certain 'a prior' expectations. We expect that imports into India during the period 1991-92 to 2002-03 depend upon exports from India and foreign investment inflows into the country. The idea is that exports pay for imports and foreign investment inflow 'facilitates' the country to import more. Our results confirm our theoretic expectations. Regression on imports on exports and investment inflows shows that R^2 = 0.966 and $\overline{R^2} = 0.958$. This shows that model has high explanatory power. It explains over 95 per cent of the variation. Typical computer output gives information on coefficients, correlation and co-variances, co-linearity diagnostics, residual analysis of variance etc. We have run another regression too – the results of which are reported above. This is regression of foreign investment

inflows on exports and imports. Theoretic expectation was that investment inflows are determined by magnitudes of exports and imports. However, this model gives rather disappointing results. $R^2 = 0.468$ and $\overline{R^2} = 0.349$ only. In other words, our model explains less than 35 per cent of the variations. It is not desirable to persist with it. The same is reflected in low values of β_1 & β_2 parameters and their high standard errors.

Check Your Progress 3

1) What is hetero-scedasticity? What are its consequences?

……

……

……

……

……

……

2) What is auto-correlation? When does it arise? What are its consequences?

……

……

……

……

……

……

3) What are maximum likelihood estimations? Why do we persist with least squares estimators most of the time.

……

……

……

……

……

……

10.10 LET US SUM UP

We began with extension of the simple linear regression model to incorporate one more explanatory variable. Our next step was to interpret the coefficients of partial regression. Afterwards, we tried to design statistical touch stone for

inclusion of more variables into the model. This also gave us some guidelines to drop the 'undesirable' variable as well. We then moved on to more tedious and mathematically more demanding extension to 'n' variable model. We then attempted to analyse the effects of multi-co-linearity, hetero scedasticity and auto-correlation respectively. We also made comments on techniques of identification of these problems and some strategies to get rid of the problems as well. Further, we discussed the class of estimators called maximum likelihood estimators (MLEs). We made comparison between MLEs and Least square estimators as well. Finally, we ran some regressions on SPSS package between India's imports, exports and foreign investment inflows to illustrate some of the points, which arose in course of our discussions. However, we must again draw the attention of learners to the fact that limitations of space did not permit us to go in for an exhaustive discussion of the concepts touched upon. Those who want to have detailed knowledge about the issues and concepts involved in multiple regression models are advised to go through our optional course MECE-001 Econometric Methods.

10.11 KEY WORDS

Multiple Regressions : Regression of one dependent variable on more than one independent variables.

Partial regression Coefficients : Coefficients of individual predictors in multiple regressions.

Coefficient of multiple determination — R^2 : Ratio of variation explained or accounted for by the regression model to the total variation.

Adjusted coefficient of multiple determination — $\overline{R^2}$: It is coefficient of determination adjusted for the loss of degrees of freedom. It helps us against increasing $\overline{R^2}$ by incorporating unnecessary explanatory variables.

Multi-co-linearity : Existence of linear relationships between explanatory variables in addition to one between dependent variable and them.

Hetero scedasticity : Differences between variances of the error terms. This problem is more serious in cross-sectional data.

Auto-co-relation : Correlations between the errors and their previous values. This problem is encountered very often when dealing with time series data.

10.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

1) Read Section 10.2

2) Read Section 10.2

3) Read Section 10.2

4) Read Section 10.3
5) Read Section 10.4
6) Read Section 10.4
7) Read Section 10.4
8) Read Section 10.6
9) Read Section 10.7
10) Read Section 10.8
11) Read Section 10.9

UNIT 11 MEASURES OF INEQUALITY

Structure

11.0 Objectives

11.1 Introduction

11.2 Positive Measures

- 11.2.1 Relative Range
- 11.2.2 Relative Inter-Quartile Range
- 11.2.3 Relative Standard Variation
- 11.2.4 Standard Deviation of Logarithms
- 11.2.5 Champernowne Index
- 11.2.6 Hirschman-Herfindahl Indices
- 11.2.7 Kolm's Index

11.3 Gini Index

- 11.3.1 Gini as a Measure of Dispersion
- 11.3.2 Simple Computational Device

11.4 Lorenz Curve

- 11.4.1 Geometrical Definition
- 11.4.2 Properties of Lorenz Curve
- 11.4.3 A Measure Based on Area
- 11.4.4 A Measure Based on Length

11.5 Normative Measures

- 11.5.1 Dalton Index
- 11.5.2 Atkinson Index
- 11.5.3 Sen Index
- 11.5.4 Theil Entropy Index

11.6 Let Us Sum Up

11.7 Key Words

11.8 Exercises

11.9 Some Useful Books

11.10 Answers or Hints to Check Your Progress Exercises

11.0 OBJECTIVES

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

- explain the various positive measures of inequality;
- discuss the computational device for construction of Gini index; and
- describe the normative measures of inequality propounded by Dalton, Atiknson, Sen and Theil.

11.1 INTRODUCTION

Improvement in well-being of the poor has been one of the important goals of economic policy and to a significant extent it is determined by the growth and distribution of its income. Distribution patterns have an important bearing on the relationship between average income and poverty levels. Extreme inequalities are economically wasteful. Further, income inequalities also interact with other life-chance inequalities. Hence reducing inequalities has become priority of public policy.

It therefore, becomes pertinent to measure income inequalities. Various measures have been developed over a period of time to study the level of inequalities in different situations. Broadly these measures can be put under two categories (i) positive measures, and (ii) normative measures. The measures which capture the inequality of income without value judgment about social well-being are known as positive measures. Range, quartile range, standard deviation, Gini ratio, Lorenz curve etc. are positive measures of inequality. On the other hand, the measures that essentially involve value judgement about social welfare are called normative measures. The index propounded by Dalton, Atkinson, Sen, Theil and Kakwani are normative measures. We shall discuss all these measures one by one in this unit. Gini Coefficient of Inequality and Lorenz Curve will receive particular attention, as they are very popular in literature.

11.2 POSITIVE MEASURES

If all values in a distribution are not equal, which means that there is dispersion in the distribution, hence, there exists inequality in the distribution. If a measure is developed to capture this non-equality in values without giving explicit consideration to its consequences with respect to social well-being or economic significance in a particular context, the measure is known as **positive**. It means that the measure is bothering about the fact whether it is measuring inequality of lengths of iron nails or incomes of wage earners in a village. Nevertheless, many of them are standard statistical measures and their social implications and/or social consequences can still be studied. Some of these measures can be arrived from normative approach as well.

Let us consider an income distribution x_i i=1,2,…,N over N persons and with mean income μ. Let the relative share of total income with person i be designated as q_i, which is naturally given by $x_i/N\mu$. The cumulative share of total income with the persons not having more than x_i income can be given as Q_i. However, when we have an income having frequency more than one, the proportion of people with income x_i can be denoted by p_i and the cumulative proportion of people with income no less than x_i as P_i. In this case, obviously the relative share of total income would be given by $f_i x_i / N\mu$ where f_i denotes the frequency of occurrence of income x_i and $N=\ f_i$. It will not be necessary to use two different subscripts for distinguishing the two cases for the purpose as the context would make it clear whether subscript i stands for a person with income x_i or for the group of persons with income x_i.

We have implicitly assumed that the data are arranged in the increasing (non-decreasing) order by magnitude of income so that symbolic representation is easy. The same could be accomplished by decreasing (non-increasing) order.

We intend to cover important measures along with their variants in this Unit. We shall also, albeit briefly, discuss their properties and weaknesses.

To recapitulate, we recount what we said above in a formal way

$$q_i = \frac{x_i}{N\mu} or \frac{f_i x_i}{N\mu}$$

$$Q_i = \sum_{i-1}^{i} q_{i,}$$

$$p_i = \frac{1}{N} or \frac{f_i}{N}$$

$$P_i = \sum_{i-1}^{i} p_{i,}$$

11.2.1 Relative Range

A measure of relative dispersion can be taken a measure of inequality. It is defined as the relative range by

$$RR_1 = \frac{Max_i x_i - Min_i x_i}{\mu} \qquad \text{(RR.1)}$$

that is, the relative difference between the highest income and the lowest income. If income is equally distributed, then $RR_1 = 0$ and if one person received all the income, then RR_1 is maximum. If one wants to make the index lie in the interval between 0 and 1, one can define it as

$$RR_2 = \frac{Max_i x_i - Min_i x_i}{N\mu} \qquad \text{(RR.2)}$$

which means it is the gap between the maximum share and the minimum share. That is,

$$RR_2 = Max_i q_i - Min_i q_i \qquad \text{(RR.3)}$$

Though Cowell has suggested division of range by $Min_i\ x_{i,}$, which does not serve, in our view, any purpose. Two other normalization or standardization procedures that make it unit-free and contain it in (0,1) interval are suggested below:

$$RR_3 = \frac{Max_i x_i - Min_i \bar{x}_i}{Max_i x_i} \qquad \text{(RR.4)}$$

and

$$RR_4 = \frac{Max_i x_i - Min_i x_i}{Max_i x_i + Min_i x_i} \qquad \text{(RR.5)}$$

The basic weaknesses of these range-based measures are that they are not based on all values and therefore they do not reflect the change in inequality if there is any transfer of income between two non-extreme recipients.

Instead of considering extreme values at either end, which may not be even known, some scholars have toyed with the idea of the ratio between the mean income of the highest fractile (percentile, quintile or decile) and that of the lowest counterpart. They term it as the extreme disparity ratio (EDR). Naturally, this ratio is not contained in the interval (0.1). This ratio is independent of μ as well. The measure will not reflect the transfer of income that does not involve the extreme fractiles.

11.2.2 Relative Inter-Quartile Range

Sometimes, extremism of the relative range is sought to be moderated by restricting the distribution between the 10^{th} and 90^{th} percentile or sometimes to interquartile range. Bowley (1937) suggested relative quartile deviation as the index of inequality:

$$B = \frac{x^{q^3} - x^{q^1}}{x^{q^3} + x^{q^1}} \qquad \text{(B.1)}$$

where x^{q^r} represents the income level which divides the population in r and (4-r) quartiles. B is zero for degenerate distribution where everybody has the same income and unity if the lowest 75 per cent people have no income at all.

Though the extremes are moderated in comparison to the measure of range, it has an obvious weakness that the measure takes into account only 50 per cent of the distribution. Further, a transfer of income between two persons without causing either or both of them cross x^{q1} or x^{q3} would not change the measured level of inequality. Thus, the index suffers from all weaknesses of the earlier proposals except that of extremism. Its highest value reaches when the lowest 75 per cent people do not possess any income.

A variant of this measure is inter-quartile ratio, which can be defined as the 75^{th} percentile (3^{rd} quartile) income minus 25^{th} percentile (1^{st} quartile) income divided by the median (x^{q2}) income.

11.2.3 Relative Standard Variation

The standard deviation divided by the mean can be used as one measure of dispersion. It is:

$$RSD = \frac{\sigma}{\mu} \qquad \text{(RSD.1)}$$

where σ and μ are standard deviation and mean of the distribution.

It can be equivalently defined as the standard deviation of relative incomes. Using definition of σ , one can find out that it lies in the interval of 0 and $(N-1)^{1/2}$, not in (0,1). The highest value depends on the size of distribution.

Since the measure uses all values, any transfer of income would be reflected in the measure. However, it should be noted that the measure is equi-sensitive to transfers at all levels. Whether a given amount d is transferred between x_j=Rs.400 and x_k=Rs.500, or between x_j=Rs.10,000 and x_k= Rs.10,100, the changed in RSD is exactly the same.

We may finally note that the square of RSD is also quite often used as another measure of inequality, which is known as the coefficient of variance. Quite a few scholars suggest use of variance as a measure of inequality but we have not considered it here primarily because it is not unit-free. We think that an inequality measure must be unit-free.

11.2.4 Standard Deviation of Logarithms

One way of attaching greater importance to transfers at lower end (as required by Sen) is to consider some transformation of incomes. This transformation can easily be attained by considering the logarithms that stagger the income at lower levels.

This measure is defined in either of the following two ways:

$$SDL_1 = \left(\frac{1}{N} \sum_{i=1}^{N} (\log \mu - \log x_i)^2 \right)^{1/2} \qquad \text{(SDL.1)}$$

$$SDL_2 = \left(\frac{1}{N} \sum_{i=1}^{N} (\log \hat{\mu} - \log x_i)^2 \right)^{1/2} \qquad \text{(SDL.2)}$$

where μ and $\hat{\mu}$ are the arithmetic and geometric means respectively. While standard statistical literature prefers use of geometric mean the more common practice in literature on income inequality is one of using arithmetic means.

Cowell (1995) prefers to define these in terms of variance and calls the square of SDL_1 as the logarithm variance (V_1) and the square of SDL_2 as the variance of logarithms (V_2). Name of the second is clear from the expression but that of the first is derived from the fact that ($log\ x$-$log\ \bar{x}$) could be written as log ($x/\bar{x}$). One can see that V_1 is equal to V_2 plus log ($\hat{\mu}$ /μ).

As these measures are in terms of ratios of incomes, any proportionate change in incomes would leave the magnitude of inequality unchanged when measured by these indices. But, unfortunately, a transfer from a richer person to a poorer person may raise the magnitude of inequality, particularly if the poorer person has income more than 2.72 times the mean of the distribution.

While the lower limit, irrespective of formula, is zero when everybody has the same income, the upper limit depends on the size of distribution and approaches infinity when N is large and when everybody except the richest, receives income equal to one unit (as zero is inadmissible in logarithmic transformation.) Further, if we face grouped data, it is convenient to use μ in place of $\hat{\mu}$ and μ_i in place of x_i.

The variance of logarithms is however decomposable. It is a property that is being given emphasis of late. It can be shown that V_2 is the sum of between-group component and within group component, latter being population-weighted sum of within-group V_2's.

11.2.5 Champernowne Index

Champernowne (1973) makes use of the idea of geometric mean. It is a well known fact of an unequal distribution that its geometric mean is smaller than the arithmetic mean. The additive inverse of the ratio of geometric mean to arithmetic mean can duly be considered as an index of inequality. Formally, the index could be written as:

$$CII = 1 - \frac{\hat{\mu}}{\mu} \qquad \text{(CII.1)}$$

where μ and $\hat{\mu}$ are, as stated earlier, arithmetic and geometric means of the income distribution. It is easy to see that its value is bound between 0 and 1.

One can obviously think of another measure where geometric mean is replaced by harmonic means.

These measures are sensitive to transfer to income and change is greater when the transfer takes place at lower end of the distribution. They are sensitive to transfer of income between two persons. One can try it by replacing x_j and x_k by (x_j-d) and (x_k+d) respectively and finding out the direction of the change. Or, one can use differential calculus.

The trouble with these indices is that they cannot be defined when any of the income is zero.

11.2.6 Hirschman-Herfindahl Indices

These indices were developed in the course of studying the commodity concentration in trade by Hirschman (1945) and in characterizing market monopoly in industry by Herfindahl (1950). Later, they were more used in capturing autonomy and dependence of units in a federation.

If each unit is a class in itself, p_i=1/N, i=1,2…,N. Then concentration could be captured through use of q_i's . As the sum of q_i's is always 1, Hirschman devised a measure which would capture the inequality among them. He proposed square root of the sum of squares of shares q_i i=1,2…,N. That is,

$$H_1 = \left(N \sum_{i=1}^{N} q_i^2 \right)^{1/2} \qquad \text{(H.1)}$$

which could be generalized as

$$H_1^* = \left(N \sum_{i=1}^{N} q_i^a \right)^{1/a}, a > 1 \qquad \text{(H.2)}$$

Herfindahl devised a very similar measure, which has been more popular than the original (H.1). This is just the sum of share squares:

$$H_2 = \sum_{i=1}^{N} q_i^2 \qquad \text{(H.3)}$$

$$H_2^* = \sum_{i=1}^{N} q_i^a, a \geq 1. \qquad \text{(H.4)}$$

It is clear that, besides inequality among the shares, the value of these measures depends on *N*—the fewness or largeness of the number of units. For *N*=2, it has been suggested that (1/*N*) could be subtracted from (H.3)

$$H_3 = \sum_{i=1}^{N} q_i^2 - \frac{1}{N} \qquad \text{(H.5)}$$

The minimum value of H_3 is zero. But it serves no great purpose. When *N*=2, for q_1=0.99 and q_2=0.01, while H_2=0.98, H_3=0.48. H_2 scores definitely better than H_3 in characterizing the scene of monopoly.

11.2.7 Kolm's Index

Let there be *N* incomes such that *N*= *nm* where *n* is the number of different incomes and each income has *m* recipients. The number of equal pairs with a given income would be *m*(*m*-1)/2 and total number of equal pairs would be *n.m*(*m*-1)/2. Total number of all pairs would obviously be *N*(*N*-1)/2-*nm*(*nm*-1)/2. One can think of an `equality' index in terms of *nm*(*m*-1)/*nm*(*nm*-1)=(*m*-1)/(*N*-1). The inequality index could then be constructed by subtracting it from1: 1-(*m*-1)/(*N*-1)-(*N*-*m*)/(*N*-1)=*m*(*n*-1)/(*nm*-1)=(*nm*-*m*)/(*nm*-1). In case, income x_1 has f_i recipients, the measure is:

$$K = 1 - \frac{\Sigma f_i^2 - N}{N(N-1)} = \frac{N^2 - \Sigma f_i^2}{N(N-1)} \qquad \text{(K.1)}$$

The purpose of developing this curiosum due to Kolm (1996) is just to make one feel that there could be a variety of simple ways to approach the issue of measurement of inequality.

Check Your Progress 1

1) Define relative range measures of inequality. List out relative merits.

……..

……..

……..

……..

……..

……..

2) How relative inter-quartile range is better than relative range?

……..

……..

……..

……..

……..

3) What is the relative mean deviation? If a transfer of income is between two persons both having income lower than the mean, will it change the magnitude of this index?

……..

……..

……..

……..

……..

4) Compare the two versions of standard logarithmic deviations.

……..

……..

……..

……..

……..

5) What is import of Champernowne Index?

……..

……..

……..

……..

……..

……..

6) What is Hirfindahl index? What are its areas of application?

……..

……..

……..

……..

……..

7) What is the message from the Kolm's index? Calculate the Kolm index for a distribution, which frequency 5 for value Rs.5 lakh and frequency 5 with value Rs.10 lakh and therefore total size 10 and the arithmetic mean 7.5.

……..

……..

……..

……..

……..

11.3 GINI INDEX

This coefficient of concentration, as it is usually called, owes to an Italian statistician by the name of Corrado Gini (1912). Modern practice is to call it just Gini. This index in its origin is positive. There are a number of ways in which this coefficient can be expressed. There are also a number of ways in which it can be interpreted. People have also derived it as a measure of inequality under plausible axioms in welfare theoretic framework. The index satisfies many axioms proposed in literature for an index of inequality.

First, we shall discuss those definitions and expressions, which can be derived as a measure of dispersion. Besides giving its expressions for its frequency data for grouped observations, we shall discuss its welfare theoretic interpretations.

11.3.1 Gini as a Measure of Dispersion

Recall that mean deviation and standard deviation, which are measures of dispersion, seek the deviation from arithmetic mean. Also recall that one of the logarithmic measure sought deviation from the geometric mean. However, one may ask why to seek dispersion in terms of deviations from any mean? Why not compare all pairs and seek the differences. In order to consider positive values of differences, we either take modal values of deviations before averaging (mean deviation) or sum the squares of deviations and take root of the mean of the squared differences.

Corrado Gini (1912) proposed to consider all the differences, that is all pairs of values. By contrast, the range measure of dispersion considers only one pair of highest value and lowest value. When x_i and x_j denote ith and jth incomes respectively and $i, j=1, 2,\ldots, N$, we can see that the aggregate of absolute differences is given by

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_{i-}x_{j}\right| \qquad \text{(G.1)}$$

and because total number of differences is N^2, the mean of absolute differences can obviously be written as

$$\frac{1}{N^2}\sum_{i=1}^{N}\sum_{i=1}^{N}\left|x_{i-}x_{j}\right| \qquad \text{(G.2)}$$

where the differences with the self have also been counted and the difference of x_i with x_j is treated as separate from that of x_j and x_i though numerically they are the same. This is also said to be the case with replacement. Expression (G.2) ranges between 0 and 2μ.

In the case of without replacement, the sum is obviously to be divided by $N(N\text{-}1)$ as there are N deviations with the self. It is not difficult to see that the numerical value of the sum remains the same.

In order to make it serve as a measure of inequality, (G.2) can be divided by μ to produce what can be called coefficient of mean difference (CMD):

$$CMD = \frac{1}{N^2\mu}\sum_{i=1}^{N}\sum_{i=1}^{N}\left|x_i - x_j\right| \qquad \text{(G.3)}$$

CMD fulfils the idea of scale independence. However, the expression (G.3) ranges between 0 when everybody has the same income and 2[=2N/(N-1)] when only one person has all the income. In order to make it satisfy the interval (0,1), we can further divide it by 2. The result is Gini coefficient of concentration or Gini index of inequality:

$$G = \frac{1}{2N^2\mu}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|x_i - x_j\right| \qquad \text{(G.4)}$$

or

$$G = \frac{1}{N^2\mu}\sum_{i=1}^{N}\sum_{x_j \le x_i}^{N}(x_i - x_j) \qquad \text{(G.5)}$$

conceived as an aggregate of only positive differences, though normalized by the number of all differences and the mean income. Kendall and Stuart define this as 'one half of the average value of absolute differences between all pairs of incomes divided by the mean income'.

The index can also be defined in terms of population proportions and income shares. If the income-share of individual i is denoted by q_i, that is,

$$q_i = \frac{x_i}{N\mu}, \qquad \text{(G.6)}$$

then the expression (G.4) can also be written as

$$G = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left|q_i - q_j\right| \qquad \text{(G.7)}$$

In the case of a discrete distribution, each individual constitutes (1/N)th of the population, that is,

$$p_i = \frac{1}{N}. \qquad \text{(G.8)}$$

Therefore one can also write (G.7) as

$$G = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \left| p_j q_i - p_i q_j \right|. \qquad \text{(G.9)}$$

An obvious question is: why not $\left| p_i q_i - p_j q_j \right|$ in the expression (G.9)? For understanding this, let us consider the Gini coefficient for the groups.

Let μ_r and μ_s denote mean incomes of r^{th} and s^{th} groups (say, families) respectively and r,s=1,2…g. Then, Gini for the groups can be defined as

$$G = \frac{1}{2N^2\mu}\sum_{r=1}^{g}\sum_{s=1}^{Ng} \left| \mu_r - \mu_s \right| f_r f_s \qquad \text{(G.10)}$$

where f_r and f_s are frequencies of the groups r and s respectively. This can obviously be written as

$$G = \frac{1}{2}\sum_{r=1}^{g}\sum_{s=1}^{g} \left| \frac{\mu_r}{\mu} - \frac{\mu_s}{\mu} \right| p_r p_s \qquad \text{(G.12)}$$

where

$$p_r = \frac{f_r}{N} \; and \;\; p_s = \frac{f_s}{N} \qquad \text{(G.12)}$$

Now, let us note the share of total income with the group r:

$$q_r = \frac{\mu_r f_r}{N\mu} = p_r \frac{\mu_r}{\mu} \qquad \text{(G.13)}$$

Then, the expression (G.12) can be written in either of the two ways (G.14) and (G.15)

$$G = \frac{1}{2}\sum_{r=1}^{g}\sum_{s=1}^{g} \left| \frac{q_r}{p_r} - \frac{q_s}{p_s} \right| \qquad \text{(G.14)}$$

or

$$G = \frac{1}{2}\sum_{r=1}^{g}\sum_{s=1}^{g} \left| p_s q_r - p_r q_s \right| \qquad \text{(G.15)}$$

It is easy to see that g=N and p_s=p_r=(1/N) when f_r and f_s are all equal to1.

In statistics literature we emphasize frequency aspects, in economics literature we find it convenient, expression-wise, to treat each individual with single income though there is no bar for x_i=x_j.

11.3.2 Simple Computational Device

Two years after giving his index to terms of relative mean differences, Gini (1914) showed that the index is exactly equal to one minus twice the area under Lorenz curve (to be discussed later). That is,

$$G = 1 - 2\overline{A} \qquad \text{(LG.1)}$$

where $\overline{A}$ is the area under the Lorenz curve, as shown in Fig. 11.1

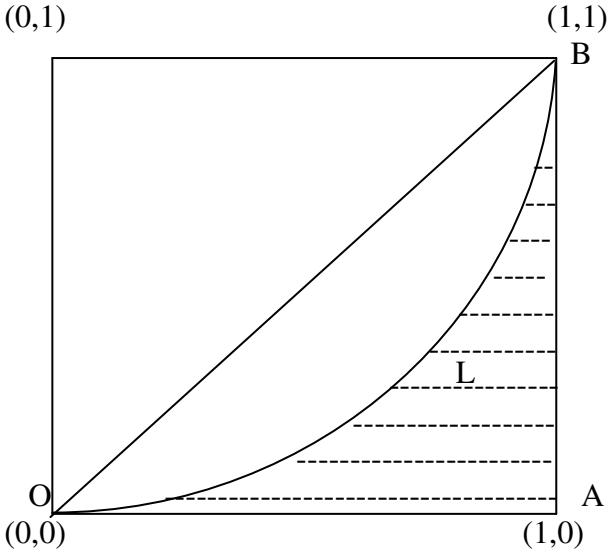


Fig. 11.1

However, normally, we do not specify and estimate a smooth relationship between Q and P. Instead, we obtain the curve by plotting cumulative proportions of people in classes and cumulative shares of their incomes, where classes are arranged according to increasing per capita income values:

$$\mu_1 \leq \mu_2 \leq \mu_3 \leq .... \leq \mu_r \leq ... \leq \mu_g . \qquad \text{(LG.4)}$$

strictly speaking, equality sign is useless in this presentation.(We have written it in deference to Theil (1967). Plotting P_r and Q_r, we obtain the Fig. 11.2. We can see that the area below the Lorenz curve can be conceived as consisting of several trapeziums. A trapezium could be seen as consisting of a rectangle and a triangle. Summing the areas of all trapeziums (say g in number), we can get the area $\overline{A}$. Substituting it in (LG.1), we get the following expression for computing Gini coefficient G:

$$G = 1 - \sum_{r=1}^{g} (P_r - P_{r-1})(Q_r + Q_{r-1}) \qquad \text{(LG.5)}$$

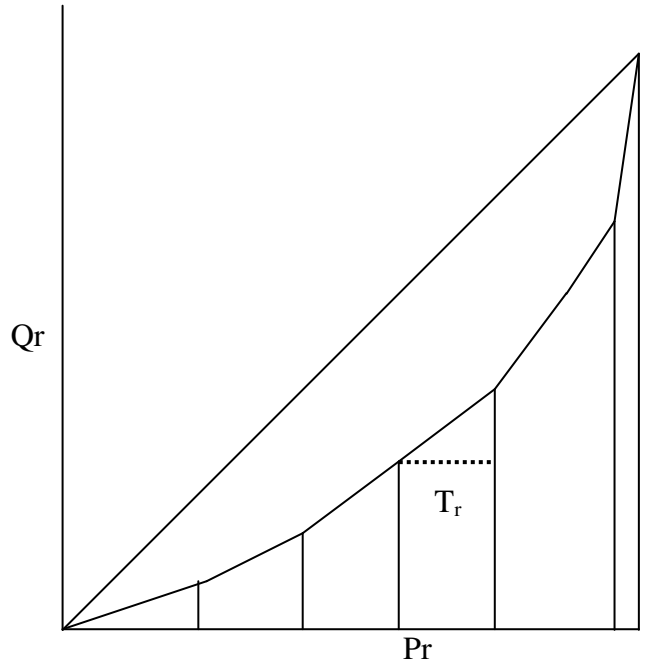


Fig. 11.2

Check Your Progress 2

1) Define Gini ratio.

……

……

……

……

……

2) Show the difference in approach in defining Gini from other measures of dispersion.

……

……

……

……

……

3) Write the expression for computing Gini.

……

……

……

……

……

4) How can you compute Gini ratio?

……

……

……

……

……

11.4 LORENZ CURVE

Lorenz curve is a powerful geometrical device to compare two situations of distribution with regard to their level of inequality. Devised some hundred years ago by Max O. Lorenz (1905) to measure concentration of wealth, it is still very widely used in empirical studies on inequality. The device can be used for comparing inequality of distribution of any measurable entity such as

income, wealth (land, capital), consumption, expenditure on an item (say, food or education), etc. The distribution may be over persons or households. But the device can also be used to measure inequality of tax collection or expenditure incurred by states or federal grants received by different states. We can compare pre-tax and post-tax distributions in order to study the efficacy of instrument of tax.

Lorenz (1905) studied a number of methods then in use to gauge the level of, or change in the level of, inequality. Most of these measures used fixed-income classes in data tabulation and made inter-temporal comparison, employed changes in percentage of recipients of class incomes in each of the fixed-income classes or movement of persons from one class to another and so on. Finding them unsatisfactory, he comes to the conclusion that changes in income and changes in population both have to be simultaneously taken into account and in a manner that 'fixed-ness' of income classes gets neutralized.

In fact, this measurement relates to comparison and in most cases, we are in a position to compare but there are situations of non-comparability. However, a few of inequality measures that are capable of numerical representation in terms of a scalar number, and therefore called summary measure, are found to be based on the Lorenz curve.

It may be pointed out that the curve was independently introduced by Gini (1914). It is therefore, quite often referred to as Lorenz-Gini curve as well. We shall, however, stick to more common usage and call it Lorenz Curve.

11.4.1 Geometrical Definition

The Lorenz curve of concentration of incomes is the relationship between the cumulative proportions of recipients, usually plotted on the abscissa, and the corresponding cumulative shares of total income with the recipients, usually plotted on the ordinate. If population proportions and income shares of class j are denoted by p_j and q_j and cumulative proportions and shares upto class i, by P_i and Q_i then

$$P_i = \sum_{j=1}^{i} p_j, \qquad 1 \geq p_j \geq 0 \qquad \text{(GD.1)}$$

and

$$Q_i = \sum_{j=1}^{i} q_j, \qquad 1 \geq q_j \geq 0 \qquad \text{(GD.2)}$$

The relationship between P_i and Q_i is given by the curve

$$Q_i = L(P_i), \qquad 1 \geq P_i \geq 0, \quad 1 \geq Q_i \geq 0 \qquad \text{(GD.3)}$$

and the point on the curve by (P_i, Q_i). Naturally, the first point is (0,0) and the last one on the curve, (1,1). It is also clear that $Q_i \leq P_i$ i=1, 2,…, N-1 if there are N classes of incomes. It means no point will make an angel of more than 45° with the abscissa at the origin. Then, one can be sure that the Lorenz curve lies in the lower triangle of Lorenz Box of the unit square. See Fig. 11.3 in which OLB shows the Lorenz curve (Fig. 11.3).

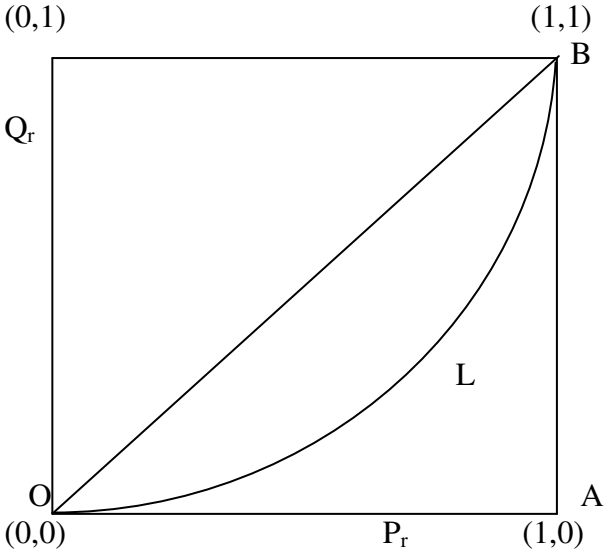


Fig. 11.3

11.4.2 Properties of the Lorenz Curve

Now it is easy to see that the extreme case of perfect equality is given by the diagonal OB which represents $P_i=Q_i$, i=1, 2,…, N. The other extreme of perfect inequality will be given by a curve OAB. The diagonal OB is often designate as the egalitarian line or line of equality. The other diagonal CA is known as the alternative diagonal and is useful to study the symmetry of the curve. The tringle OAB with sharp kink of 90˚ at A can be said to be the line of perfect inequality. See Fig. 11.4.

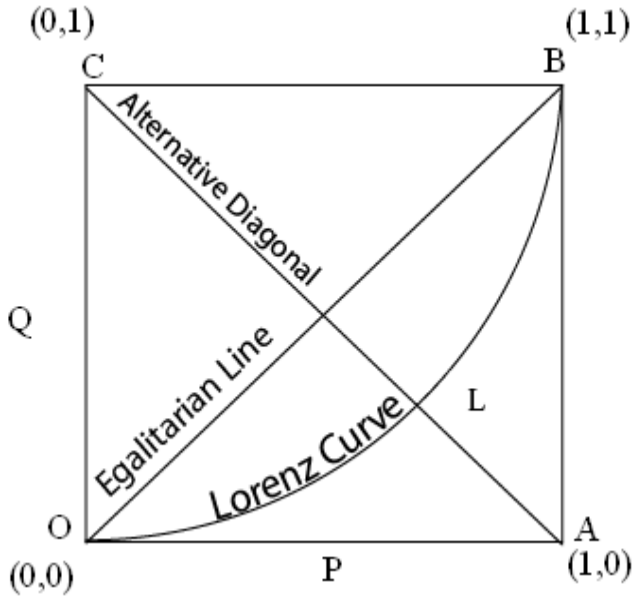


Fig. 11.4

We can finally note the following properties:

i) $1 \geq p_i \geq 0; 1 \geq q_i \geq 0,\ i = 1,2,...,N$

ii) $1 \geq P_i \geq 0; 1 \geq Q_i \geq 0, i = 1,2,...,N - 1$

iii) $P_0 = Q_0 = 0; P_N = Q_N = 1$

iv) $P_i \geq Q_i, i = 1,2,..,\ N - 1$

By drawing a Lorenz Curve, we can know whether a given distribution is equal or unequal. We do not yet know how much unequal a given distribution is. When we draw two or more Lorenz Curves, we can compare the distributions as regards their levels of inequality. The curve closer to the diagonal of equality has lower level of inequality than the one away from it (Fig. 11.5). But we do not know yet the level of inequality. And even this comparison is possible only when the curves do not intersect (Fig. 11.6).

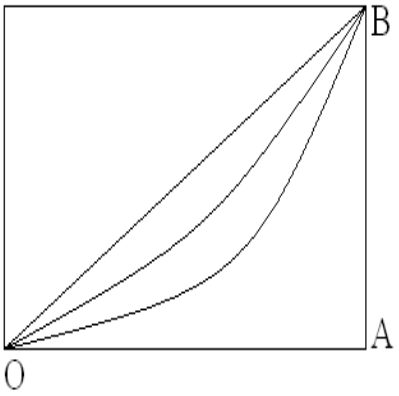


Fig. 11.5

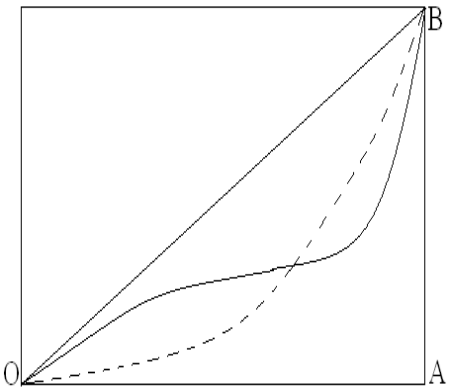


Fig. 11.6

However we can devise some measures, which are based on the Lorenz curve. In case the Lorenz curves intersect, reducing the distributions into single real number is the only option. So we shall discuss only two such proposals.

11.4.3 A Measure Based on Area

We have noted that if Lorenz curve coincides with the diagonal of equality, the inequality is nil and if Lorenz curve coincides with the two sides of the square, the inequality is full.

In the case of non-intersecting Lorenz curves, it is clear that the curve closer to the diagonal of equality will circumscribe smaller area between itself and the diagonal of equality than the one, which is farther. Which is what it should be. We can therefore devise a measure of inequality by dividing the area OLB (Fig. 11.4) by the area of triangle OAB, which is the maximum possible area between the diagonal of equality and Lorenz curve. As the area of OAB is (1/2), the measure turns out to be twice the area between the diagonal of equality and the Lorenz curve. In other words, Lorenz coefficient of concentration (LCC) is:

$$LCC = \frac{Area\,OLB}{\Delta OAB} = 2\,Area\,OLB$$

Since this turns out to be exactly equal to Gini coefficient, we are not elaborating it any further.

11.4.4 A Measure Based on Length

This is a measure proposed by Kakwani (1980). The length of the Lorenz curve, denoted by ℓ, cannot fall below $\sqrt{2}$, which is the length of the egalitarian line and cannot exceed 2, which is the sum of the lengths of the two arms of the lower triangle. In order to produce a measure with the minimum value 0 and the maximum value 1, following exercise can be suggested:

	Minimum	Actual	Maximum
Length of the Curve	$\sqrt{2}$	ℓ	2
Length of the Curve - $\sqrt{2}$	0	$\ell-\sqrt{2}$	$2-\sqrt{2}$
$\frac{\text{Length of the Curve } -\sqrt{2}}{\text{Maximum length } -\sqrt{2}}$	0	$\frac{\ell-\sqrt{2}}{2-\sqrt{2}}$	1

So this measure is clearly:

$LK = (\ell-\sqrt{2})/(2-\sqrt{2})$

In both the cases, one can draw actual graphs and actually measure the area and the length and calculate the indices for level of inequality. Those who wish to carry out a more sophisticated exercise will have to estimate smooth functions.

Check Your Progress 3

1) Enumerate the properties of Lorenz curve.

……

……

……

……

2) When will comparison between two Lorenz curve fail to compare inequality in two distribution?

……

……

……

……

3) What is the relationship between Lorenz curve and Gini coefficient.

……

……

……

……

4) What is Kakwani's measure of inequality, which is based on the Lorenz curve.

……

……

……

……

……

11.5 NORMATIVE MEASURES

The measures that essentially involve judgement about values through specification of social welfare function are called normative measures. The arguments of this nature were first advanced by Dalton, pretty eight decades ago in 1920 for constructing what are today called normative measures of inequality.

Reacting to an observation by Pearsons (1909) that 'the statistical problem before the economists in determining upon a measure of inequality in the distribution of wealth is identical with that of the biologist in determining upon a measure of the inequality in the distribution of any physical characteristic', Dalton (1920) pointed out that 'economist is interested, not in distribution as such, but in effects of the distribution upon the distribution (and total amount) of economic welfare which may be derived from income'. The objection to great inequality of income, he further points out, is due to the resulting loss of potential economic welfare that could accrue to people in the absence of it.

Yet, it has to be noted that inequality though defined in terms of economic welfare, has to be measured in terms of income. This idea due to Dalton has been conceded by subsequent contributions. Using the notion of social welfare function in construction gives rise to **normative** measures of inequality.

It may be instructive to remember that the discussion would revolve around three issues:

1) the relationship between income of a person and his welfare;
2) the relationship between personal income-welfare functions; and
3) the relationship between personal welfare and social welfare.

It may be noted that utility is the word mostly used for personal welfare whereas for welfare of society the phrase social utility is rarely used.

There are two major indices in this category: Dalton's index and Atkinson's index. In Atkinson's index a new idea is introduced and that is of equally distributed equivalent income. Actually there are two sub-approaches within normative approach. One is Dalton's and the other is Atkinson's. While in Dalton's approach present social welfare is compared with that could be obtained by equally distributing the total income, in Atkinson's approach the present level of income is compared with that of equally distributed level of income, which generates the present level of social welfare. Sen has generalised the Atkinson's index. Theil's index based on information theory could be suggested here only to sort of complete the unit.

11.5.1 Dalton Index

For each individual, Dalton assumes, marginal economic welfare diminishes as income increases. It means income-welfare function

$$U_i = U_i(x_i), i = 1,2,...N \qquad \text{(D.1)}$$

(where U_i is welfare of person i possessing income $x_{i)}$

is concave, suggesting that $(\partial U_i/x_i) > 0$ but $(\partial^2 U_i/x_i^2) < 0$. Dalton further assumes that economic welfare of different persons is additive. Thus, in his scheme, social welfare is a simple aggregation of personal welfares. In other words, social welfare W is given by

$$W = \sum_{i=1}^{N} U_i(x_i) \qquad \text{(D.2)}$$

He further assumes that the relation of income to economic welfare is the same for all members of the community. That is,

$$U_i = U(x_i), \text{ i=1,2,…,N} \qquad \text{(D.3)}$$

In that case, the relation (D.2) can be expressed as

$$W = \sum_{i=1}^{N} U(x_i) \qquad \text{(D.4)}$$

which makes it clear that whosoever gains in welfare, the addition to the social welfare is the same. For any given level of social welfare, any distribution of welfare among the members of the society is permissible. However, one must remember that the relation of individual income to their welfares is concave. Therefore, transfer of income from A to B will not lead to symmetric change in welfares of those two persons involved in the transaction. The result is some impact on W the measure of social welfare.

From Fig. 11.7, we may compare the situation when two individuals, both possessing the same relation, have two different income levels, with that when they have the same (mean) income. We may note that the sum of the welfare of person 1(BB') and the welfare of person 2(DD') is less than the twice of CC' which is the level of welfare enjoyed by both the persons when they have equal income. It is easy to see that the loss suffered by person 2, that is D'E, is overcompensated by the gained by person 1, which is C'F.

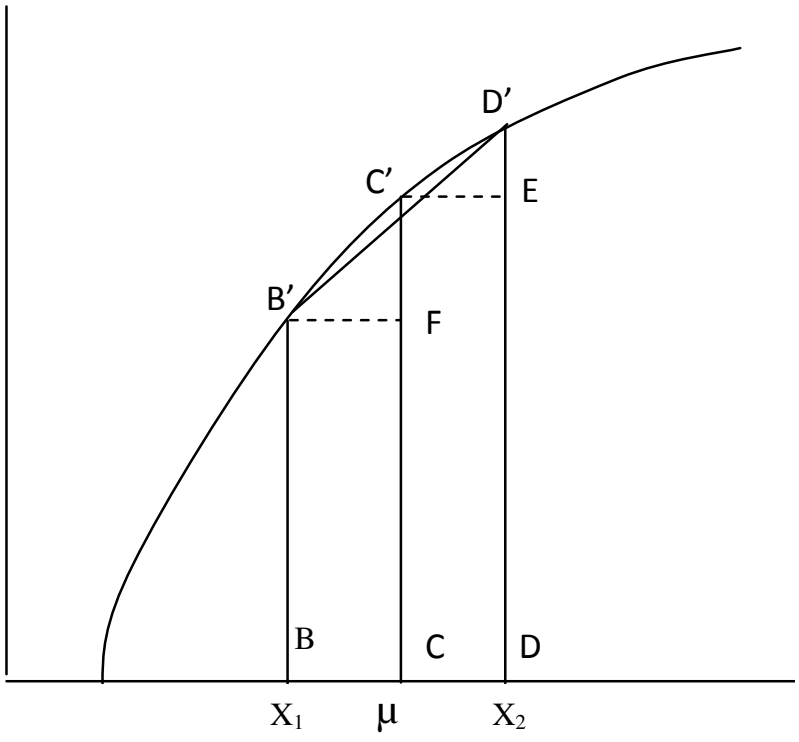


Fig. 11.7

This demonstrates that, under assumptions by Dalton, an equal distribution is preferable to an unequal one for a given amount of total income from the viewpoint of social welfare. In fact, for a given total of income, the economic welfare of the society will be maximum when all incomes are equal. The inequality of any given distribution may therefore be defined as

$$D_1 = \frac{\sum_{i=1}^{N} U(\mu)}{\sum_{i=1}^{N} U(x_i)} = \frac{NU(\mu)}{\sum_{i=1}^{N} U(x_i)} \qquad \text{(D.7)}$$

which is equal to unity for an equal distribution and greater than unity for an unequal one. It may therefore be preferred to define the Dalton's index as

$$D_2 = \frac{NU(\mu)}{\sum_{i=1}^{N} U(x_i)} \qquad \text{(D.8)}$$

which is obviously zero for an equal distribution. How large can it be? It will depend on the values of *U(0)*, *U(μ)* and *U(Nμ)* when *N* and μ are given, not necessarily 1. Later writers have therefore preferred to define Dalton's index in the following form, which inverts the arguments of D_2 subtract it from 1:

$$D = 1 - \frac{\sum_{i=1}^{N} U(x_i)}{NU(\mu)} = 1 - \frac{\overline{U}}{U(\mu)} \qquad \text{(D.9)}$$

It looks as if the index is contained in the interval (0,1). However, there are many valid concave functions where it may not hold true. For example, if we have $U(x_i) = \log x_i$, then $D = 1 - \{\log \hat{\mu} / \log \mu\}$. Given the fact that $\hat{\mu} < \mu$, D would turn out to be a negative number for $\mu < 1$. And μ could be less than 1 as *x* can be measured in any unit. It would be the same case $U(x_i) = 1/x_i$.

However, in order to obtain numerical magnitude, it is not sufficient to define the index. Dalton (1920) points out that though defined in terms of economic welfare, inequality has to be measured in terms of income. Then, no unique measure of inequality will emerge. It will verily depend on the particular functional relationship assumed. Dalton himself considered two such functions for the purpose of illustration. The first is related to Bernaulli's hypothesis. It holds that proportionate additions to income (in excess of that required for bare subsistence-poverty line) make equal additions to personal welfare, That is,

$$dU_i = \frac{dx_i}{x_i} \; or \; U_i = \log x_i + c_i \qquad \text{(D.10)}$$

Under the assumption that every person has the same functional relationship, the Dalton's index can be given as

$$D = 1 - \frac{\log \hat{\mu} + c}{\log \mu + c} \qquad \text{(D.11)}$$

where $\hat{\mu}$ is the geometric mean of personal incomes. The other formulation he discusses is given as

$$dU_i = \frac{dx_i}{x_1^2} or U_i = c - \frac{1}{x_i} \quad \text{(D.12)}$$

where c is the maximum welfare one can obtain when $x - \propto$. Dalton's index in this case would turn out to be:

$$D = 1 - \frac{c - (1/\tilde{\mu})}{C - (1/\mu)} \quad \text{(D.13)}$$

where $\tilde{\mu}$ is the harmonic mean.

11.5.2 Atkinson Index

Atkinson (1970) objects to Dalton's measure because *D* is not invariant with respect to positive linear transformations of personal income-welfare functions. This was pointed out by Dalton himself but he could not resolve it.

Atkinson seeks to redefine the index in such a way that measurement would be invariant with respect to permitted transformations of welfare numbers. Atkinson does it through devising what he calls 'equally distributed equivalent income'. Both the distributions, the original and the new one, are supposed to yield the same level of welfare.

In order to make the concept clear, we put a few artifacts along with the actual distribution. Let us first note that for an actually distributed income vector x_i, i=1,2…,N (call it vector a), there is only one equally distributed income vector with each element equal to μ (call it vector b) but there are a number of equivalently distributed, vectors (call them vectors c). See Chart 1. An equivalent income distribution is one, which has the same level of welfare as that of currently given distribution. However, one of these equivalent distributions (vectors c) is `equal' as well. This is called equally distributed equivalent income vector, shown as vector (d) in the Chart 1. As μ is the mean level of current distribution, μ^* may be used for designating the level of equally distributed equivalent income.

CHART-I

Vector (a) Actually distributed income vector $x_1, x_2 \ldots, x_i, \ldots, x_N.$

Vector (b) Equally distributed income vector μ, μ, …., μ, …, μ.

Vector (c) Equivalently distributed income vector $x_1^*, x_2^*, \ldots, x_1^*, \ldots, x_N^*$.

Vector (d) Equally distributed equivalent income vector μ*, μ*, …, μ*, …, μ*

If should be obvious that $W_b \geq W_a = W_c$ *and* $W_c = W_d$. *Then*, $W_a = W_d$. W represents social welfare with respective distributions of income vectors. It is clear that $\mu \geq \mu^*$. μ^* is defined by the additive social welfare function having symmetric individual utility functions such as:

$$U(\mu) = \frac{1}{N}\sum_{i=1}^{N} U(x_i) \quad \text{(A.1)}$$

or equivalently

$$\mu^* = U^{-1}\left[\frac{1}{N}\sum_{i=1}^{N} U(x_i)\right] \quad \text{(A.2)}$$

The index due to Atkinson is then defined as the additive inverse of the ratio of equivalent mean income to actual mean income:

$$A = 1 - \frac{\mu^*}{\mu} \tag{A.3}$$

which is said to lie between zero (complete equality) and 1 (complete inequality). We can see that A cannot be 1 unless μ^* is zero, which is an impossibility for any distribution with $\mu>0$. If complete inequality is defined as the situation when only one person grabs all the income, we can see that

$$A = 1 - \frac{\mu^*_m}{\mu} \tag{A.4}$$

where

$$NU(\mu^*) = \sum_{i=1}^{N} U(x_i)$$

and

$$NU(\mu^*_m) = (N - 1)U(0) + U(N\mu).$$

This index is not scale independent unless some restriction is imposed on the relationship U. If this requirement has to be met, Atkinson points out, we may have to have the following form

$$U(x_i) = \begin{cases} \alpha + \dfrac{\beta}{1-\varepsilon} x_i^{1-\varepsilon}, & \varepsilon \neq 1 \\ \log_e x_i, & \varepsilon = 1 \end{cases} \tag{A.5}$$

Note that we need $\epsilon \geq 0$ for ensuring concavity and $\epsilon > 0$ for ensuring strict concavity. This is a homothetic function and is linear when $\epsilon=0$. We may note that ϵ cannot exceed 1 as in that case the varying component assumes inverse relationship. α is usually negative so that $U(x_i)$ is not positive for $x_i = 0$. Otherwise, when $x_i = 0$, Ui = α which means that welfare is positive even when income is zero. This is generally not acceptable. On the contrary, a negative α would be more acceptable. When $\epsilon = 1$, α is infinitely large and negative.

Since ϵ can be zero, Atkinson's requirement is not strict concavity. Sen (1973) has a question. He asks to consider two distributions (0,10) and (5,5) along with

$$U(x_i) = \alpha + \beta(x_i) \tag{A.6}$$

Then, he points out the level of social welfare would be $(2\alpha + 10\beta)$ whatever the distribution. μ^* would be 5 in both the cases. μ is of course 5. The measure of inequality A is therefore zero. So, both the distributions are ethically equal. This is obviously absurd. Therefore, the relation (A.5) should be defined with the restriction $\epsilon>0$. We should also note that (A.4) is an iso-elastic marginal utility function.

ϵ is the inequality-aversion parameter and has close resemblance with risk-aversion premium. Atkinson proposed to draw on the parallel formally with the problem of measuring risk. He finds his concept of equally distributed equivalent income very closely resembles with risk-premium or certainty equivalent income as used in the theory of decision-making under uncertainty.

In case we seek to introduce this restrictive personal income-welfare function along with simple aggregation of individual welfares to constitute the social welfare, into the inequality measure A, we will have

$$A = 1 - \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i}{\mu} \right)^{1-\varepsilon} \right]^{1/(1-\varepsilon)} \quad , \varepsilon \neq 1 \qquad \text{(A.7)}$$

The question is now narrowed down to choosing ϵ. As ϵ rises, more weight is attached to transfers at the lower end of the distribution and less to that at the top. When ϵ rises, (A.7) assumes the function $\min_i (x_i)$, which only takes account of transfers to the very lowest income group (and is therefore not strictly concave). When $\epsilon = 0, U_i$ is linear. As a consequence, A is always zero. This means A has no descriptive content at all. When $\epsilon \to 1$, A turns out to be

$$A = 1 - \prod_{i=1}^{N} \left(\frac{x_i}{\mu} \right)^{1/N} = 1 - \frac{\hat{\mu}}{\mu} \qquad \text{(A.8)}$$

which is the same as Champernowne index (CII.1). For values of ϵ between 0 and 1, the expressions may not be very neat. Parameter ϵ is often chosen to be ½ or 1/3 or 2/3.

11.5.3 Sen Index

There are people who feel rather strongly that the social valuation of the welfare of individuals should depend crucially on the incomes of their neighbours too. Then, why should society add simply individual welfares? One may also question the assumption of one welfare function for all individuals. If we do so, we should go for broad social welfare function such as

$$W = W(x_i, x_2, ..., x_N) \qquad \text{(S.1)}$$

which is just symmetric, quasi-concave and increasing in individual income levels. Then, a more general normative measure of inequality can be defined by devising the concept of 'generalized equally distributed equivalent income'. This is obviously the level of per capita income x^* which, if shared by all, would produce the same level of W as is generated by the present actual distribution. That is,

$$x^* = x \,|\, W(x^*, x^*, .., x^*) = W(x_1 x_2, ..., x_N) \qquad \text{(S.2)}$$

Under the assumption that (S.1) is quasi-concave, $x^* \leq \mu$ for every distribution of income. The index S would then be

$$S = 1 - \frac{x^*}{\mu} \qquad \text{(S.3)}$$

which is but a generalized version of *A*. If utilitarian framework is employed, then *S* and *A* turn out to be indistinguishable.

These measures, it may be noted, clearly suggest that there exists a redistribution equivalent of growth so far as the concern is about raising the welfare.

11.5.4 Theil Entropy Index

Theil (1967) poses a question: Does information theory supply us with a 'natural' measure of income inequality among N individuals, which is based on income shares? He answers: Yes. Here is a short introduction.

Let us start with income share of individual i:

$$q_i = \frac{x_i}{N_\mu} > o \quad such\,that \sum_{i=1}^{N} q_i = 1 \tag{T.1}$$

When $x_i = \mu$, i=1, 2,…, N, that is, when distribution is equal, we have

$$q_i = \frac{1}{N} \quad i = 1,\ 2,...N \tag{T.2}$$

We have complete inequality when some $x_i = N\mu$ and $x_j = 0, j \neq i$. It implies that $q_i = 1$ for some i and $q_j = 0, i \neq j$.

In information theory, one way of defining entropy of probabilities p_i is

$$H = \sum_{i=1}^{N} p_{i.} \log \frac{1}{P_i.} \tag{T.2}$$

Replacing probabilities by shares, we have

$$H = \sum_{i=1}^{N} q_{i.} \log \frac{1}{q_i.} \tag{T.3}$$

which can be taken as a measure of equality. For the situation of complete equality, we can see that H is equal to $\log N$ and for that of complete inequality H is zero. We can therefore define Theil index T as

$$T = \log N \sum_{i=1}^{N} q_{i.} \log \frac{1}{q_i.}$$

$$= \sum_{i=1}^{N} q_i \log N - \sum_{i=1}^{N} q_{i.} \log \frac{1}{q_i}$$

$$= \sum_{i=1}^{N} q_i \log N.q_i \tag{T.4}$$

This measure is motivated by the notion of entropy in information theory. But one can see that it can be interpreted in the traditional normative framework with

$$U_i = q_i \log \frac{1}{q_i} \qquad \text{(T.5)}$$

and

$$W = \sum_{i=1}^{N} U_i(q_i). \qquad \text{(T.6)}$$

We may note that (T.5) depends on x_i as well as on μ along with N and U that it is concave with respect to x_i.

While the lower limit of T is zero, its upper limit *log N* increases as the number of individuals increases. To many people, it is objectionable. However Theil (1967) chooses to defend it. When society consists of two crore persons and one grabs all and when society consists of two persons and one grabs all, cannot have the same level of inequality. The former case is equivalent to the situation in which one crore out of two crore people have nothing and the other one crore have equal income. Maximum value for two-person society is log 2, and that for two crore-person society is 7 log 2. Some researchers still insist that the measure should be normalized by dividing it by log N.

Check Your Progress 4

1) What is social welfare function, according to Dalton?

……

……

……

……

……

2) Discuss Dalton index of inequality.

……

……

……

……

……

3) Give the logic behind Atkinson index.

……

……

……

……

……

4) How is Sen index distinct from Atkinson's index?

……

……

……

……

……

5) Discuss Theil's entropy index of inequality.

……

……

……

……

……

11.6 LET US SUM UP

Owing to adverse impacts of economic inequality both on poverty and on growth, reducing inequality has been a priority of public policy. Various measures of income inequality can be put under two categories: positive measures and normative measures. The positive measures capture the inequality of income without value judgements. These include range quartile range, standard deviation, Gini ratio, etc. Lorenz curve belongs to this category. It measures inequality to the extent of comparing two distributions. The measures, which essentially involve value judgements about social welfare, are normative measures. These include indices propounded by Dalton, Atkinson, Sen, and Theil.

Easy comprehension and easy computation, range of variation and amount of information needed the desirable properties of the measures of economic inequality. In order to judge the efficacy of an inequality index, several axiom have been set up. However, these axioms have been relegated to the Appendix.

11.7 KEY WORDS

Co-efficient of Mean Difference : Mean of all pair-wise differences divided by the mean of differences has been termed as coefficient of mean difference in this text.

Dispersion : The fact that values of a variable are not all the same is known as dispersion. The spread or scattering of the distribution is measured by a measure of dispersion.

Extreme Disparity Ratio : The ratio of the highest value to the lowest value is known as extreme disparity ratio.

Normative Measures of Inequality : Measures of inequality, which are articulated through the explicit incorporation of social welfare function or social welfare considerations, are known as the normative measures of inequality.

Positive Measures of Inequality : Measures of inequality, which are based in statistical properties of distribution, are known as the positive measures of inequality.

Relative Standard Deviation : Standard deviation of a distribution divided by its mean is known as Relative Standard Deviation.

Standard Logarithmic Deviation : Standard deviation of logarithms of values in a distribution is known as Standard Logarithmic Deviation. Though logically the deviations of logarithmic values should be taken from the logarithm of geometric mean but at times they are taken from logarithm of arithmetic mean. Therefore, there are two versions.

Social Welfare Function : An index of social well-being, often articulated as a function of individual utilities or individual incomes or individual consumption baskets, with or without labour disposition, or individual rankings of potential state of affairs.

11.8 EXCERCISES

1) Following the adapted distribution of monthly per capita expenditure (in Rs.) in rural India in the 60th round of the NSS over January to June 2004:

Class	50-225	225-255	255-300	300-340	340-380	380-420	420-470	470-525	525-615	615-755	755-950	950-1200
Avg. Exp.	100	240	280	325	365	405	450	500	580	700	850	1100
Percentage of persons	2.4	2.7	6.4	8.3	9.6	9.6	10.8	10.0	12.2	12.6	6.7	8.8

Calculate as many positive measures as you can.

2) Following is distribution data of operational holdings from agriculture census 1976-77:

	Holding	Marginal	Small	Semi-medium	Medium	Large	All
Definition	Unit	0.0-1.0 Ha	1.0-2.0 Ha	2.0-4.0 Ha	4.0-10.0 Ha	Above 10.0 Ha	
Number	'000	44523	14728	11666	8212	2440	81569
Area	'000 Ha	17509	20905	32428	49628	42673	163343

Draw Lorenz curve and compute Gini ratio.

11.9 SOME USEFUL BOOKS

Chaubey, P. K. (2004), *Inequality: Issues and Indices*, Kanishka Publishers, Distributors, Delhi.

Cowell, Frank A. (1995), *Measuring Inequality*, Prentice Hall/Harvester Wheatsheaf, London;

Sen, A.K. (1997) *On Economic Inequality*, Oxford University Press, Oxford. (with Annexe by James E. Foster.

11.10 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Range is the difference between the maximum and minimum value. With a view to ensuring that the index of inequality based on this measure of dispersion is unit-free and/or is confined in the interval of (0,1), various ways of normalizations could be considered. See Sub-section 11.2.1.

2) In the relative range only extreme values are considered, which may not be representatives of the distribution. It is like comparing the poorest (who may be a few) with the richest (who may be one). Inter-quartile measures take into account the middlemost distribution with 50 percent recipients. See Sub-section 11.2.2

3) In the mean deviation, all values in a way are considered. The mean of absolute deviations is divided by the mean of the distribution to yield relative mean deviation. See Sub-section 11.2.3. Since the sum of deviations on one side of the mean will not change with the transfer of income contemplated, the magnitude of the index would not change.

4) See Sub-section 11.2.4. In one version the deviations are taken with respect to logarithm of geometric mean and in the other with respect to that of arithmetic mean. With some mathematical manipulation, one can find out that former is smaller than the latter by square of the difference between the logarithms of geometric mean and arithmetic mean.

5) It is a straight application of the fact that geometric mean of a distribution is smaller than its arithmetic mean. Of course, when the values are greater than 1!

6) Hirfindahl index is sum of squares of the shares with each recipient, which of course varies with the number of recipients. Equal distribution of shares between two recipients will yield a value of 0.5 and between three recipients, 0.333. It is therefore more used as a measure of concentration. See Sub-section 11.2.5.

7) The message is that one can try on one's own to devise new methods. See Sub-section 11.2.6. For second part, the answer is 5/9. Try it out.

Check Your Progress 2

1) Gini ratio is one half of the average value of absolute differences between all pairs, including with self, of values divided by the arithmetic mean.

See the definition by Kendall and Stuart in Sub-section 11.3.1. Have a look at expression (g.4) and (G.7).

2) The basic difference lies Gini index from other measures based on dispersion that in articulating all differences are considered while in others either few differences are considered or deviations from arithmetic/ geometric means are considered.

3) See Sub-section 11.3.2.

4) By writing out in a table, columns for class intervals or values, frequencies, class total values, cumulative frequencies, cumulative total values, cumulative proportions and cumulative shares in respect of each class. For using expression (LG.5), consecutive moving differences (or sums) of proportions and consecutive moving sums (or differences) need to be computed in two additional columns. MS-Excel will do well.

Check Your Progress 3

1) Look at the Fig. 11.4 and Sub-section 11.4.2. Try writing the properties in language.

2) When the two Lorenz curve will intersect, it will not be possible to say on balance which distribution is more unequal. In fact, one section in that case will be more unequal and the other section less unequal in distribution 1 in comparison to their counterparts in the distribution 2.

3) The value of Gini coefficient is equal to twice the areas inscribed between line of equality and the Lorenz curve. It is equivalent to saying that G is equal to (1-A) where A is the areas below Lorenz curve in the unit box.

4) Kakwani's measure of inequality is normalized length of the Lorenz curve.

Check Your Progress 4

1) Dalton's social welfare function is a simple aggregation of welfare (utility) functions of the individuals constituting the society. In addition, all individuals are supposed to have the same income utility function.

2) Sub-section 11.5.1. Write Dalton's proposal and its modern version. Also point out that though conceived in terms of utility, Dalton held that the index has to be measured in terms of income only.

3) An inequality measure should not change with linear transformation of personal utility function. Since Dalton's index does not respect this property, Atkinson is not happy. He therefore devises a new artifact called 'equally distributed equivalent income' and suggests a new utility function called iso-elastic marginal utility function.

4) Sen index is different from the Atkinson's index in one respect that he opts for a social welfare function, which has as individual incomes its arguments and is symmetric and quasi-concave. See Sub-section 11.5.3.

5) See Sub-section 11.5.5.

Appendix

AXIOMS OF INEQUALITY MEASURES

For any statistical measure, some of the desirable properties that are described in standard textbooks are (i) simplicity of comprehension, (ii) ease of computation, (iii) range of variation, and (iv) amount of information needed. However, we discuss below only those properties, which are peculiar to the measures of economic inequality.

We are often faced with situations where we have to compare two distributions with the help of an index with regard to their level of inequality. The two distributions may belong to two different countries at a point of time, to a country at two points of time, or to two situations—say, one before tax and the other after tax or before and after interpersonal transfers etc.

People have set up some intuitively appealing properties in order to judge the efficacy of an inequality index. The first set of properties was given as 'principles' by Dalton (1920). Today, in literature, they are known as axioms. We propose to discuss some common axioms. It may be pointed out at the outset that these axioms almost ignore the question whether inequality is an issue, which matters more (or less) in an affluent society or in a poor one. The whole discussion will assume that all incomes are positive though we know for sure in case of business failure or crop failure, incomes can well be negative or zero.

1) **Axiom of Scale Independence**

If there are two distributions of equal size such (N) that each element of one distribution is a multiple θ of the corresponding element of the other distribution, i.e.,

$x_1^2 = \theta x_1^1$ i=1,2,…,N,

then the numerical magnitudes of inequalities of both the distributions should be the same, i.e.,

$$I(x_1^1, x_2^1, ..., x_N^1) = I(x_1^2, x_2^2, ..., x_N^2)$$

where the inequality measure I is shown as a function of the distribution

$(x_1, x_{2,} ...., x_N)$.

Obviously, it also satisfies the idea that the level of inequality should not change when the scale of measurement changes, say, from rupees to paise or bushels to quintals.

It does also mean that equal proportionate additions to all incomes would not change the level of inequality for

$x_i(1 + \lambda) = \theta x_i$ i =1,2,….,N,

The proportionate addition could even be negative. Thus, it is a question of shares in the cake, not the size of the cake. It is very obvious that an inequality

measure is defined in terms of shares s_i because a proportionate change in all incomes leaves the shares unchanged.

However, this axiom goes against Dalton's principle of proportionate additions to income, which stated that equal proportionate additions (subtractions) should diminish (increase) the level of inequality. Perhaps, Dalton could not see that, equal proportionate addition is theoretically, equivalent to change in the scale of measurement. It should so happen in the case of a measure of relative dispersion is obvious enough.

The axiom covers the cases of proportionate taxation/subsidies. It may be noted that such additions do not change individual (class) shares of total income and, therefore, the Lorenz curve remains unchanged. All measures based on the Lorenz curve shall therefore satisfy this axiom.

It should not mean that change in the size of cake is immaterial. In the social welfare, size of cake and distribution of cake both matter. It is only in a limited context of measurement of inequality that this property is considered desirable.

Lorenz (1997) mentions an objection raised against this axiom in terms of non-proportionate increase in well-being of different income holders, which means diffusion of well-being when incomes increase but concentration when incomes decrease. Thus, this idea existed much before Dalton mentioned it. One could easily see that this objection incorporates the idea of diminishing marginal utility. The true province of the axiom then is the unit of measurement.

2) **Axiom of Population Size Independence**

The level of inequality remains unaffected if a proportionate number of persons is added to each income level.

This suggests that the magnitude of inequality in the distribution of the cake should depend on the relative number of receivers with different levels of income. If we merge two economies of identical distributions of the same size *N*, then, in the consequent economy of size *2N*, there shall be the same proportion of the merged population for any given income. Such replications will leave the inequality level unchanged. The axiom is also known as the Principle of Population Replication.

This exactly corresponds to Dalton's principle of proportionate additions of persons. Since the Lorenz curve remains unchanged so long as proportions of people in each class remains the same, the measures based on the Lorenz curve would satisfy this axiom.

Let us have however a counter-intuitive example. Let us a have two-person world in which one person is having no income and the other having is having all. Let us replicate the economy. Now there is a four-person world in which two are sharing destitution with zero income but the other two are sharing positive income equally. Earlier there was no equality; now each 50 per cent of population is sharing the income equally. So, some scholars do not accept it.

3) **Axiom of Equal Income Addition**

If the distribution 2 x_i^2,i=1,2,….,N is obtained by addition of equal amount d (say, through pension) to each element of distribution 1 x_i^1, i= 1,2,…N, i.e.,

$$x_i^2 = x_i^1 + d$$

then inequality level of distribution 2 should be lower than that of distribution 1 i.e.,

$$I(x_1^2, x_2^2, ..., x_N^2) < I(x_1^1, x_2^1, ...x_N^1)$$

naturally, subtractions (say, taxation) of equal amount from each income would reverse the inequality sign. It can be noted that in the former situation, the shares of the poorer persons increase and in the latter, they decrease. This axiom exactly corresponds to Dalton's principle of equal additions to incomes.

Now, we propose to discuss two very important axioms relating to transfer of an income from a person to another when other things remain the same. The former may be called Pigou-Dalton condition and the latter, Sen condition.

4) **First Axiom of Income Transfer (Pigou-Dalton Condition)**

If an equal transfer from a richer person to a poorer person takes place, then the level of inequality is strictly diminished, provided that the equalizing transfer amount is not more than the difference between two incomes involved. Any number of such transfers taking place between any two consecutive income units will not cause any change in the ranking of income units and therefore such a process of transfers may be called the rank-preserving equalization.

This axiom requires an inequality measure to be sensitive to transfers at all levels of income and, thus, at least a function of all incomes.

This axiom corresponds to Dalton's principle of income transfer. Dalton (1920) argued that an inequality measure must have this minimal property. Since in this context Pigou's contribution (1912) is found significant, Sen (1973) designated this axiom as Pigou-Dalton condition. Following him, a number of contributors in the field have given it the name of 'P-D condition'.

Most of the indices, barring relative range and relative mean deviation, pass this test. This axiom is also known as weak transfer axiom because it suggests the direction but not the magnitude of change in the level of inequality.

5) **Second Axiom of Income Transfer (Sen Condition)**

If we consider two transfers, one at a time, at different points of scale, then the transfer at lower end of scale should have greater impact than its counterpart at higher end of the scale. According to Sen, (1973), the impact on the index should be greater if the transfer takes place from a person with an income level of, say Rs.1000 to someone with Rs.900 than a similar transfer from a man with Rs.1000100 to someone with Rs.1000000.

We may see many measures do not satisfy any of the two conditions of transfer and some satisfy only the first one. Those that satisfy the second transfer axiom automatically satisfy the first transfer axiom.

6) **Axiom of Symmetry**

If distribution $(x_1^p, x_2^p, ..., x_N^p)$ were a permutation of distribution $(x_1, x_2, ..., x_N)$, then the inequality level of both the distributions would be the same.

This implies that if two persons interchange their income positions, inequality measure does not change. Thus the axiom ensures impartiality between individuals for non-income characteristics: The evaluator does not distinguish between Amar, Akabar and Anthony; nor between Shiela and Peter; or between Mr. Pygmy and Ms. Dwarfy. Further, it means that the inequality depends only on the frequency distribution of incomes.

7) **Axiom of Interval**

The inequality measure should lie in the closed interval of (0,1).

The measure is supposed to assume the value of zero when all incomes are equal, which means when all persons have equal income and the value of unity when only one individual gets all the income (and other have zero incomes, not negative incomes).

Most people tend to agree with the axiom. A few, notably Theil (1967) and Cowell (1995), disagree. They hold that the situation of one person grabbing all the income in a society of 2 persons cannot be described by the same level of inequality as that of one person doing so in a society of 2 crore persons. It would not be easy to assert that in the case of 2-person society the level of inequality is unity when one person has all the income and the other has none. In one case, 50 percent population is having positive income, in the other only 0.00000005 percent. Some people therefore qualify the axiom by saying that when one person gets all the income the measure approaches unity in the limit as the number increases.

When a measure has a finite maximum, it is easy to transform such an index into the one, which has maximum value 1. Most measures, though not all, have zero as their minimum value. But question that Cowell (1995) raises is that there are many ways in which the measure could be transformed so that it lies in the zero-to-one range but each transformation has different cardinal properties.

8) **Axiom of Decomposability**

Suppose population can be sub-divided into several groups and an over-all index of inequality was a function of group-wise indexes and if the population mean can be expressed as weighted average of group means, the population index of inequality can be regarded as decomposable. The groups might be defined as comprising of people in different occupations, residents of different areas, with different religious or educational backgrounds etc.

However we find a lot of overlapping in these groups. This leads sum of the weights to differ from unity.

Not all indices are found to be decomposable. Gini coefficient, a very popular measure is decomposable only if the constituent groups are non-overlapping. Cowell (1995) has conducted a beautiful experiment. First, he computes four inequality measures for two distributions of same size and same mean-each divided into two groups of equal size in a manner that there is no overlapping:

Population A: (60,70,80), (30,30,130)

Population B: (60,60,90), (10,60,120)

Now, it is found that the group means and population means in two distributions are the same and group inequalities in B are higher than their counterparts in A. But when we compute overall inequalities, one of the measures suggests that the magnitude of inequality in B is lower than that in A. And the measure used is Gini, which is very popular among economists. As he says, 'strange but true'. If the component inequality magnitudes are higher and the weights are the same, how could overall measure be lower? It is therefore impossible to express overall inequality (change) as some consistent function of inequality change in the consisted groups.

These are all intuitively appealing axioms. There does remain scope for formulating other axioms. In the literature on poverty measurement one finds a plethora of axioms developed by a number of contributors working in that areas. But we shall be content with these only.

UNIT 12 CONSTRUCTION OF COMPOSITE INDEX IN SOCIAL SCIENCES

Structure

12.0 Objectives

12.1 Introduction

12.2 Composite Index: The Concept

12.3 Steps in Constructing Composite Index

12.4 Dealing with Missing Values and Outliers

12.5 Methods to Construct Composite Index

- 12.5.1 Simple Ranking Method
- 12.5.2 Indices Method
- 12.5.3 Mean Standardization Method
- 12.5.4 Range Equalization Method

12.6 Principal Component Analysis (PCA)

- 12.6.1 Conducting PCA and Analyses of Results
- 12.6.2 Use of Output Indicators
- 12.6.3 Use of Weight
- 12.6.4 Limitations of Principal Component Analysis

12.7 Merits and Limitations of Composite Index

12.8 Let Us Sum Up

12.9 Exercises

12.10 Some Useful Books/References

12.11 Answer or Hints to Check Your Progress Exercises

12.0 OBJECTIVES

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

- state the concept of composite index;
- describe the process of constructing the composite index;
- explain the various methods of composite index;
- discuss the merits and limitations of composite index; and
- learn how to interpret the results derived from composite indexes.

12.1 INTRODUCTION

In social sciences research, many a times the complex social and economic issues like child deprivation, food security, human well-being, human development etc. are difficult to measure in terms of single variable. The reason being that such issues have several dimensions and indicators. For example, it is difficult to explain the status of development of a district in terms

of a single variable because development is reflected in terms of several indicators. Some of such variables/indicators are quantitative type while others are of qualitative nature. In such situations, Composite index plays an important role to express the single value of several inter-dependent or independent variables. Further, the composite index makes it possible to compare the performance among region/states or districts etc. Hence, composite indexes are being recognized as a useful tool for policy analysis. Composite indexes can also be resorted to make comparison among different regions/sectors where wide range of variables are used.

In this Unit, we shall therefore discuss the concept of composite index, the process of their construction, various methods, their uses, and limitations. The study of the unit will also enable you to learn how to interpret the results. Let us begin to explain the concept of composite index.

12.2 COMPOSITE INDEX: THE CONCEPT

A composite index is an expression of a single score made by combination of different scores to measure a given variable or a group of variables. It expresses quantity or place (position) of multi facet aspects of a concept. The UNDP (2005) has explained that 'a composite index expresses a quantity or a position on a scale of qualitative multi-faceted aspects….. which is relevant for information of the society'. An index can be a combination of independent indicators, or the average of an accumulation of selected indicators. The index represent specific concept or highlighting specific sector or areas like status of dalit (dalit deprivation index), food situation (food security/insecurity index), status of child (child deprivation/development index), quality of human development (human development index) etc.

The index that we construct is the outcome of some unidirectional variables or indicators. If an index is constructed by taking positive indicators, the higher value of index implies higher development and lower values imply lower development. For example, in case of index related to child development, if the variables are positive directional, the final index can be termed as 'child development index'. On the other hand, if the variables are negative directional, it is called 'Child Deprivation Index'. Let us take an example that child mortality is a negatively directional variable whereas percentage of children immunized is a positively directional variable. Such index is very useful in case of qualitative data. An index is more robust than a single indicator or variable.

The choice of indicator is a big challenge for researchers. The major issues in identifying measurable indicators are: whether data are available or not, whether we have reliable data, whether the data are cross-section or time series, and minimization of double counting raised from overlap or redundancy. Again if the variables to be chosen are easy to understand, they are more acceptable to a wider audience.

12.3 STEPS IN CONSTRUCTING COMPOSITE INDEX

The following steps are required to follow in construction of composite index:

Step	Why it is needed
1. Theoretical framework A theoretical framework need to be developed because it provides the basis for the selection and combination of variables into a meaningful composite indicator under a fitness-for-purpose principle (involvement of experts and stakeholders is envisaged at this step).	• To get clear understanding and definition of the multidimensional phenomenon to be measured. • To structure the various sub-groups of the phenomenon (if needed). • To compile a list of selection criteria for the underlying variables, e.g., input, output, process.
2. Data selection Indicators should be selected on the analytical soundness, measurability, country coverage, and relevance of the indicators to the phenomenon being measured and relationship to each other. The use of proxy variables should be considered when data are scarce (involvement of experts and stakeholders is envisaged at this step).	• To check the quality of the available indicators. • To discuss the strengths and weaknesses of each selected indicator. • To create a summary table on data characteristics, e.g., availability (across country, time), source, type (hard, soft or input, output, process).
3. Imputation of missing data Consideration should be given to different approaches for imputing missing values. Extreme values should be examined as they can become unintended benchmarks.	• To estimate missing values. • To provide a measure of the reliability of each imputed value, so as to assess the impact of the imputation on the composite indicator results. • To discuss the presence of outliers in the dataset.
4. Multivariate analysis Should be used to study the overall structure of the dataset, assess its suitability, and guid subsequent methodological choices (e.g., weighting, aggregation).	• To check the underlying structure of the data along the two main dimensions, namely individual indicators and countries (by means of suitable multivariate methods, e.g., principal components analysis, cluster analysis). • To identify groups of indicators or groups of countries that are statistically "similar" and provide an interpretation of the results. • To compare the statistically-determined structure of the data set to the theoretical framework and discuss possible differences.
5. Normalisation Should be carried out to render the variables comparable	• To select suitable normalization procedure(s) that respect both the theoretical framework and the data properties.

	• To discuss the presence of outliers in the dataset as they may become unintended benchmarks. • To make scale adjustments, if necessary. • To transform highly skewed indicators, if necessary.
6. Weighting and aggregation Should be done along the lines of the underlying theoretical framework.	• To select appropriate weighting and aggregation procedure(s) that respect both the theoretical framework and the data properties. • To discuss whether correlation issues among indicators should be accounted for. • To discuss whether compensability among indicators should be allowed.
7. Uncertainty and sensitivity analysis Should be undertaken to assess the robustness of the composite indicator in terms of e.g., the mechanism for including or excluding an indicator, the normalization scheme, the imputation of missing data, the choice of weights, the aggregation method.	• To consider a multi-modeling approach to build the composite indicator, and if available, alternative conceptual scenarios for the selection of the underlying indicators. • To identify all possible sources of uncertainty in the development of the composite indicator and accompany the composite scores and ranks with uncertainty bounds. • To conduct sensitivity analysis of the inference (assumptions) and determine what sources of uncertainty are more influential in the scores and/pr ranks.
8. Back to the data Is needed to reveal the main drivers for an overall good or bad performance. Transparency is primordial to good analysis and policymaking.	• To profile country performance at the indicator level so as to reveal what is driving the composite indicator results. • To check for correlation and causality (if possible). • To identify if the composite indicator results are overly dominated by few indicators and to explain the relative importance of the sub-components of the composite indicator.
9. Links to other indicators Should be made to correlate the composite indicator (or its	• To correlate the composite indicator with other relevant measures, taking into consideration the results of sensitivity analysis.

dimensions) with existing (simple or composite) indicators as well as to identify linkages through regressions.	• To develop data-driven narratives based on the results.
10. Visualisation of the results Should receive proper attention, given that the visualization can influence (or help to enhance) interpretability.	• To identify a coherent set of presentational tools for the targeted audience. • To select the visualization technique which communicates the most information. • To present the composite indicator results in a clear and accurate manner.

Source: OECD (2008), 'Handbook on Constructing Composite Indicators Methodology and User Guide'

Caution on choosing variable

1) Whenever we choose any variable or a particular dimension for the index, we have to justify the inclusion of the variable into the index. This justification should come from empirical evidences or policy based research studies or from theoretical explanation.

2) If the variable is not unidirectional, the entire variables used should be converted to unidirectional. For example, in construction of the food security index, two variables like 'proportion of agricultural worker to total workers' and per capita value of agricultural output' is used for index. Here the first variable is a negatively directional whereas the second variable is positively directional. In this case we have to convert the entire variables into either positive direction or negative direction. If we want to convert this to positive direction, the first variable which has a negative direction should be deducted from 100. On the other hand, if we want to convert the entire variables into negative direction, we have to work out the reciprocal of per capita value of agricultural output.

12.4 DEALING WITH MISSING VALUES AND OUTLIERS

After selecting the indicators, you have to have a clear idea on missing values for each selected variable. Data can be missing in a random or non random fashion. In such situation, easy solution is to drop cases for which data is missing. However, before dropping the variable, you have to see the number of cases for which values are missing, because exclusion of such households for which data is missing could significantly lower sample size. Further, deleting such household may lead to some biases also. For example, in livelihood study of households, if there is a missing value, we have to find out the frequency of missing value because in many cases, the chances of missing value is high for lower economic class as compared to upper economic class. In such cases if we delete the households having missing value, it may result biases towards upper economic classes of household. In case, inclusion or exclusion of households having missing value put a little impact on the final result, we can delete such cases.

The second solution is to impute the missing value by applying some methodology or logic. But use of imputation is more common in academic indices and datasets. Here you can try to impute the missing value which should be as accurate as possible..For example, in calculating the consumption expenditure of 50 households of which about 5 cases are missing. In this case in imputing we can substitute the average value of consumption expenditure and replace the missing value by the averages. But for a more accurate estimation we can group theses 50 households on the basis of value of assets holding and see to which asset category these 'MPCE missing' households belonged to. Then we can substitute the average consumption expenditure of the respected asset group.

Such approach has been used in a number of more recent indices such as the Corruptions Perceptions Index (CPI) developed by Transparency International and the World Bank's Worldwide Governance Indicators (WGI). Such approaches carry a dual advantage. They allow scores to be estimated for a maximal number of countries, and can use a broader range of indicators to triangulate indices for nebulous constructs.

Many a time,outliers may also disturb the analysis. For example if the income of four persons is 18000, 17000, 18500 and 19000 then the average is 18125. If we include the income of fifth person as 60000 then the average turns out to be 26500. In such matters, you have to drop the cases with high extreme value. Both in case of missing value and outlier you are expected to document and explain the selected imputation procedures and the results in detail.

Check Your Progress 1

1) What do you understand by the term Composite Index?

……..

……..

……..

……..

……..

2) Give some examples of Composite Index.

……..

……..

……..

……..

……..

3) List the steps involved in construction of Composite Index.

……..

……..

……..

……..

……..

4) State the alternative methods to deal with the missing values of the selected variables.

……..

……..

……..

……..

……..

……..

……..

……..

12.5 DIFFERENT METHODS OF COMPOSITE INDEX

The various methods are used in construction of composite index. These include: Simple Ranking Method, Indices Method, Mean standardization method, Range Equalization Method, and Principal Component Analysis Method. Let us discuss one by one with simple examples.

12.5.1 Simple Ranking Method

Rank Method is one of the simplest methods to analyse the status of a region/district/state/country. In this method, the first step is to arrange/convert the variable into unidirectional for each district or state. Let us take an example of development status of different districts in Orissa state. First, we have to convert the relevant variables in either positive or negative direction. For example, the variable says 'the proportion of agricultural labour is negatively associated with the development of a region. Hence we have to convert this variable to positive direction by deducting this variable from 100 and change the labeling of variable as 'proportion of other than agricultural labour to total labour'. Higher the value of this variable, higher the development of the district or region. Depending on the value, each variable was ranked in the similar manner. Highest rank (1^{st}) was given to the variables with high value and vice-a-versa. Alternatively, we can do reverse ranking of the variable i.e. highest rank is given to the variables with lowest value. The individual ranks are added to get the total rank value for the district. This has been illustrated in the example given in table 12.1. Our objective is to find out most backward tribal district in Odisha (having more than 50 per cent of tribal population to total population). A total of 11 districts qualify for tribal dominant area having 50 per cent or more proportion of population. We have selected 10 variables for identifying most backward tribal areas. Each of the variables for 11 tribal dominated districts was ranked according to total value of variable. The individual rank of all the variables of all the districts is given in table 12.2. As all the variable are unidirectional, rank 1 is given to district having higher value and vice-a-versa. After doing so we can add together the value of all the variables and find out the average rank for each district (Sum of total rank divided by number of variable i.e. 10). The average value of variable shows the status of district on development indicators.

Table 12.1: Districtwise Development Indicators of Tribal Concentrated Districts in Odisha

District	% of other than agricultural labour to all labour	% of net irrigated area to total sown area	Per capita value of agrictural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	% of household getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to Phcs	Average casual wage rate
Actual Value										
Malkangiri	73	31	1304	201	18	67	82	21	16	36
Rayagada	50	23	479	201	18	67	78	34	13	37
Sundargarh	61	19	537	280	43	58	57	38	28	32
Nabarangapur	46	6	839	201	18	66	80	45	28	37
Kandhamal	62	14	410	201	33	68	32	21	19	36
Koraput	55	31	611	201	16	67	67	22	22	40
Mayurbhanj	60	30	572	280	35	62	44	42	21	31
Gajapati	52	25	529	331	24	77	43	29	20	34
Sambalpur	62	34	1075	280	50	62	56	31	13	39
Kendujhar	60	23	537	280	44	46	52	46	15	34
Jharsuguda	68	20	662	280	54	43	63	47	32	39
Average	59	23	687	249	32	62	60	34	21	36

Table 12.2: Districtwise Development Scenario of Tribal Orissa by Simple Ranking Method

District	% of other than agrictural labour to all labour	% of net irrigated area to total sown area	Per capita value of agrictural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	% of household getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Average Rank/indices
Malkangiri	1	2	1	7	8	5	1	10	8	7	5.0
Rayagada	10	7	10	7	9	3	3	6	11	4	7.0
Sundargarh	5	9	7	2	4	9	6	5	3	10	6.0
Nabarangapur	11	11	3	7	10	6	2	3	2	5	6.0
Kandhamal	3	10	11	7	6	2	11	11	7	6	7.4
Koraput	8	3	5	7	11	4	4	9	4	1	5.6
Mayurbhanj	7	4	6	2	5	7	9	4	5	11	6.0
Gajapati	9	5	9	1	7	1	10	8	6	8	6.4
Sambalpur	4	1	2	2	2	8	7	7	10	3	4.6
Kendujhar	6	6	8	2	3	10	8	2	9	9	6.3
Jharsuguda	2	8	4	2	1	11	5	1	1	2	3.7

By this method, the most developed tribal district is Jharsuguda and most backward district is kandhamal.

12.5.2 Indices Method

The indices method is another simple method for calculating the status of development of an area/district or state. Like rank method, here also, at the first stage the variables are converted into one direction. After converting the variable into one (positive or negative) direction, we calculate the index. In this method we have to convert the district figures based on the average figure for the entire 11 districts as 100. Let us take an example of table 12.1. If the proportion of other than agricultural labour to total labour in malkangiri district is 73 and the average of that variable for all the 11 districts is 59, then the indices of that variable for Malkangiri district will be: $\frac{100}{59} x73 = 124$

In the same manner, we can find out the indices for all the variables for 11 districts as is given in Table 12.3. The final index of development for the districts can be obtained by taking the arithmetic mean for all the indicators (given in last column of table 12.3). After working out the average indices value for all the districts, we can rank all the districts. The district with highest indices will be most developed and the district with lowest indices will be least developed.

Table 12.3: Districtwise Development Scenario of Tribal Orissa by Indices Method

District	% of other than agrictural labour to all labour	% of net irrigated area to total sown area	Per capita value of agrictural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	% of households getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Average Rank/indices
Malkangiri	124	133	190	81	57	108	138	62	78	99	107.0
Rayagada	85	99	70	81	57	108	131	98	64	103	89.6
Sundargarh	104	83	78	113	135	93	96	111	135	90	103.6
Nabarangapur	78	25	122	81	56	107	135	132	137	102	97.4
Kandhamal	106	59	60	81	102	109	54	60	92	100	82.3
Koraput	93	133	89	81	49	108	113	65	108	113	95.1
Mayurbhanj	101	129	83	113	109	101	74	123	103	86	102.2
Gajapati	89	106	77	133	76	123	73	86	95	95	95.3
Sambalpur	105	147	157	113	155	100	95	90	65	108	113.3
Kendujhar	102	99	78	113	136	74	88	136	71	94	99.1
Jharsuguda	114	88	96	113	167	69	105	138	153	109	115.3

In our example Kandhamal is the most backward with indices (82.3) and Jharsuguda (115.3) is the most developed district.

Let us make a comparison between the results obtained by rank method and indices method. Results in Table 12.4 shows that Jharsuguda, Sambalpur, and Malkangiri are most developed in both the methods whereas Raygada and Kandhamal are most backward in both the methods. While comparing both rank and indices method, we can also run correlation between the two set of indices. If the correlation is high then we can say that the finding in both methods is almost similar.

Table 12.4: A Comparison of Simple Ranking Method and Indices Method

District	Status of District in Rank Method	District	Status of District in indices Method
Jharsuguda	3.7	Jharsuguda	115
Sambalpur	4.6	Sambalpur	113
Malkangiri	5.0	Malkangiri	107
Koraput	5.6	Sundargarh	104
Sundargarh	6.0	Mayurbhanj	102
Nabarangapur	6.0	Kendujhar	99
Mayurbhanj	6.0	Nabarangapur	97
Kendujhar	6.3	Gajapati	95
Gajapati	6.4	Koraput	95
Rayagada	7.0	Rayagada	90
Kandhamal	7.4	Kandhamal	82

Note: Correlation of status of district in both rank and indices method is 0.927

12.5.3 Mean Standardization Method

Mean standardization Method is another simple method used both as a process of normalization and also a composite index. In this method we normalize the value of each variable and than work out the average of the normalized value for all the variables. The average of normalized value will be the composite index value. The normalization is done by dividing the actual value of variables by their respective means. Let us take an example from Table 12.1. For example the standardized value of the variable '% of other then agricultural labour to all workers' for Malkangiri district can be found out by

Normalized value MS Method=Actual value (73)/Mean value (59) =1.237

In the same process we can find out the normalized value of the entire variable for 11 selected districts. The standardized value of each selected indicator for different district is given in Table 12.5. The composite index value of each district given in the last column of Table 12.5 has been worked out by taking average of all the 10 selected indicators (row vector of the each district).

Table 12.5: Index Value in Mean Standardization Method

District	Percentage of other than agricultural labour to all labour	Percentage of net irrigated area	Per capita value of agricultural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	Percentage of household getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Composite Index
Malkangiri	1.237	1.332	1.899	0.808	0.561	1.079	1.379	0.614	0.775	1.003	1.069
Rayagada	0.847	0.988	0.697	0.808	0.561	1.079	1.312	0.995	0.630	1.030	0.895
Sundargarh	1.034	0.816	0.782	1.126	1.340	0.934	0.959	1.112	1.357	0.891	1.035
Nabarangapur	0.780	0.258	1.222	0.808	0.561	1.063	1.346	1.316	1.357	1.030	0.974
Kandhamal	1.051	0.602	0.597	0.808	1.028	1.095	0.538	0.614	0.921	1.003	0.826
Koraput	0.932	1.332	0.890	0.808	0.499	1.079	1.127	0.644	1.066	1.114	0.949
Mayurbhanj	1.017	1.289	0.833	1.126	1.091	0.999	0.740	1.229	1.018	0.863	1.020
Gajapati	0.881	1.074	0.770	1.331	0.748	1.240	0.723	0.848	0.969	0.947	0.953
Sambalpur	1.051	1.461	1.565	1.126	1.558	0.999	0.942	0.907	0.630	1.086	1.132
Kendujhar	1.017	0.988	0.782	1.126	1.371	0.741	0.875	1.346	0.727	0.947	0.992
Jharsuguda	1.153	0.859	0.964	1.126	1.683	0.693	1.060	1.375	1.551	1.086	1.155

12.5.4 Range Equalization Method

Range Equalization (RE) method otherwise known as max-min approach is adopted by UNDP in computation of Human Development Index. Under this approach, an index is constructed for each variable by applying Range Equalization formula derived by UNDP. This is worked out by subtracting an indicator's minimum value from each observation and then dividing it by its range.

$$\text{RE Index} = \frac{X_i - Min\ X}{Max\ X - Min\ X}$$

where X_i Value of the variable, min X- Minimum value of X in the scaling , max X- Maximum value of X in the scaling. The RE index is also a normalization technique. Without it, a composite index can be biased towards an indicator with very high range. For example the per capita value of agricultural output which is measured in rupees ranges from Rs. 410/- to Rs. 1304/-. On the other hand, the variable proportion of female literacy rate is measured in percentage. Different variables measured in different units sometimes give upward or downward biases. To ensure the index value biasfree, we convert all the variables into equal scaling from 0 to 1.

In RE method, as a first step we undertake scaling exercise. In undertaking the scaling procedure, desirable norms are followed for each indicator. In some cases, scaling of indicators is self-selected, while in others element of value judgment is involved. This scaling exercise is called goal post where we identify the maximum and minimum goal. The goalpost basically visualizes the extent of minimum value and maximum value in the future time period (say next 5 years). The scaling norm that we have adopted is given in Table 12.6. In the table the third and fourth column shows the maximum and minimum district value of selected 10 indicators. The first and second column is the maximum and minimum goal post.

Table 12.6: Construction of Food Security Radar

Variable	Description of Variables	Goalposts		District Value	
		Minimum	Maximum	District Minimum	District Maximum
Oth_agl	Percentage of other than agricultural labour to all labour	30	85	46	73
Irr	Percentage of net irrigated area to net cropped area	2	55	6	34
pcvao	Per capita value of agricultural output	200	2500	410	1304
Mpcce_ia	Inequality adjusted per capita consumption expenditure	150	450	201	331
Lit_f	Female Literacy (adult) rate	10	70	16	54
Wfpr_f	women work force participation rate	30	85	43	77
Hhsdw	Percentage of household having access to safe drinking water	20	90	32	82
Paved_r	Percentage of villages having access to paved road	10	60	21	47
v_phcs	Percentage of villages having access to Primary health centre	10	50	13	32
Wage_c	average Casual wage rate	25	70	31	40

Based on the goalpost, we normalize the value of the variable for all the districts and all the indicators. In our example, the values have been presented in Table 12.7.

Table 12.7: Index Value in Range Equalization Method

District	Percentage of other than agrictural labour to all labour	Percentage of net irrigated land to net sown area	Per capita value of agrictural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	Percentage of households getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate	Composite Index
Malkangiri	0.782	0.547	0.480	0.170	0.133	0.673	0.886	0.220	0.150	0.244	0.429
Rayagada	0.364	0.396	0.121	0.170	0.133	0.673	0.829	0.480	0.075	0.267	0.351
Sundargarh	0.564	0.321	0.147	0.433	0.550	0.509	0.529	0.560	0.450	0.156	0.422
Nabarangapur	0.291	0.075	0.278	0.170	0.133	0.655	0.857	0.700	0.450	0.267	0.388
Kandhamal	0.582	0.226	0.091	0.170	0.383	0.691	0.171	0.220	0.225	0.244	0.300
Koraput	0.455	0.547	0.179	0.170	0.100	0.673	0.671	0.240	0.300	0.333	0.367
Mayurbhanj	0.545	0.528	0.162	0.433	0.417	0.582	0.343	0.640	0.275	0.133	0.406
Gajapati	0.400	0.434	0.143	0.603	0.233	0.855	0.329	0.380	0.250	0.200	0.383
Sambalpur	0.582	0.604	0.380	0.433	0.667	0.582	0.514	0.420	0.075	0.311	0.457
Kendujhar	0.545	0.396	0.147	0.433	0.567	0.291	0.457	0.720	0.125	0.200	0.388
Jharsuguda	0.691	0.340	0.201	0.433	0.733	0.236	0.614	0.740	0.550	0.311	0.485

After calculating the index of each variable, we have averaged them to give each of the five dimensions of food security. The composite food security index is again derived by averaging the five dimensions.

The normalized value for the variable percentage of other than agricultural labour to total worker' for Malkangiri district is found out in the following manner.

$$\frac{\text{(Actual value (73) – Minimum goalpost (30))}}{\text{(Maximum goalpost (85) – Minimum goalpost (30))}} = 0.782$$

Likewise we have converted and worked out the normalized value of all the variables and for all the districts shown in Table 12.7.

Comparing MS Method and RE Method

The composite index value worked out by both the RE and MS methods can be compared and analyzed (Table 12.8). We can also examine correlation between two indices. Here the correlation between final index value worked out by both the methods is very high (0.991).

Table 12.8: Comparison between RE and MS Methods

District	Composite Index Value by RE method	Composite Index Value by MS method
Jharsuguda	1.155	0.485
Sambalpur	1.132	0.457
Malkangiri	1.069	0.429
Sundargarh	1.035	0.422
Mayurbhanj	1.020	0.406
Kendujhar	0.992	0.388
Nabarangapur	0.974	0.388
Gajapati	0.953	0.383
Koraput	0.949	0.367
Rayagada	0.895	0.351
Kandhamal	0.826	0.300

Note: Correlation of index value is 0.991

The high degree of the correlation between the values of composite index computed by applying both methods separately indicate the high degree of homogeneity between the two approaches. However, the Range equalization method is preferred because it accounts for wider variations and strong correlations to the PCA composite.

One important point should be kept in mind that the index value as discussed above are based on equal weight. We can also give weights to variables/ components discussed in section 12.5.

Check Your Progress 2

1) State the procedure to workout Composite Index by way of Simple Ranking Method.

   ……

   ……

   ……

2) How will you compute the final value of Composite Index by Indices Method?

   ……

   ……

   ……

3) What is the Distinction between Range Equalization Method and Mean Standarisation Method?

   ……

   ……

   ……

   ……

12.6 PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is one of the important composite index method used to analyse the various problems in social sciences. PCA is a mathematical technique that transfers a number of correlated variables into smaller number of uncorrelated variables called the principal components. For our analysis we have an array of variables which may have high correlation with each other. In case where there is high relation between variables, the PCA is a suitable technique where the correlation between different components is low.

Categorical data are not suitable for PCA analysis because in such cases categories converted into quantitative scale has no meaning. To avoid this type of problems, we should recode these variables into binary variables. Let us take an example of social category of household that are Scheduled caste (1), Scheduled tribe (2), Other backward caste (3) and General (4). In PCA analysis such type of variables have no meaning. Such type of categorical variables can be converted into a bianary variable. For example if we want to study the dalit status, variable can be converted to a bianary variable as dalit (code 2) and non-dalit (code 2, 3, and 4). Likewise, if we want to study the impact of reservation for other backward castes we can categorize as OBC (code 3) and other than OBC (code 1, 2 and 4).

The PCA is a data reduction technique. Under this technique, the original data set is transformed into a new set of uncorrelated variables called principal components. PCA reduces the number of variables in a data set to smaller number of dimension/s. In a series of variables if a_{mn} is the weight of m_{th} principal component and nth variable then the matrix will be:

$$\begin{Bmatrix} PC_1 \\ PC_2 \\ PC_3 \\ . \\ . \\ . \\ PC_n \end{Bmatrix} = \begin{Bmatrix} a_{11} & a_{12} & a_{13} & ..... & a_{1n} \\ a_{21} & a_{22} & a_{23} & ..... & a_{2n} \\ . & & & & \\ . & & & & \\ . & & & & \\ a_{m1} & a_{m2} & a_{m3} & ..... & a_{mn} \end{Bmatrix} \begin{Bmatrix} X_1 \\ X_2 \\ X_3 \\ \\ \\ \\ X_n \end{Bmatrix}$$

The matrix can be transformed into equation for principal component i.e.

$$PC_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_n$$
$$PC_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2n}x_n$$
.....
$$PC_n = a_{m1}x_1 + a_{m2}x_2 + ... + a_{mn}x_n$$

The components are ordered so that the first component (PC_1) explains the largest possible amount of variation in the original data, subject to the constraint that the sum of the squared weights of the vectors $(a_{11}^2 + a_{12}^2 + \cdots + a_{1n}^2)$ is equal to one.

In the output the eigen value is one of the important determinant of PCA. The eigen value is a number that tell you how much variance there is in the data set. In other words the eigen value is a number telling us how spread out the data is on the line. The eigen vector with the highest eigen value is therefore the principal component.

As the sum of the Eigen values equals the number of variables in the initial data set, the proportion of the total variation in the original data set accounted by each principal component is given by λ_i/n. The second component (PC_2) is completely uncorrelated with the first component, and explains additional but less variation than the first component, subject to the same constraint. Each subsequent component captures additional dimension in the data and adds smaller and smaller proportion of variation in the original variables.

Before conducting PCA, we should run correlation between variables. If the correlation between two variables is very high, we may remove one of those variables. The reason being that the two variables seem to measure same thing. The higher the degree of correlation among variables, the lower will be the number of components required to capture common information. Sometimes two variables may be combined in some ways (taking average). The PCA can be applied either to the original values of variables or to the normalized values of the variables.

In general, normalization can be done by three methods, i.e (i) by deviation of the variables from their respective means (i.e.); (ii) by dividing the actual values by their respective means; (iii) and deviation of value of a variable from the mean which is then divided by standard deviation {i.e. ()/σ}.We are applying here the second method.

Let us try to apply and analyse PCA by using the database given in Table 12.1. We are applying here the second method for normalization that is found out in Table 12.5 column 2 to 10.

Table 12.9: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.394
Bartlett's Test of Sphericity	Approx. Chi-Square	47.933
	Df	45
	Sig.	.355

When the normalized database is prepared, you have to apply Kaiser-Meyer-Olkin (KMO) technique and Measure of Sampling Adequacy and Bartlett's Test of Sphericity (BTS). In KMO technique, the value varies between 0 and 1 and the value closer to1 is better. However, according to some statisticians, the minimum value should be 0.6. The BTS tests the null hypothesis that the correlation matrix is an identity matrix. Let us remember that identity matrix is a matrice in which the main diagonal elements are 1 and off-diagonal elements are 0.

The above two tests provide the minimum standard before conducting any PCA.

12.6.1 Conducting PCA and Analyzing Results

For the purpose of Principal Component analysis and interpreting its results, SPSS is user friendly software. However, you can try to run the same in STATA also. Let us take our previous example of Table 12.5. In the table, there are 10 standardised variables and we can run our PCA model. Table 12.10 give the first output called 'communalities'

Table 12.10: (Output 1) Communalities: Extraction by Principal Component Analysis.

Variable name	Initial	Extraction
% of other than agricultural labour to all labour	1.000	.721
% of net irrigated area to net sown area	1.000	.748
Per capita value agricultural output	1.000	.710
Monthly per capita consumption expenditure	1.000	.702
Female Literacy rate (adult)	1.000	.876
women work force participation rate	1.000	.858
% of household access to safe drinking water	1.000	.789
% of villages having access to paved road	1.000	.773
% of villages having access to primary health centres	1.000	.572
Average Casual wage rate	1.000	.491

Communalities represent the percentage of variance explained by the extracted components. This explains the proportion of each variable's variation explained by PCA. The short notation of communalities is 'h^2' and defined as sum of squared factor loading. In table 12.10, the first column is the name of variables used in our PCA. The second column is the 'initials'. The initial value of the communalities in PCA is 1. The second column 'extraction' shows the proportion of each variable's variation captured by PCA. This value ranges between 0 and 1. When the value is near to 1 that means the variable is well represented and the reverse for value is closer to '0'.

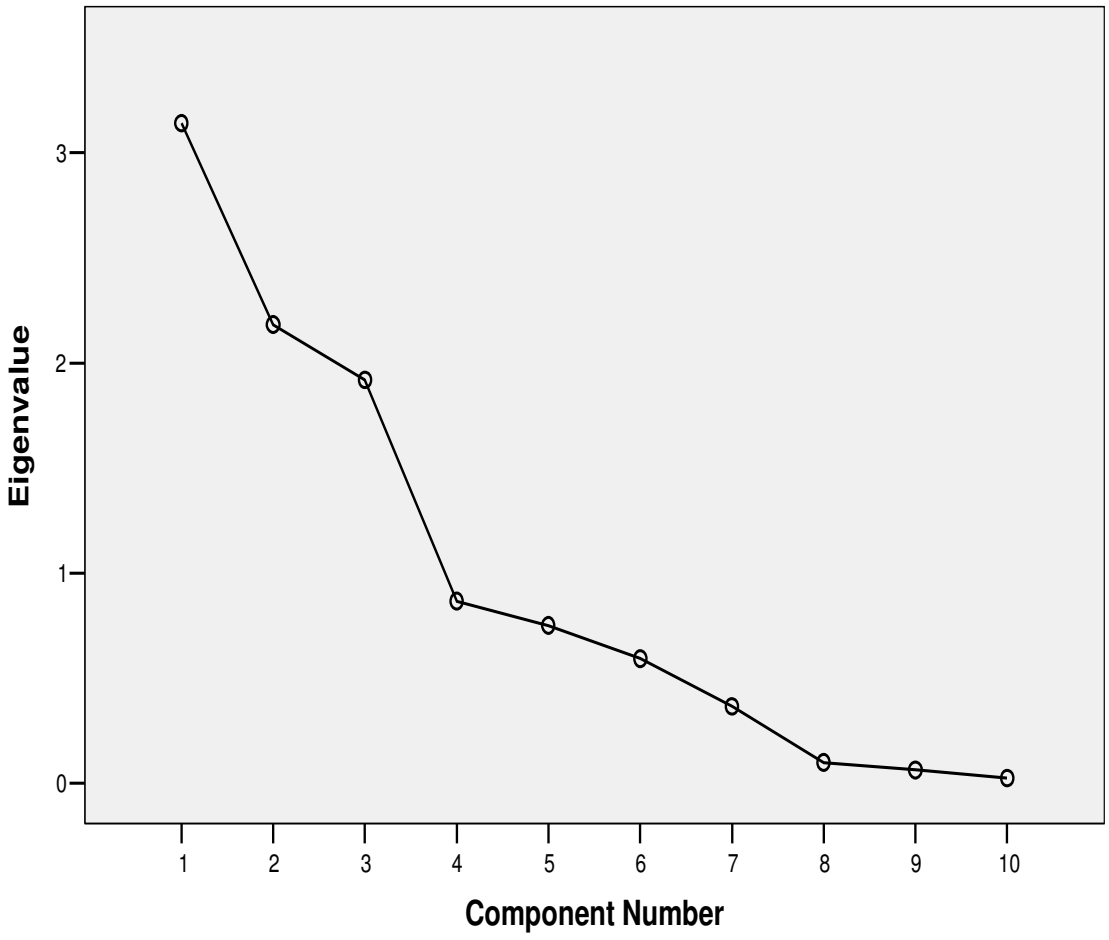
If the communality is low for an item, the reason might be that the item was poorly designed. If the item has very little variance the communality is low. If the different items usually resulting from large positive or negative skewed the communalities is low. We can take an example that if everyone ticks strongly agree, the variation within the variable is low and the communality is low. If the communality is low, we can either remove the item from the analysis to exclude it from any further analyses or we can treat it as a stand alone variable.

Table 12.11: (Output 2) Total Variance Explained: Extraction Method: Principal Component Analysis.

	Initial Eigen values			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.141	31.413	31.413	3.141	31.413	31.413
2	2.182	21.820	53.233	2.182	21.820	53.233
3	1.919	19.195	72.428	1.919	19.195	72.428
4	.866	8.658	81.086			
5	.750	7.502	88.587			
6	.593	5.927	94.514			
7	.365	3.645	98.160			
8	.098	.976	99.135			
9	.062	.625	99.760			
10	.024	.240	100.000			
	10.0					

The Table 12.11 shows 'total variance explained'. The first column shows the number of components. The number of components should be equal to number of variables used for PCA. In our example, we have used 10 variables and hence the total number of component is 10. The second part of the table reflect initial Eigen value which consists of three columns i.e. Total, percentage of Variance, Cumulative percentage. The initial Eigen values are the variances of principal components. Here as we use the standardized values of variable, the variance becomes equivalent to 1 and the total variance is equal to the total number of variables i.e. '10'. The second column 'total' contains the Eigen values. Here it can easily be seen that the Eigen values of subsequent components gradually reduce implying that the successive components add less and less variation. The third column shows the percentage variation on the components and the fourth column explains the cumulative addition of percentage variation of cumulative percentage components. In our example the first component explain 31.4 per cent of total variation and the second component explain 21.8 percent of total variation and so on. The third row of fourth column shows that the first three components explain 72 per cent of total variation.

The second part of the table is 'Extraction Sums of Squared Loadings' which consists of three columns – Total, percentage of Variance and Cumulative percentage. This has actually reproduced the three rows of figures of the first part of table reflecting the components whose eigen value is greater than 1.



Graph 12.1: Scree plot

The 'scree plot' is a graph whose 'X axis' represent the component number and 'Y axis' represents 'Eigen values'. In other words scree plot demonstrates the graph of first two column of 'output 2' table. From the graph one can select the number of components to be taken for analysis. It can be clearly seen from the above graph that from the fourth component onward the line becomes flat indicating that when the successive addition is less and less, the line of the graph becomes more and more flatter.

Table 12.12: (Output 4) Component Matrix(a)

	Component[a]		
	1	2	3
Percentage of other than agricultural labour to all labour	.202	.807	.174
Percentage of net irrigated area to total sown area.	-.185	.793	-.293
Per capita value of agricultural output	-.403	.596	.438
Monthly per capita consumption expenditure	.717	.217	-.378
Female Literacy rate	.837	.406	.105
women work force participation rate	-.721	-.233	-.533
Percentage of households having access to safe drinking water	-.504	-.041	.730
Percentage of villages having access to paved road	.733	-.291	.389
Percentage of villages having access to primary health centre	.441	-.417	.451
Average Casual wage rate	-.430	.145	.535

Extraction Method: Principal Component Analysis.
a 3 components extracted.

Table 12.12 (output 4) shows Component Matrix of first three components. The first column is the name of variable used in PCA. The values in the output table are the component loading which are the correlation between the variables and components. The values in the tables range from -1 to +1. One can note that the sum of square of all three component matrix of a variable is equal to the extraction value of the communalities of that variable. In this example three component has been extracted as the eigen value for these component is greater than 1.

Table 12.13: (Output 5) Reproduced Correlations

	Percentage of other than agricultural labour to all labour	Percentage of net irrigated area to total sown area	Per capita value of agricultural output	Per capita consumption expenditure	Female Literacy rate	women work force participation rate	Percentage of of households getting safe drinking water	Percentage of villages having access to paved road	Percentage of villages having access to primary health centre	Average casual wage rate
Reproduced Correlation										
Percentage of other than agricultural labour to all labour	0.721	0.554	0.477	0.253	0.512	-0.424	-0.007	-0.022	-0.171	0.120
Percentage of net irrigated to total sown area	0.554	0.748	0.419	0.151	0.137	0.102	-0.152	-0.479	-0.543	0.032
Per capita value of agricultural output	0.477	0.419	0.710	-0.324	-0.049	-0.083	0.500	-0.297	-0.228	0.491
Per capita consumption expenditure	0.253	0.151	-0.324	0.702	0.648	-0.366	-0.645	0.314	0.055	-0.479
Female Literacy rate	0.512	0.137	-0.049	0.648	0.876	-0.754	-0.362	0.536	0.247	-0.246
women work force participation rate	-0.424	0.102	-0.083	-0.366	-0.754	0.858	-0.015	-0.668	-0.460	-0.010
Percentage of households having access to safe drinking water	-0.007	-0.152	0.500	-0.645	-0.362	-0.015	0.789	-0.074	0.123	0.603
Percentage of villages having access to paved road	-0.022	-0.479	-0.297	0.314	0.536	-0.668	-0.074	0.773	0.620	-0.144
Percentage of village access to primary health centre	-0.171	-0.543	-0.228	0.055	0.247	-0.460	0.123	0.620	0.572	-0.005
Average Casual wage rate	0.120	0.032	0.491	-0.479	-0.246	-0.010	0.603	-0.144	-0.005	0.491
Residual(a)										
Percentage of other than agricultural labour to all labour		-0.149	-0.039	-0.161	-0.040	0.002	-0.087	-0.161	0.154	-0.134
Percentage of net irrigated area to total sown area	-0.149		-0.069	0.096	-0.066	-0.030	0.121	0.111	0.016	0.033
Per capita value of agricultural output	-0.039	-0.069		0.139	-0.028	0.145	0.045	0.089	0.046	-0.180
Monthly per capita consumption expenditure	-0.161	0.096	0.139		-0.039	0.119	0.120	0.084	0.067	0.045
Female Literacy rate	-0.040	-0.066	-0.028	-0.039		0.008	-0.082	-0.054	-0.052	0.151
women work force participation rate	0.002	-0.030	0.145	0.119	0.008		0.001	-0.016	0.147	-0.025
Percentage of households having access to safe drinking water	-0.087	0.121	0.045	0.120	-0.082	0.001		0.143	-0.070	-0.173
Percentage of villages having access to paved road	-0.161	0.111	0.089	0.084	-0.054	-0.016	0.143		-0.183	-0.113
Percentage of villages having access to primary health centre	0.154	0.016	0.046	0.067	-0.052	0.147	-0.070	-0.183		-0.007
Average Casual wage rate	-0.134	0.033	-0.180	0.045	0.151	-0.025	-0.173	-0.113	-0.007	

Extraction Method: Principal Component Analysis.

a) Residuals are computed between observed and reproduced correlations. There are 29 (64.0%) no redundant residuals with absolute values greater than 0.05.

b) Reproduced communalities

Table 12.13 (Output 5) has two parts of analysis: reproduced correlation and residuals. The reproduced correlation is the correlation among the extracted components. From this table, we intend to ensure that the correlation between the original variables and reproduced matrix should be as close as possible. The lower part of the table 'residual matrix' shows the difference between original variable and residual matrix. For a good PCA, it was expected that the difference between original variable and extracted matrix should be near to zero. Once the difference is near to zero, it can be said that the extracted components accounted a larger variation in the original correlation matrix. Here we can take an example from the table that the original correlation between irrigation and other than agricultural labour is 0.403, Whereas the extracted correlation between these two variable is 0.554 and the difference between the two correlation given in residual part is 0.403-0.554= -0.149.

Final PCA Index Value

The final index is calculated by the addition of multiplication of normalized value of the variable and the Eigen vector of that variable (first component). In our example the PCA index value for all the 11 selected district can be calculated by using two Table 12.5 and Table 12.12 given above. The final PCA index of first district (Malkangiri) is

PCA index Malkangiri District = a11x1+a12x2+……+a1nxn

= (0.202*1.237) + (-0.185*1.332) + (-0.403*1.899) + (0.*717*0.808) + (0.*837*0.561) + (-0.721*1.079) + (0.504*1.379) + (0.*733*0.614) + (0.*441*0.775) + (-.430*1.003)= 0.049

Likewise the PCA Index value of all other districts can be calculated as given in Table 12.14

Table 12.14: PCA Index Value

District	PCA Index
Malkangiri	0.049
Rayagada	1.273
Sundargarh	1.322
Nabarangapur	1.145
Kandhamal	1.641
Koraput	0.517
Mayurbhanj	1.333
Gajapati	1.757
Sambalpur	0.465
Kendujhar	0.290
Jharsuguda	1.551

The table 12.14 reveals that Gajapati is the most developed district followed by Kandhamal district. On the other hand, Malkangiri is most backward district.

12.6.2 Use of Output Indicators

After finding out the Index findings need to be validated. This can be done by comparing the result with some output indicators. Let us illustrate with example discussed in section 12.6. District wise final index as arrived out by PCA method is given in Table 12.14. This finding can be validated be taking an output indicator say infant mortality rate (IMR). We can run a correlation between PCA index and the output indicator 'Infant mortality rate'. This has been provided in Table 12.15 which is 0.893. The table shows that the most backward district has a higher degree of mortality. Hence our final index is validated. If suppose the correlation between these two variable is -0.120, then our index is not validated as the correlation between the two variable is very low.

Table 12.15: PCA Index and Comparing with Output Indicators

District	PCA Index	Infant Morality Rate*
Malkangiri	0.049	120
Kendujhar	0.290	100
Sambalpur	0.465	110
Koraput	0.517	92
Nabarangapur	1.145	89
Rayagada	1.273	100
Sundargarh	1.322	80
Mayurbhanj	1.333	75
Jharsuguda	1.551	65
Kandhamal	1.641	60
Gajapati	1.757	55

Note: Imaginary numbers

12.6.3 Use of Weight

Assigning weight i.e. all the indicators/dimensions are treated equally or differentially is an important issue in construction of composite index.. Sometimes different dimensions are given differential weight but in construction of overall index, an equal weight approach is followed. The human development index assigns differential weights to indicators/dimension but for overall index, an equal weight approach is followed. In construction of human development index, in 2001 by Planning Commission, the overall index was calculated by using equal weight whereas the different sub-indices like Composite indicator on educational attainment, Composite indicator on health attainment differential weights were given. In calculating educational attainment, two variables i.e. 'literacy rate for the age group 7 years and 'adjusted intensity of formal education' were used where the first and second variable were given 1/3 and 2/3 weight respectively. The other 'health attainment was calculated by taking two variables 'life expectancy at age one,'

and 'infant mortality rate'. The first variable has a weight of 2/3 whereas the second variable 1/3. But the overall weight is calculated by taking equal weight for both the variables. Take another example of our range equalization method where we have given equal weight to all variables. In some other cases, the individual variables of sub indices were given equal weights but the overall index was calculated by giving differential weights to sub-indicators.

Let us denote the weight in mathematical notation:

$$I=XW,$$

where 'X' is a matrix with m rows and n column, 'I' is identity matrix and 'W' is weight. Hence the final indices can be arrived at by taking the weighted component/variables. There are basically two ways in assigning weights. According to Munda and Nardo (2005), we can define weight by taking the importance of the particular variable or group of variables. In this process, the weights are arrived at by the past knowledge or observation of individual and the probable effects of that variable in the overall analysis. Let us illustrate with an example. Suppose we want to find out the sanitation and hygiene Index. For this we have selected some variables like 'proportion of people having access to toilet', 'proportion of people having drainage', 'proportion of people with wash hand', 'proportion of people safely disposed child exgratia' etc . Here either on the basis of literature or on the basis of our observation from some villages, we found that not having toilet is very important then wash hand before eating. In this case we can put higher weight for having toilet and less weight for washing hand.

In case of literature, weights are derived from on the basis of theoretical or statistical consideration.

12.6.4 Limitations of Principal Component Analysis

The major criticisms against the use of PCA is that the technique is arbitrary in constructing indices. The number of components and the number of variables used are not well defined. This method entirely depends on the first component. If the first component does not explain large variation, this method is not useful. Such possibility depends on nature of data and on relationship of variable.

Alternative methods to PCA that can reduce the dimensionality of the data include: correspondence analysis, multivariate regression or factor analysis. The details about these methods have been provided in Unit 13 and Unit 16 of Block 4.

12.7 MERITS AND LIMITATIONS OF COMPOSITE INDEX

One of the important merits of a composite index is that this can summarize the complex and multidimensional indicators into one indicator which helps the policy makers for implementation of a particular programme or policy decision. A composite index can easily be taken to interpret about the development or backwardness of a particular region or area or sector. The progress of a state in India say, e.g. Bihar can easily be accessed and we can compare the status of Bihar at two or more points of time. The composite index

can reduce the number of indicators without reducing the underlying information base. The composite index facilitates communication to general public and promote accountability. The composite index is not free from limitations. It has the risk of misleading policy message if it is poorly constructed or misinterpreted. The calculation may be misleading and faulty if sound statistical or conceptual principle is not applied. The assignment of weight many times creates a debate. One of the principles of assigning weight is on the basis of value judgment. This always being a source of criticism. Again allocation of upper and lower bound (mainly in range equalization method) is a point of criticism.

Check Your Progress 3

1) In the context of Composite Index, what are the uses of Principal Component Analysis (PCA).

……

……

……

……

……

2) What is explained by the term 'communality'?

……

……

……

……

……

3) How is Final Index Value calculated?

……

……

……

……

4) In construction of Human Development Index, how weights are assigned to the different indicators?

……

……

……

……

5) State the limitations of Composite Index.

……

……

……

……

……

12.8 LET US SUM UP

Composite index is an important analytical technique to analyse the developmental related issues like status of development, food security, human development etc. at the district/region/state level. It is an expression of single score made of different scores measuring the various dimensions and indicators of particular issue. This Unit describes the process to derive composite index by applying the various methods. These methods include simple ranking method, indices method, mean standarisation method, and range equalization method. The main advantage of the rank and indices method is that they are easy to understand and interpret. The RE, MS and PCA methods uses all the variables in reducing the dimensionality of the data. These methods are very much useful in comparing across districts, states, countries or areas such as rural and urban. Again these methods are useful to compare over time by constructing composite index at two points of time by taking same indicators/variables. The principal component analysis enables the students to reduce the indicators that are uncorrelated to explain the variation in the original data set. To run the PCA, SPSS and STATA softwares are preferred.

12.9 EXERCISES

1) Think and collect some gender related variables from the data sources like women literacy, women workforce participation rate in the age group 15-59, mean year schooling, mortality rate, immunization rate etc and find out the gender development index.

2) Explain the process of using Principal Component Analysis (PCA) to reduce the number of variables in a data set to smaller number of dimensions.

3) By using range equalization method find out the food security index from the report prepared by IHD-WFP given in website

   'http://122.180.7.122/displaymorePub.asp?itemid=84&subchkey=11&chname=Publications.'

12.10 SOME USEFUL BOOKS/REFERENCES

A. Saltelli, G. Munda and M. Nardo (2006), 'From Complexity to Multidimensionality: the Role of Composite Indicators for Advocacy of EU Reform', Tijdschrift voor Economie en Management Vol. LI, 3

Downloaded from http://www.econ.kuleuven.be/rebel//jaargangen/2001-2010/2006/TEM%202006-3/TEM_2006-3_03_Saltelli.pdf

Planning Commission (2001), 'National Human Development Report', Government of India downloaded from http://planningcommission.nic.in/reports/genrep/index.php?repts=nhdcont.htm

OECD (2008), 'Handbook on Constructing Composite Indicators Methodology and User Guide', Organization For Economic Co-operation and Development.

12.11 ANSWER OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 12.2
2) See Section 12.2
3) See Section 12.3
4) See Section 12.4

Check Your Progress 2

1) See Sub-section 12.4.1
2) See Sub-section 12.4.2
3) See Sub-section 12.5.4

Check Your Progress 3

1) See the heading 'conducting PCA and analyzing result' under Section 12.6
2) See Sub-section 12.6.1 (Under the head conducting the PCA and analyzing results)
3) See Sub-section 12.6.1 (Under the head final PCA index value)
4) See Sub-section 12.6.3
5) See Section 12.7

Indira Gandhi
National Open University
School of Social Sciences

MEC-109
Research Methods in Economics

Block

4

QUANTITATIVE METHODS-II

UNIT 13
Multivariate Analysis: Factor Analysis **5**

UNIT 14
Canonical Correlation Analysis **31**

UNIT 15
Cluster Analysis **43**

UNIT 16
Correspondence Analysis **59**

UNIT 17
Structural Equation Modeling **75**

Expert Committee

Prof. D.N. Reddy
Rtd. Professor of Economics
University of Hyderabad, Hyderabad

Prof. Harishankar Asthana
Professor of Psychology
Banaras Hindu University
Varanasi

Prof. Chandan Mukherjee
Professor and Dean
School of Development Studies
Ambedkar University, New Delhi

Prof. V.R. Panchmukhi
Rtd. Professor of Economics
Bombay University and Former
Chairman, ICSSR, New Delhi

Prof. Achal Kumar Gaur
Professor of Economics
Faculty of Social Sciences
Banaras Hindu University, Varanasi

Prof. P.K. Chaubey
Professor, Indian Institute of
Public Administration, New Delhi

Shri S.S. Suryanarayana
Former Joint Advisor
Niti Ayoug, New Delhi

Prof. Romar Korea
Professor of Economics
University of Mumbai, Mumbai

Dr. Manish Gupta
Sr. Economist
National Institute of Public Finance and Policy
New Delhi

Prof. Anjila Gupta
Professor of Economics
IGNOU, New Delhi

Prof. Narayan Prasad **(Convenor)**
Professor of Economics
IGNOU
New Delhi

Prof. K. Barik
Professor of Economics
IGNOU
New Delhi

Dr. B.S. Prakash
Associate Professor in Economics
IGNOU, New Delhi

Shri Saugato Sen
Associate Professor in Economics
IGNOU, New Delhi

Course Coordinator and Editor: Prof. Narayan Prasad

Block Preparation Team

Unit	Resource Person	IGNOU Faculty (Format and Language Editing)	
13	Dr. Darvinder Kumar (Sr. Scale) Assistant Professor in Statistics PGDAV College (University of Delhi), Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	**Block Editor**
14	Vivek N. Sharma Assistant Professor in Mathematics SGTB Khalsa College, (Univ. of Delhi), Delhi Ms. Neha Research Assist., IGNOU, New Delhi	Prof. Narayan Prasad IGNOU, New Delhi	Prof. G.K. Shukla Rtd. Prof. of Statistics IIT, Kanpur
15	Dr. Nausheen Nizami Assistant Professor in Economics Gargi College (University of Delhi), Delhi Ms. Neha Bailwal Research Assistant, IGNOU, New Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi	
16	Mr. Vivek N. Sharma Assistant Professor in Mathematics SGTB Khalsa College (Univ. of Delhi)	Shri B.S. Bagla PGDAV College (University of Delhi)	
17	Prof. Hari Shankar Asthana Professor of Psychology, BHU, Varanasi	Prof. Narayan Prasad Professor of Economics, IGNOU, New Delhi	

Print Production

Mr. Manjit Singh
Section Officer (Pub.), SOSS, IGNOU, New Delhi

Secretarial Assistance

Shri Vinay Kumar

December, 2015
© *Indira Gandhi National Open University, 2015*
ISBN-978-93-85911-28-6
All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.
Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.
Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.
Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi
Printed at :

BLOCK 4 QUANTITATIVE METHODS-II

Many a times a social scientist or a policy maker is required to identify the factors causing variation in any social or economic issue of policy importance like food security, child deprivation etc. in order to take corrective policy measures. To handle such issues, factor analysis technique, in case of interval scale or ratio scale data and correspondence analysis method in nominal or categorized variables are used. Another Multi-Variate statistical model namely canonical correlation analysis is used to examine the interrelationships among set of multiple dependent variables and multiple independent variables. Cluster analysis enables us to classify the units or observations into sub-groups or clusters based on characteristics or information contained in the variables. Structural Equation Modeling (SEM) attempts to examine the causal relations by including directly observed variables and latent variables. Keeping in view the widening role of interdisciplinary nature of economics, the mixed methods research is gaining popularity among researchers. All these five Multi-Variate statistical techniques are increasingly being used in Mixed Methods Research and hence have been covered in this block.

Unit 13 on '**Multivariate Analysis: Factor Analysis**' throws light on various concepts used in factor analysis, algorithm involved in factor analysis and the process to interpret the results of factor analysis.

Unit 14 on '**Canonical Correlation Analysis**' deals with the procedure to compute the canonical correlation, its application in social sciences, interpretation of its results and limitation.

Unit 15 entitled '**Cluster Analysis**' discusses the concept and purpose of cluster analysis, the various steps and algorithm in computation of its results and the different approaches followed in it.

Unit 16 on '**Correspondence Analysis**' covers the concept and special features of correspondence analysis, its computation algorithm, interpretation of its results and the process of application of this technique in social sciences.

Unit 17 entitled '**Structural Equation Modeling (SEM)**' throws light on the meaning and importance of SEM, its various models, steps involved in it, its advantages and disadvantages.

UNIT 13 MULTIVARIATE ANALYSIS: FACTOR ANALYSIS

Structure

13.0 Objectives

13.1 Introduction

13.2 Factor Analysis: Concept and Meaning

13.3 Historical Background of Factor Analysis

13.4 The Orthogonal Factor Model

- 13.4.1 Notations and Terminology
- 13.4.2 The Factor Model
- 13.4.3 Model Assumptions

13.5 Communalities

13.6 Methods of Estimation

- 13.6.1 Principal Component Method
- 13.6.2 Maximum Likelihood Method

13.7 Factor Rotation

13.8 Oblique Rotation

13.9 Factor Scores

13.10 Methods for Estimation of Factor Scores

- 13.10.1 Ordinary Least Square Method
- 13.10.2 Weighted Least Square Method
- 13.10.3 Regression Method

13.11 Let Us Sum Up

13.12 Key Words

13.13 Some Useful Books

13.14 Answers or Hints to Check Your Progress Exercises

13.15 Exercises

13.0 OBJECTIVES

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

- explain the terms used in factor analysis model like factor loadings, specific variances, communalities, factor rotation and oblique rotation;
- know the various uses of factor analysis model;
- elucidate how to apply principle component analysis and maximum likelihood methods for estimating the parameters of a factor model;
- discuss methods for estimating factor score; and
- learn how to interpret the results of factor analysis based on rotated factor loadings, rotated eigenvalues and scree test.

13.1 INTRODUCTION

We have seen in Unit 12 that how the principal component analysis (PCA) is a useful technique to reduce the indicators or data for construction of composite index and identify the variables for policy decisions. Factor analysis is a further extension of principal component analysis. Many a times, a researcher in social sciences is confronted with entangled human behaviour, unknown interdependencies and masses of qualitative and quantitative variables. In such situations, he may intend to uncover major social, institutional and international pattern. Factor analysis enables him to manage over a hundred variables, explore a content area, structure a domain, map unknown concepts, classify or reduce data, illuminate causal nexus, define relationships, test hypotheses, formulate theories and make inferences. In short, a social scientist can discern the regularity and order in a phenomena under investigation by applying factor analysis. Factor analysis takes thousands of measurements and qualitative observations and resolves them into distinct patterns of occurrence. Factor analysis as a constituent of multi-variate analysis involves a large number of issues ranging from conceptual, application, modeling, inference and interpretation, factor rotation etc. Keeping in view, the limited scope of this unit, we shall confine here to explain the concepts used in factor analysis, methods of estimation in factor model and the techniques of extraction such as principal component method and maximum likelihood method. Let us begin with explaining the concept of factor analysis.

13.2 FACTOR ANALYSIS: CONCEPT AND MEANING

The term factor analysis refers to anyone of a number of similar but distinct multi-variate statistical models that model observed variables as linear functions of a set of latent or hypothetical variables (also known as factors) not directly observed.

Factor analysis models are similar to regression models in the sense that both types of models possess dependent variables as linear functions of independent variables. However, they are distinct to each other in the sense that in factor analysis models, the independent variables are not observed independently of the observed dependent variables.

The factor variables contained in factor analysis models may be determinate or indeterminate. The determinate models encompass the various component analysis models such as Principal Components Analysis, weighted principal components and Gultman's image analysis. Indeterminate models are represented by the common factor model that seeks to account for co-variation between the observed variables and a set of common factor variables. Determinate factor analysis are more useful in data reduction role by finding a smaller number of variables that capture most of the information of variation and co-variation among the observed variables. In contrast, the common factors in the indeterminate models are indeterminate from the observed variables and are not linear combinations of them. In this unit, we shall confine our discussion to determinate factor analysis.

Factor analysis is based on a model in which the observed vector is partitioned into an unobserved systematic part and an unobserved error part. The

components of the error vector are considered as uncorrelated or independent, while the systematic part is taken as a linear combination of a relatively small number of unobserved factor variables. The analysis separates the effects of the factors, which are of the basic interest from the errors. From another point of view the analysis gives a description or explanation of the interdependence of a set of variables in terms of the factors without regard to the observed variability.

13.3 HISTORICAL BACKGROUND OF FACTOR ANALYSIS

The early development of factor analysis was due to Charles Spearman in the year 1904. He published a paper that is thought to be where factor analysis originated. Spearman was working with examination scores of school children and he was able to notice certain systematic effects in the matrix of correlation between the scores the children made on different tests. For example, in one case he obtained the following matrix of correlations for school children in a preparatory school for their scores on tests in Classics (C), French (F), English (E), Mathematics (M), Discrimination of pitch (D), and Music (Mu) as follows:

$$\begin{array}{c} \\ C \\ F \\ E \\ M \\ D \\ Mu \end{array} \begin{array}{c} \begin{array}{cccccc} C & F & E & M & D & Mu \end{array} \\ \begin{bmatrix} 1.00 & 0.83 & 0.78 & 0.70 & 0.66 & 0.63 \\ 0.83 & 1.00 & 0.67 & 0.67 & 0.65 & 0.57 \\ 0.78 & 0.67 & 1.00 & 0.64 & 0.54 & 0.51 \\ 0.70 & 0.67 & 0.64 & 1.00 & 0.45 & 0.51 \\ 0.66 & 0.65 & 0.54 & 0.45 & 1.00 & 0.40 \\ 0.63 & 0.57 & 0.51 & 0.51 & 0.40 & 1.00 \end{bmatrix} \end{array}$$

It can be noted out that this matrix has the interesting property that any two rows are almost proportional if the diagonals are ignored. Thus for rows C and E there are ratios:

$$\frac{0.83}{0.67} \simeq \frac{0.70}{0.64} \simeq \frac{0.66}{0.54} \simeq \frac{0.63}{0.51} \simeq 1.2$$

Spearman proposed the idea that the six test scores are all of the form

$$X_i = l_i F + \varepsilon_i$$

Where X_i is the i^{th} standardized score with a mean of zero and a standard deviation of one, l_i is a constant, F is a factor value which has mean zero and standard deviation one for individuals as a whole and ε_i is the part of X_i that is specific to the i^{th} test only. On the basis of his work Spearman formulated his two-factor theory of mental tests i.e each test result is made up of two parts, one that is common to all tests is general intelligence and another that is specific to the test. Later on this theory was modified to allow for each test result to consist of a part due to several common factors plus a part specific to test. This gives the general factor analysis model

$$X_i = l_{i1}F_1 + l_{i2}F_2 + \cdots + l_{im}F_m + \varepsilon_i$$

13.4 THE ORTHOGONAL FACTOR MODEL

In this section, first we will define some notations and terminology along with the underlying assumptions.

13.4.1 Notations and Terminology

Multivariate population and sample can be defined by mean vector μ and covariance matrix Σ. These are defined as follows. Suppose that there are p variables $X_1, X_2, \ldots\ldots X_p$ then the vector of sample mean is given by

$$\overline{X} = \begin{bmatrix} \overline{X}_1 \\ \overline{X}_2 \\ \vdots \\ \overline{X}_p \end{bmatrix}$$

This can be thought of as the centre of the sample. It is an unbiased estimate of the population vector of means

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

The matrix of the sample variance and covariances is

$$S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1m} \\ S_{21} & S_{22} & \cdots & S_{2m} \\ \vdots & \vdots & & \vdots \\ S_{p1} & S_{p2} & & S_{pm} \end{bmatrix}$$

where S is called the sample covariance matrix or sometimes sample dispersion matrix. It is an unbiased estimate of the population covariance matrix Σ.

Consider m unobserved common factors $F_1, F_2, \ldots\ldots F_m$. The i^{th} common factor is F_i. Generally, m is substantially less than p.

The common factors are also formed into a vector as

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_m \end{bmatrix}$$

13.4.2 The Factor Model

The factor model can be observed as a series of multiple regressions, predicting each of the observable variable X_i from the unobservable common factors F_i. In general the factor analysis model is given by

$$X_1 = \mu_1 + l_{11}F_1 + l_{12}F_2 + \cdots l_{1m}F_m + \varepsilon_1$$
$$X_2 = \mu_2 + l_{21}F_1 + l_{22}F_2 + \cdots l_{2m}F_m + \varepsilon_2$$
$$\vdots$$
$$X_p = \mu_p + l_{p1}F_1 + l_{p2}F_2 + \cdots l_{pm}F_m + \varepsilon_p$$

Here, the means $\mu_1, \mu_2, \ldots\ldots, \mu_p$ can be regarded as the intercept for the multiple regression models. The regression coefficients l_{ij} (slopes) for all the multiple regressions are called factor loadings. Here, l_{ij} is the loading of the i^{th} variable on the j^{th} factor. These will be shown in the matrix form as follows:

Matrix of the factor loading

$$\mathbf{L}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1m} \\ l_{21} & l_{22} & \cdots & l_{2m} \\ \vdots & \vdots & & \vdots \\ l_{p1} & l_{p2} & & l_{pm} \end{bmatrix}$$

and finally the errors ε_i are called the specific factors. Here, ε_i is the specific factor for variable i. The specific factors are also shown into a vector form as

Vector of specific factors

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix}$$

In general, the basic factor analysis model is like a regression analysis model. Each of our response variable X is expressed as a linear function of the unobserved common factors F_1, F_2,….F_m. Therefore, here one can say that we have m unobserved factors that control the variation among the data.

We will write the above factor model in matrix notation as

$$X = \mu + L\,F + \varepsilon \qquad (13.1)$$

In general we want m < p

13.4.3 Model Assumptions

Some assumptions are necessary to uniquely estimate the parameters. Otherwise, an infinite number of equally well fitted models with different values for the parameters may be obtained if these assumptions are not made. Therefore, the following assumptions are made about the unobserved random vectors F and ε and its certain covariance relationships:

i) The specific factors or random errors all have mean zero i.e.,

$$E(\varepsilon_i) = 0; \qquad i = 1, 2, \ldots\ldots, p$$

ii) The common factors also have mean zero i.e.,

$$E(F_j) = 0; \qquad j = 1, 2, \ldots\ldots, m$$

iii) The variance of specific factor i is Ψ_i, where Ψ_i is called the specific variance, and they are independent, i.e. $Cov\left(\varepsilon_i, \varepsilon_i'\right) = 0$. for $i \neq i'$.

$$\text{Cov}\,(\varepsilon_i) = E(\varepsilon_i, \varepsilon_i') = \Psi = \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix}; \quad i = 1, 2, \ldots\ldots, p$$

iv) The common factors F_j have variance one and are uncorrelated, i.e.

$$V(F_j) = \text{Cov}\,(F_j, F_j') = 0 \text{ for } j \neq j'.$$

Thus co-variance matrix of F is an identify matrix I, i.e. cov (F) = I

v) Since the common factors and specific factors are independent, then the specific factors are uncorrelated with the common factors i.e

$$\text{Cov}\left(\varepsilon_i, F_j\right) = E\left(\varepsilon_i, F_j'\right) = 0; \quad i = 1, 2, \ldots\ldots, p; \;\; j = 1, 2, \ldots\ldots, m$$

Now, the factor analysis model implies a covariance structure for the random variable X is given by equation

$$\begin{aligned}(X - \mu)\,(X - \mu) &= (L\,F + \varepsilon)\,(L\,F + \varepsilon) \qquad (13.1)\\ &= (L\,F + \varepsilon)\,((L\,F) + \varepsilon)\\ &= L\,F(L\,F) + \varepsilon(L\,F) + L\,F\,\varepsilon + \varepsilon\varepsilon\end{aligned}$$

so that

$$\begin{aligned}\Sigma = \text{Cov}\,(X) &= E(X - \mu)\,(X - \mu)\\ &= L\,E(F\,F)L + E(\varepsilon\,F)L + LE(F\,\varepsilon) + E(\varepsilon\varepsilon)\\ \therefore \quad \Sigma &= L\,L + \Psi \qquad (13.2)\end{aligned}$$

This is the matrix of factor loadings, times its transpose plus a diagonal matrix containing the specific variances. Hence a simplified model for the covariance matrix is obtained which is used for estimation.

Also, on multiplying F in equation (13.1) on both sides, we have

$$\begin{aligned}(X - \mu)F &= (L\,F + \varepsilon)F\\ &= L\,F\,F + \varepsilon\,F\end{aligned}$$

so, $\quad \text{Cov}\,(X, F) = E(X - \mu)F = LE(F\,F) + E(\varepsilon F) = L$

or $\quad \text{Cov}\left(X_i, F_j\right) = l_{ij}$

Under this model, the variance for the i^{th} observed variable is equal to the sum of the squared loadings for that variables and specific variance. The variance of variables X_i is:

$$\sigma_{ii} = \sum_{j=1}^{m} l_{ij}^2 + \psi_i \qquad (13.3)$$

This is a derivation which is based on the previous assumptions and the quantity $\sum_{j=1}^{m} l_{ij}^2$ is called the communality for variable i.

Important Remarks:

1) The model $X - \mu = L\,F + \varepsilon$ assumes that the data is a linear function of the common factors. However, since the common factors are not observable, we cannot check the linearity of the model.

2) The covariance matrix is a symmetric matrix, that is the variance between variables i and j is the same as the variance between j and i. For this model:

$$\Sigma = L\,L' + \Psi$$

The covariance matrix is going to have p(p+1)/2 unique elements of Σ which are approximated by mp factor loadings in the matrix L and the p specific variances ψ_i. When m= p, any covariance matrix Σ can be reproduced exactly as $L\,L'$, so Ψ can be the zero matrix. However, when m is smaller relative to p, that factor analysis is most useful. In this case,

the factor model provides a simple explanation of the covariation in X with fewer parameters than the p(p+1)/2 parameters in Σ.

For example, if X contains p=12 variables and the factor model with m= 2 is appropriate, then the p(p+1)/2= 78 elements of Σ are described in terms of the mp+ p= 36 parameters l_{ij} and ψ_i of the factor model.

Unfortunately, for the factor analyst, most covariance matrices cannot be factored as $L\,L' + \Psi$, where the number of factors m is much less than p. There are some problems, where it is not possible to get a unique consistent solution for the parameters l_{ij} and ψ_i from the variances and covariances of the observable variables.

3) If m > 1, there is always some inherent ambiguity associated with the factor model. To explain that, let T be any m × m orthogonal matrix. A matrix is orthogonal if its inverse is equal to the transpose of the original matrix.

$$TT' = T'T = I$$

Hence, we can write the factor model in matrix notation as:

$$X - \mu = L\,F + \varepsilon = L\,T\,T'F + \varepsilon = L^*\,F^* + \varepsilon$$

Note that this does not change the calculations since the identity matrix times any matrix just gives back the original matrix. This results in an alternative factor model, where the relationship between the new factor loadings and the original factor loadings is:

$$L^* = L\,T$$

and the relationship between the new common factors and the original common factor is:

$$F^* = T'F$$

This gives a model that fits equally well. Moreover, since there are infinite numbers of orthogonal matrices, then there are an infinite number of alternative models. This model, as it turns out, satisfied all the assumptions that we discussed earlier as:

$$E(F^*) = E(T'F) = T'E(F) = 0$$

$$Cov\,(F^*) = Cov\,(T'F) = T'Cov(F)T = T'I\,T = T'T = I$$

and

$$Cov\,(F^*, \varepsilon) = Cov\,(T'F, \varepsilon) = T'Cov(F, \varepsilon) = T'0 = 0$$

On the basis of observations on X, it is impossible to distinguish the loadings L from the loadings L^*. That is, the factors F and $F^* = T'F$ have the same statistical properties, and even though the loadings L^* are, in general different from the loadings L, they both have the same covariance matrix Σ. That is

$$\Sigma = L\,L' + \Psi = L\,T\,T'L' + \Psi = (L^*)\,(L^*)' + \Psi \qquad (13.4)$$

This ambiguity is going to be used to justify the factor rotation, since orthogonal matrices correspond to rotation of the coordinate system for X. We will use later to obtain more thrifty description of the data.

13.5 COMMUNALITIES

The communalities are computed by taking the sum of squares of the loadings for the i^{th} variable on the m common factors. This can be shown as below

Variance due to the specific factor is often called the uniqueness or specific variance. It is denoted by h_i^2, we have from equation (13.2)

$$\sigma_{ii} = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2 + \Psi_i$$

or $$\sigma_{ii} = h_i^2 + \Psi_i \; ; \qquad i = 1, 2, \ldots\ldots, p \qquad (13.5)$$

here, $$h_i^2 = l_{i1}^2 + l_{i2}^2 + \cdots + l_{im}^2$$

Now, to understand the computation of communalities in a simple way, let us have the table of three factor loadings. The following dataset involves the Places Rated Almanac (Boyer and Savageau) rates 329 communities according to nine criteria:

These data are only recorded for the first three factors only. It should also note that these factor loadings are the correlations between the factors and the variables. For example, the correlation between the Arts and the first factor is about 86%. Similarly the correlation between climate and that factor is only about 28%.

	Factor		
Variable	**1**	**2**	**3**
Climate	0.286	0.076	0.841
Housing	0.698	0.153	0.084
Health	0.744	-0.410	-0.020
Crime	0.471	0.522	0.135
Transportation	0.681	-0.156	-0.148
Education	0.498	-0.498	-0.253
Arts	0.861	-0.115	0.011
Recreation	0.642	0.322	0.044
Economics	0.298	0.595	-0.533

For example, to compute the communality for climate, the first variable which is given in the above table, we square the factor loadings for climate and then add the results:

$$h_1^2 = 0.286^2 + 0.076^2 + 0.841^2 = 0.795$$

In the similar way we can obtain the communalities for all the 9 variables and the total communality is obtained as 5.617. Therefore, all the communalities computed are placed into a table shown below:

Variable	Communality
Climate	0.795
Housing	0.518
Health	0.722
Crime	0.512
Transportation	0.510
Education	0.561
Arts	0.754
Recreation	0.517
Economics	0.728
Total	**5.617**

These values are as similar as multiple R^2 values for regression models predicting the variables of interest from the 3 factors. The communality for a given variable can be interpreted as the proportion of variation in that variable explained by the three factors. In other words, if we perform multiple regression of climate against the three common factors, we obtain an R^2= 0.795, indicating that about 79% of the variation in climate is explained by the factor model. The results suggest that the factor analysis does the best job of explaining variation in climate, the arts, economics, and health.

One assessment of how well this model is doing can be obtained from the communalities. What you want to see is values that are close to one. This would indicate that the model explains most of the variation for those variables. In this case, the model does better for some variables than it does for others. The model explains Climate the best, and is not bad for other variables such as Economics, Health and the Arts. However, for other variables such as Crime, Recreation, Transportation and Housing the model does not do a good job, explaining only about half of the variation. If you take all of the communality values and add them up you can get a total communality value:

$$\sum_{i=1}^{p} h_i^2 = \sum_{i=1}^{m} \lambda_i$$

Here, the total communality is 5.617. The proportion of the total variation explained by the three factors is

$$\frac{5.617}{9} = 0.624$$

This gives us the percentage of variation explained in our model. This might be looked at as an overall assessment of the performance of the model. However, this percentage is the same as the proportion of variation explained by the first three eigenvalues, obtained earlier. The individual communalities tell how well the model is working for the individual variables, and the total communality gives an overall assessment of performance. These are two different assessments that can be used.

Since the data are standardized in this case, the variance for standardized data is going to be equal to one. Then the specific variances can be computed by subtracting the communality from the variance as expressed below:

$$\Psi_i = 1 - h_i^2$$

Recall, that the data were standardized before analysis, so the variances of the standardized variables are all equal to one. For example, the specific variance for Climate is computed as follows:

$$\Psi_1 = 1 - 0.795 = 0.205$$

The specific variance for housing is Ψ_2 =0.482 and the other values of specific variances can be calculated accordingly.

Source: http://onlinecourses.science.psu.edu

EXAMPLE 1 Calculating Correlations from Factors

Factor analysis aims to explain the correlations among the observed variables in terms of small number of factors. To estimate the success of factor solution, we reproduce the original correlation matrix by using the loadings on the common factors and see how large a discrepancy is there between the original matrix and reproduced correlation matrix. The greater the discrepancy, the less successful the factor solution in preserving the information in the original correlation matrix. And, lesser the discrepancy, more successful the factor solution in preserving the information in original correlation matrix.

In order to calculate the correlations from the factor solution the process is quite simple when the factors are uncorrelated. The correlation between any two variables X_1 and X_2 is obtained by adding products of coefficients for these variables (X_1 and X_2) across all the common factors. Let us say for the three factor solution F_1, F_2 and F_3 the quantity would be $(l_{11} \times l_{21}) + (l_{12} \times l_{22}) + (l_{13} \times l_{23})$. This process will become clearer in the description of the hypothetical two-factor solution based on five observed variables as follows:

A Hypothetical Solution

Variables	**Loadings/ Correlations**		**Communality**	**Reproduced Correlations**					
	Factor 1	**Factor 2**			**X_1**	**X_2**	**X_3**	**X_4**	**X_5**
X_1	.5	.2	.29	X_1					
X_2	.6	.3	.45	X_2	.36				
X_3	.7	.4	.65	X_3	.43	.54			
X_4	.3	.6	.45	X_4	.27	.36	.45		
X_5	.4	.7	.65	X_5	.34	.45	.56	.54	
	$\sum x^2 = 1.35$	$\sum x^2 = 1.14$							

The Coefficients: According to the above solution,

$X_1 = 0.5F_1 + 0.2F_2$

$X_2 = 0.6F_1 + 0.3F_2$

⋮

$X_5 = 0.4F_1 + 0.7F_2$

As well as being weights, the coefficients above show that the correlation between X_1 and F_1 is .50, and that of between X_1 and F_2 is .20, and so on.

Variance Accounted For: The quantities at the bottom of each factor column are the sums of the squared loadings for that factor, and show how much of the total variance of the observed variables is accounted for by that factor. For Factor 1, the quantity is $.5^2 + .6^2 + .7^2 + .3^2 + .4^2 = 1.35$. Because in the factor analyses discussed here, the total amount of variance is equal to the number of observed variables (the variables are standardized, so each has a variance of one), the total variation here is five, so that Factor 1 accounts for (1.35/5)×100 = 27% of the variance.

The quantities in the communality column show the proportion of the variance of each variable accounted for by the common factors. For X_1 this quantity is $.5^2 + .2^2 = 0.29$, for X_2 it is $.6^2 + .3^2 = 0.45$, and so on.

Reproducing the Correlations: The correlation between variables X_1 and X_2 as derived from the factor solution is equal to (.7 X .8) + (.2 X .3) = 0.36, while the correlation between variables X_3 and X_5 is equal to (.7 X .4) + (.4 X .7) = .56. These values are shown in right-hand side of the above table.

EXAMPLE 2 To verify the relation $\Sigma = L\,L' + \Psi$ for two factors:

Consider the following covariance matrix

$$\Sigma = \begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix}$$

The equality equation of the model is

$$\begin{bmatrix} 19 & 30 & 2 & 12 \\ 30 & 57 & 5 & 23 \\ 2 & 5 & 38 & 47 \\ 12 & 23 & 47 & 68 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix} \begin{bmatrix} 4 & 7 & -1 & 1 \\ 1 & 2 & 6 & 8 \end{bmatrix} + \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

or

$$\Sigma = L\,L' + \Psi$$

may be verified by the matrix algebra. Therefore, Σ has the structure produced by an m= 2 orthogonal factor model. Since

$$L = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \\ l_{31} & l_{32} \\ l_{41} & l_{42} \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 7 & 2 \\ -1 & 6 \\ 1 & 8 \end{bmatrix}$$

$$\Psi = \begin{bmatrix} \psi_1 & 0 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 \\ 0 & 0 & \psi_3 & 0 \\ 0 & 0 & 0 & \psi_4 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

The communality of X_i can be obtained by using equation (13.5), we have

$$h_i^2 = l_{11}^2 + l_{12}^2 = 4^2 + 1^2 = 17$$

and the variance of X_i can be decomposed as

$$\sigma_{11} = (l_{11}^2 + l_{12}^2) + \psi_1 = h_1^2 + \psi_1$$

or

$$19 = 4^2 + 1^2 + 2 = 17 + 2$$

i.e. variance = communality+ specific variance

Similarly, we can find out the results for the other variables also.

Check Your Progress 1

1) What do you mean by a factor in the context of factor analysis?

……

……

……

……

……

2) Show that the covariance matrix

$$\Sigma = \begin{bmatrix} 1.0 & .63 & .45 \\ .63 & 1.0 & .35 \\ .45 & .35 & 1.0 \end{bmatrix}$$

For the p=3 standardized random variables Z_1, Z_2 and Z_3 can be generated by the m=1 factor model:

$$Z_1 = .9F_1 + \varepsilon_1$$

$$Z_2 = .7F_1 + \varepsilon_2$$

$$Z_3 = .5F_1 + \varepsilon_3$$

where, $\text{Var}(F_1) = 1, \quad \text{Cov}(\varepsilon, F_1) = 0$

and $$\Psi = \text{Cov}(\varepsilon) = \begin{bmatrix} .19 & 0 & 0 \\ 0 & .51 & 0 \\ 0 & 0 & .75 \end{bmatrix}$$

Formulate the factor model for the given problem and also calculate

(a) Communalities h_i^2; i= 1, 2, 3 and interpret the result.

(b) Cor (Z_i, F_i) for i= 1, 2, 3. Which variable carry the greatest weight in naming the common factor? And why?

……

……

……

……

3) Let the factor model with p=2 and m=1. Show that

$$\sigma_{11} = l_{11}^2 + \psi_1$$

$$\sigma_{12} = \sigma_{21} = l_{11}l_{21}$$

$$\sigma_{22} = l_{21}^2 + \psi_2$$

for given σ_{11}, σ_{22} and σ_{12}, there is an infinite choices for L and Ψ i.e., the factor model need not be unique.

…….....

…….....

…….....

…….....

…….....

4) Consider an m=1 factor model for the population with covariance matrix

$$\Sigma = \begin{bmatrix} 1 & .4 & .9 \\ .4 & 1 & .7 \\ .9 & .7 & 1 \end{bmatrix}$$

Show that there is a unique choice of L and Ψ with $\Sigma = L L' + \Psi$, but that $\psi_3 < 0$, so the choice is not admissible i.e., the solution is unique but improper.

…….

…….

…….

…….

…….

13.6 METHODS OF ESTIMATION

For the observations $x_1, x_2, \ldots\ldots, x_n$ on p generally correlated variable, does the factor model of equation (13.1) with a small number of factors, effectively represent the data? The solution of such kind of statistical problems may be tackled by using the covariance relationships. The sample covariance matrix S, is an unbiased estimate of the unknown population covariance matrix Σ. If Σ appears to deviate significantly from a diagonal matrix, then a factor model can be entertained, and the initial problem is one of estimating the factor loadings l_{ij} and specific variances ψ_i.

In this section, we shall consider two of the most popular methods of parameter estimation, the principal component (and the related principal factor) method and the maximum likelihood method. The solution from either method can be rotated in order to simplify the interpretation of factors.

13.6.1 Principal Component Method

A vector of observations for the X_i variable may be defined as

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

The sample covariance matrix is denoted by S and is expressed as

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})\,(X_i - \bar{X})'$$

We have p eigenvalues for the covariance matrix S as well as corresponding eigenvectors for this matrix. Let the eigenvalues of the covariance matrix S are $\hat{\lambda}_1, \hat{\lambda}_2, \ldots\ldots\hat{\lambda}_p$ and eigenvectors of the covariance matrix S are $\hat{e}_1, \hat{e}_2, \ldots\ldots\hat{e}_p$. The covariance matrix can be re-expressed in the following form as a function of the eigenvalues and the eigenvectors:

Spectral Decomposition of Σ: The spectral decomposition allows us to express the inverse of a square matrix in terms of its eigenvalues and eigenvectors, and this leads to a useful square-root matrix. Let the population covariance matrix Σ have eigenvalues and eigenvectors pairs (λ_i, e_i) with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, then

$$\Sigma = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \cdots + \lambda_p e_p e_p' = \sum_{i=1}^{p} \lambda_i e_i e_i'$$

$$= \left[\sqrt{\lambda_1}e_1 \;\vdots\; \sqrt{\lambda_2}e_2 \;\vdots\; \cdots\cdots \;\vdots\; \sqrt{\lambda_p}e_p\right] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \cdots\cdots \\ \sqrt{\lambda_2}e_2' \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ \sqrt{\lambda_p}e_p' \end{bmatrix} \qquad (13.6)$$

This fits the prescribed covariance structure for the factor analysis model having as many factors as variables i.e., m= p and specific variances $\Psi_i = 0$ for all i. The loading matrix has j^{th} column given by $\sqrt{\lambda_j}e_j$. Then we can write

$$\Sigma = L\,L + 0 = L\,L \qquad (13.7)$$

Apart from the scale factor $\sqrt{\lambda_j}$, the factor loadings on the j^{th} factor are the coefficients for the j^{th} principal component of the population. The factor analysis representation of Σ in equation (13.7) is exact, it is not particularly useful. It employs as many common factors as there are variables and does not allow any variation in the specific factors ε in the main model of equation (13.1). But we prefer models that explain the covariance structure in terms of just a few common factors. One approach when the last $p - m$ eigenvalues are small is to neglect the contribution of $\lambda_{m+1}e_{m+1}e_{m+1}' + \cdots + \lambda_p e_p e_p'$ to Σ in eq (13.6), we obtain the approximation

$$\Sigma = \left[\sqrt{\lambda_1}e_1 \;\vdots\; \sqrt{\lambda_2}e_2 \;\vdots\; \cdots\cdots \;\vdots\; \sqrt{\lambda_m}e_m\right] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \cdots\cdots \\ \sqrt{\lambda_2}e_2' \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ \sqrt{\lambda_m}e_m' \end{bmatrix} = L\,L \qquad (13.8)$$

The approximation shown in eq (13.8) assumes that the specific factors ε in eq (13.1) are of minor importance and can also be ignored in the factoring of Σ.

Allowing for specific factors, we find that the approximation becomes

$$\Sigma = L\,L' + \Psi$$

$$= \left[\sqrt{\lambda_1}e_1 \;\vdots\; \sqrt{\lambda_2}e_2 \;\vdots\; \cdots\cdots \;\vdots\; \sqrt{\lambda_m}e_m\right] \begin{bmatrix} \sqrt{\lambda_1}e_1' \\ \cdots\cdots \\ \sqrt{\lambda_2}e_2' \\ \cdots\cdots \\ \vdots \\ \cdots\cdots \\ \sqrt{\lambda_m}e_m' \end{bmatrix} + \begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \psi_p \end{bmatrix} \quad (13.9)$$

where

$$\psi_i = \sigma_{ii} - \sum_{j=1}^{m} l_{ij}^2\,; \qquad i = 1, 2, \ldots\ldots, p$$

This yields the following estimator for the factor loadings

$$\hat{l}_{ij} = \hat{e}_{ij}\sqrt{\hat{\lambda}_i}$$

When the units of the variables are not appropriate, it is better to use the standardized variables. If the standardized measurements are used, we replace sample correlation matrix S by R. This in turn suggests that the specific variances, which are the diagonal elements of the matrix Ψ, can be estimated by using the expression:

$$\hat{\psi}_i = S_{ii} - \sum_{j=1}^{m} \hat{l}_{ij}^2\,; \qquad i = 1, 2, \ldots\ldots, p \qquad (13.10)$$

Here, the estimated specific variances are equal to the sample variance for the i^{th} variable minus the sum of the squares of factor loadings (i.e the communality).

The equation (13.10), when applied to the sample covariance matrix S or the sample correlation matrix R, is known as the principal component solution.

13.6.2 Maximum Likelihood Method

Maximum likelihood estimation requires that the data are sampled from a multivariate normal distribution. But if the data have been collected on a likert scale, which is most often the case in the social sciences, these kinds of data cannot really be normally distributed. This is the main drawback of this method. Therefore, using the maximum likelihood estimation method we must assume that the data are independently sampled from a multivariate normal distribution with mean vector μ and the covariance matrix Σ that takes the particular form

$$\Sigma = L\,L' + \Psi$$

where L is the matrix of factor loadings and Ψ is the diagonal matrix of specific variances.

Again, let us assume that the p×1 vectors $X_1, X_2, \ldots\ldots X_n$ represents a random sample which is drawn from a multivariate normal population $N_p(\mu, \Sigma)$, then the joint probability function or the likelihood function is given by

$$l(\mu,\Sigma)=\prod_{j=1}^{n}\left\{\frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\,e^{-(x_j-\mu)'\Sigma^{-1}(x_j-\mu)/2}\right\}$$

$$=\frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}}\,e^{-\sum_{j=1}^{n}(x_j-\mu)'\Sigma^{-1}(x_j-\mu)/2} \qquad (13.11)$$

Maximum likelihood estimation involves estimating the mean, the matrix of factor loadings and the specific variances.

The maximum likelihood estimator for the mean vector μ , the factor loadings L and the specific variances Ψ are obtained by finding the values of $\hat{\mu}$, $\hat{L}$ and $\hat{\Psi}$ that maximizes by taking logarithm of the likelihood function in equation (13.11) which is given by the following expression:

$$\text{Log } l(\mu,L,\Psi)=-\frac{np}{2}\log 2\pi$$

$$-\frac{n}{2}\log|LL'+\Psi|-\frac{1}{2}\sum_{j=1}^{n}(x_j-\mu)'(LL'+\Psi-1)\,(x_j-\mu) \qquad (13.12)$$

The log of the likelihood function of the data is to be maximized. We want to find the values of the population parameters μ, L and Ψ. As mentioned earlier, the solution for these factor models are not unique. Equivalent models can be obtained by rotation. To obtain a unique solution an additional constraint to be imposed is that $L'\Psi^{-1}L$ is a diagonal matrix.

Computationally this process is complex. In general, there is no closed- form solution to this maximization problem, so iterative methods must be applied.

Source: Adapted from Applied Multivariate Statistical analysis by the Johnson R.A. and Wichern D.W. (2002).

13.7 FACTOR ROTATION

Broadly speaking, there are two kinds of rotations i.e. orthogonal rotation and oblique rotation. Orthogonal rotation assumes that the factors are uncorrelated. This is less realistic since factors generally are correlated with each other to some extent or the other. Two commonly used orthogonal rotation techniques are Quartimax and Varimax rotation. Quartimax involves the minimization of the number of factors needed to explain each variable whereas Varimax minimizes the number of variables that have high loadings on each factor and works to make small loadings even smaller.

All factor loadings obtained from the initial loadings by an orthogonal transformation have the same ability to reproduce the covariance matrix. From matrix algebra, we know that an orthogonal transformation corresponds to a rigid rotation of the coordinate axis. For this reason, an orthogonal transformation of the factors is called factor rotation.

If $\hat{L}$ is the p×m matrix of estimated factor loadings obtained by any one of the method i.e., principal component or maximum likelihood, then

$$\hat{L}^{*}=\hat{L}\,T \qquad \text{where} \quad TT'=T'T=I \qquad (13.13)$$

is a p×m matrix of rotated loadings. Moreover, the estimated covariance matrix remains unchanged, since

$$\hat{L}\hat{L}'+\hat{\Psi}=\hat{L}TT'\hat{L}+\hat{\Psi}=\hat{L}^{*}\hat{L}^{*'}+\hat{\Psi} \qquad (13.14)$$

Equation (13.14) shows that the residual matrix, $S_n - \hat{L}\hat{L}' - \hat{\Psi} = S_n - \hat{L}^*\hat{L}^{*'} - \hat{\Psi}$, remains unchanged. Moreover, the specific variances $\hat{\Psi}_i$ and hence the communalities $\hat{h}_i^2$ are unaltered. Thus from a mathematical viewpoint, it is immaterial whether $\hat{L}$ or $\hat{L}^*$ is obtained.

The various factor rotation methods have, as a guiding principle, the simple structure concepts. That is to say, the results after rotation should become simple in their appearance. To put it another way, these simple structure concepts should be considered when trying to determine whether or not a given factor rotation has clarified the underlying structure of the data.

Factors are rotated for better interpretation since un-rotated factors are ambiguous. The goal of rotation is to attain an optimal simple structure which attempts to have each variable load on as few factors as possible, but maximizes the number of high loadings on each variable. Ultimately, the simple structure attempts to have each factor define a distinct cluster of interrelated variables so that interpretation becomes easier. For example, variables that relate to language should load highly on language ability factors but should have close to zero loadings on mathematical ability.

Varimax Rotation: Perhaps the most widely used is the varimax criterion. It seeks the rotated loadings that maximize the variance of the squared loadings for each factor; the goal is to make some of these loadings as large as possible, and the rest as small as possible in absolute value. The varimax method encourages the detection of factors each of which is related to few variables. It discourages the detection of factors influencing all variables. The quartimax criterion, on the other hand, seeks to maximize the variance of the squared loadings for each variable, and tends to produce factors with high loadings for all variables.

Kaiser [5] has suggested an analytical measure of simple structure known as the varimax (normal varimax) criteria. Let us define $\tilde{l}_{ij}^{*4} = \hat{l}_{ij}^*/\hat{h}_i$ to be the rotated coefficients scaled by the square root of the communalities. Then the (normal) varimax procedure selects the orthogonal transformation T that makes

$$V = \frac{1}{p}\sum_{j=1}^{m}\left[\sum_{i=1}^{p}\tilde{l}_{ij}^{*4} - \left(\sum_{i=1}^{p}\tilde{l}_{ij}^{*2}\right)^2/p\right] \qquad (13.15)$$

as large as possible.

Scaling the rotated coefficients $\hat{l}_{ij}^*$ has the effect of giving variables with small communalities relatively more weight in the determination of simple structure. After the transformation T is determined, the loadings $\tilde{l}_{ij}^*$ are multiplied by $\hat{h}_i$ so that the original communalities are preserved.

In the above equation, it has a simple interpretation. In words

$$V \propto \sum_{j=1}^{m}\left(\text{variance of the squares of (scaled)loadings for } j^{th} \text{ factor}\right)$$

One can maximize V that corresponds to the squares of the loadings on each factor as much as possible. Therefore, we hope to find groups of large and negligible coefficients in any column of the rotated loadings matrix $\hat{L}^*$.

13.8 OBLIQUE ROTATION

Orthogonal rotations are appropriate for a factor model in which the common factors are assumed to be independent. Many investigators in social sciences consider oblique (non-orthogonal) rotations, as well as orthogonal rotations. The former rotations are often suggested after one view of the estimated factor loadings that do not follow the assumption made in the factor model. Nevertheless, an oblique rotation is frequently a useful aid in factor analysis.

In other words, oblique rotation is suitable when the factors are considered to be correlated. Oblique rotation is more complex than orthogonal rotation, since it can involve one of two coordinate systems, a system of primary axes or a system of reference axes. Additionally, oblique rotation produces a pattern matrix that contains the factor or item loadings and factor correlation matrix that includes the correlations between the factors. The commonly used oblique rotation techniques are Direct Oblimin and Promax. Direct Oblimin attempts to simplify the structure and the mathematics of the output of the problem under consideration, while Promax is used because of its speed in larger datasets. Promax involves raising the loadings to a power of four which ultimately results in greater correlations among the factors and achieves a simple structure.

13.9 FACTOR SCORES

A factor score can be considered to be a variable describing how much an individual would score on a factor. One of the methods to produce factor score is called Bartlett method (or regression approach) which produces unbiased scores that are correlated only with their own factor. Another method is called the Anderson-Rubin method which produces scores that are uncorrelated and standardized. In factor analysis, interest is usually centered on the parameters in the factor model. However the estimated value of the common factors, called factor scores, may also be required. These quantities are often used for diagnostic purposes, as well as inputs to a subsequent analysis.

Factor scores are not estimates of unknown parameters in the usual sense. Rather, they are estimates of values for the unobserved random factors F_j, j = 1, 2,……,m. We try to find out the vectors of m unobserved common factors that underlie our model to estimate those factors. Therefore, for the factor model

$$X - \mu = L\,F + \varepsilon$$

We may wish to estimate the vectors of the factor scores F_1, F_2,……..,F_m.

13.10 METHODS FOR ESTIMATING FACTOR SCORE

A number of different methods are used for estimating factor scores from the data. These include

1) Ordinary Least Square Method
2) Weighted Least Square Method
3) Regression Method

Now, the above methods are discussed below in brief.

13.10.1 Ordinary Least Squares Method

If the factor loadings are estimated by the principal component method, it is customary to generate factor scores using an unweighted (ordinary) least squares procedure. The L's are factor loading and the f are unobserved common factors. The vector of common factors for subject i, is found by minimizing the sum of the squared residuals:

$$\sum_{j=1}^{p} \varepsilon_{ij}^2 = \sum_{j=1}^{p} \left(X_{ij} - \mu_j - l_{j1}F_1 - l_{j2}F_2 - \cdots - l_{jm}F_m\right)^2 = (X - \mu - L\,F)'(X - \mu - L\,F)$$

This is like a least squares regression, except in this case we already have estimates of the parameters (the factor loadings), but we wish to estimate the explanatory common factors. In matrix notation the solution is expressed as:

$$F_j = (L'L)^{-1}\,L'(X_j - \mu) \qquad (13.16)$$

In practice, the parameters are unknown in nature so we take the estimates of the factor loadings as follows:

$$\hat{F}_j = \left(\hat{L}'\,\hat{L}\right)^{-1}\hat{L}'(X_j - \bar{X})$$

or

$$\hat{F}_j = \left(\hat{L}_z'\,\hat{L}_z\right)^{-1}\hat{L}_z'\,z_j \qquad (13.17)$$

For the standardized data, since $L = \left[\sqrt{\hat{\lambda}_1}e_1 \vdots \sqrt{\hat{\lambda}_2}e_2 \vdots \cdots \vdots \sqrt{\hat{\lambda}_m}e_m\right]$, we have

$$\hat{F}_j = \begin{bmatrix} \frac{1}{\sqrt{\hat{\lambda}_1}}\,\hat{e}_1'(X_j - \bar{X}) \\ \frac{1}{\sqrt{\hat{\lambda}_2}}\,\hat{e}_2'(X_j - \bar{X}) \\ \vdots \\ \frac{1}{\sqrt{\hat{\lambda}_m}}\,\hat{e}_m'(X_j - \bar{X}) \end{bmatrix}$$

We see that the $\hat{F}_j$ are approximately same as that of obtained in the principal component method with the unrotated factor loadings. Here, we have for the factor scores

$$\frac{1}{n}\sum_{j=1}^{n}\hat{F}_j = 0 \qquad \text{(sample mean)}$$

$$\frac{1}{n-1}\sum_{j=1}^{n}\hat{F}_j\hat{F}_j' = I \qquad \text{(sample variance)}$$

13.10.2 Weighted Least Squares Method

The difference between weighted least square and the ordinary least square is that the squared residuals are being divided by the specific variances as shown below. Let us suppose that the mean vector μ, the factor loading L, and the specific variance Ψ are known for the factor model

$$X - \mu = L\,F + \varepsilon$$

Further, regard the specific factors $\varepsilon' = [\varepsilon_1, \varepsilon_2, \ldots\ldots, \varepsilon_p]$ as errors. Since, $Var(\varepsilon_i) = \Psi_i$; i= 1, 2,……,p, need not be equal. Bartlett [2] has suggested that weighted least squares be used to estimate the common factor values.

The sum of squares of the errors, weighted by the reciprocal of their variance is

$$\sum_{i=1}^{p} \frac{\varepsilon_i^2}{\Psi_i} = \sum_{j=1}^{p} \varepsilon' \Psi^{-1} \varepsilon = (X - \mu - LF)' \Psi^{-1} (X - \mu - LF) \qquad (13.18)$$

The solution is given by the expression where Ψ is the diagonal matrix whose diagonal elements are equal to the specific variances.

$$\hat{F} = \left(L' \Psi^{-1} L\right)^{-1} L' \Psi^{-1} (X - \mu)$$

We take the estimates $\hat{L}, \hat{\Psi}$ and $\hat{\mu} = \overline{X}$ as the true values and obtain the factor scores for the j^{th} case as

$$\hat{F}_j = \left(\hat{L}' \hat{\Psi}^{-1} \hat{L}\right)^{-1} \hat{L}' \hat{\Psi}^{-1} \left(X_j - \overline{X}\right) \qquad (13.19)$$

When $\hat{L}$ and $\hat{\Psi}$ are determined by the maximum likelihood method, these estimates must satisfy the uniqueness condition, $\hat{L}' \hat{\Psi}^{-1} \hat{L} = \hat{\Delta}$, a diagonal matrix.

13.10.3 Regression Method

This method is used when you are calculating maximum likelihood estimates of factor loadings. A vector of the observed data, supplemented by the vector of factor loadings for the ith subject, is considered.

In factor model $X - \mu = L\,F + \varepsilon$, we initially treat the loadings matrix L and specific variance matrix Ψ is known. When the common factors F and specific factors ε are jointly distributed with means and covariances, the linear combination $X - \mu = L\,F + \varepsilon$ has an $N_p(0, LL' + \Psi)$ distribution. Moreover, the joint distribution of $(X - \mu)$ and F is $N_{m+p}(0, \Sigma^*)$, where

$$\Sigma^* = \begin{bmatrix} \Sigma = LL' + \Psi & \vdots & L \\ \ldots\ldots\ldots\ldots\ldots\ldots & \vdots & \ldots\ldots \\ L' & \vdots & I \end{bmatrix} \qquad (13.20)$$

and 0 is an (m×p)×1 vector of zeros. We find that the conditional distribution of $(F|x)$ is multivariate normal with

$$\text{mean} = E(F|x) = L'\Sigma^{-1}(x - \mu) = L'(LL' + \Psi)^{-1}(x - \mu) \qquad (13.21)$$

and

$$\text{Covariance} = E(F|x) = I - L^{\Sigma^{-1}} L = I - L'(LL' + \Psi)^{-1} L \qquad (13.22)$$

The quantities $L'(LL' + \Psi)^{-1}$ are the coefficients in a multivariate regression of the factors on the variables. Estimates of these coefficients produce factor scores that are analogous to the estimates of the conditional mean values in multivariate regression analysis. Consequently, given any vector of observations x_j, and taking the maximum likelihood estimates $\hat{L}$ and $\hat{\Psi}$ as the true values, we see that the j^{th} factor score vector is given by

$$\hat{f}_j = \hat{L}'\hat{\Sigma}^{-1}(x_j - \bar{x}) = \hat{L}'(\hat{L}\hat{L}' + \hat{\Psi})^{-1}(x_j - \bar{x}); \qquad j = 1,2, \ldots\ldots n \qquad (13.23)$$

The calculation of $\hat{f}_j$ in equation (13.23) can be simplified by using the matrix identity

$$\hat{L}'(\hat{L}\hat{L}' + \hat{\Psi})^{-1} = \left(I + \hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1}\hat{L}'\hat{\Psi}^{-1} \qquad (13.24)$$

This identity allows us to compare the factor scores in equation (13.23), generated by the regression argument, with those generated by the weighted least squares method in equation (13.19). Temporarily, If we denote the former by $\hat{F}_j^R$ and the latter by $\hat{F}_j^{LS}$. Then by using equation ( ), we have

$$\hat{F}_j^{LS} = \left(\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1}\left(I + \hat{L}'\hat{\Psi}^{-1}\hat{L}\right)\hat{F}_j^R = \left(I + \left(\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1}\right)\hat{F}_j^R \qquad (13.25)$$

For maximum likelihood estimates $\left(\hat{L}'\hat{\Psi}^{-1}\hat{L}\right)^{-1} = \hat{\Delta}^{-1}$ and if the elements of this diagonal matrix are close to zero, the regression and generalized least squares methods will give nearly the same factor scores.

There may be made an attempt to reduce the effect of a incorrect determination of the number of factors, to calculate the factor scores in equation (13.23) by using S (the sample covariance matrix) instead of $\hat{\Sigma} = \hat{L}\hat{L}' + \hat{\Psi}$. Then we have the following results:

$$\hat{f}_j = \hat{L}'S^{-1}(x_j - \bar{x}); \qquad j = 1,2, \ldots\ldots n \qquad (13.26)$$

or, if a correlation matrix is factored

$$\hat{f}_j = \hat{L}_z' R^{-1} z_j; \qquad j = 1,2, \ldots\ldots n \qquad (13.27)$$

where

$$z_j = D^{-1/2}(x_j - \bar{x}) \quad \text{and} \quad \hat{\rho} = \hat{L}_z\hat{L}_z' + \hat{\Psi}_z$$

Again, if rotated loadings $\hat{L}^* = \hat{L}T$ are used in place of the original loadings in eq (13.26), the subsequent factor scores $\hat{f}_j^*$ are related to $\hat{f}_j$ by

$$\hat{f}_j^* = T'\hat{f}_j; \qquad j = 1,2, \ldots\ldots, n$$

A numerical measure between the factor scores generated from two different calculation methods is provided by the sample correlation coefficient between scores on the same factor.

Source: Adapted from Applied multivariate statistical analysis by Johnson R.A. and Wichern D.W. (2002)

Check Your Progress 2

1) The correlation matrix for chicken- bone measurements is

$$\begin{bmatrix} 1.000 & & & & & \\ .505 & 1.000 & & & & \\ .569 & .422 & 1.000 & & & \\ .602 & .467 & .926 & 1.000 & & \\ .621 & .482 & .877 & .874 & 1.000 & \\ .603 & .450 & .878 & .894 & .937 & 1.000 \end{bmatrix}$$

The following estimated factor loadings were extracted by the maximum likelihood procedure:

Variable	Estimated factor loadings		Varimax rotated estimated factor loadings	
	F_1	F_2	F_1^*	F_1^*
Skull length	.602	.200	.484	.411
Skull breadth	.467	.154	.375	.319
Femur length	.926	.143	.603	.717
Tibia length	1.000	.000	.519	.855
Humerus length	.874	.476	.861	.499
Ulna length	.894	.327	.744	.594

Using the unrotated estimated factor loadings, obtain the maximum likelihood estimates of the following:

a) The specific variances

b) The communalities

c) The proportion of variance explained by each factor

d) The residual matrix $R - \hat{L}_z \hat{L}_z' - \hat{\Psi}_z$

……

……

……

……

……

……

1) From the above exercise 1, Compute the value of the varimax criterion using both unrotated and rotated estimated factor loadings Also interpret the result.

……

……

……

……

……

……

……

……

2) The following table provides data of the Students entering in a certain MBA program must take three required courses in Finance, Marketing and Business policy. Let X_1, X_2, and X_3, respectively, represent a student's grades in these courses. The available data consist of the grades of five students (in a 10-point numerical scale above the passing mark), as shown in Table 13.1. Using an appropriate statistical package (SPSS, Minitab, SAS etc) run a factor analysis and interpret the results.

Table: 13.1: Students Grade

Grade in (→) Student No.(↓)	Finance (X_1)	Marketing (X_2)	Business Policy (X_3)
1	3	6	5
2	7	3	3
3	10	9	8
4	3	9	7
5	10	6	5

(Example adapted from Peter Tryfos, 1997, version printed: 14-3-2001).

13.11 LET US SUM UP

Factor analysis is something like an art. Factor analysis is used to study the patterns of relationship among many dependent variables to have an idea about the nature of the independent variables not measured directly.

Factor analysis is used to identify latent constructs or factors. It is commonly used to reduce variables into a smaller set factors to save time and facilitate easier interpretations. There are many extraction techniques such as Principal component method and Maximum Likelihood. The factor analysis and the principal component analysis are among the oldest of the multivariate statistical methods of data reduction. Mathematically, factor analysis is complex and the criteria used to determine the number and significance of factors are vast. There are two types of rotation techniques – orthogonal rotation and oblique rotation. Orthogonal rotation (e.g., Varimax and Quartimax) involves uncorrelated factors whereas oblique rotation (e.g., Direct Oblimin and Promax) involves correlated factors. The interpretation of factor analysis is based on rotated factor loadings, rotated eigenvalues, and scree test. In reality, investigators often use more than one extraction and rotation technique based on pragmatic reasoning rather than theoretical reasoning.

13.12 KEY WORDS

Communality : It shows the variance of an observed variable accounted for by the common factors; in an orthogonal factor model. It is equivalent to the sum of the squared factor loadings.

Common Factor : It implies the unmeasured (or hypothetical) underlying variable which is the source of variation in at least two observed variables under consideration.

Confirmatory Factor Analysis (CFA) : This technique of the factor analysis attempts to confirm hypotheses and uses path analysis diagrams to represent variables and factors.

Exploratory Factor Analysis (EFA) : The Exploratory Factor Analysis technique tries to uncover complex patterns by exploring the dataset and testing predictions.

Diagonal Matrix	:	It is a square matrix in which the entries outside the main diagonal are all zero.
Eigenvalue (characteristic root)	:	It is a mathematical property of a matrix; used in relation to the decomposition of a covariance matrix, both as a criterion of determining the number of factors to extract and a measure of variance accounted for by a given dimension.
Eigenvector	:	It is a vector associated with its respective eigenvalue; obtained in the process of initial factoring; when these vectors are appropriately standardized, they become factor loadings.
Factor Extraction	:	It is the initial stage of factor analysis in which the covariance matrix is resolved into a smaller number of underlying factors or components.
Factors	:	It implies hypothesized, unmeasured, and underlying variables which are presumed to be the sources of the observed variables; often divided into unique and common factors.
Factor Loading	:	It is a general term referring to a coefficient in a factor pattern or structure matrix.
Factor Pattern Matrix	:	It refers to a matrix of coefficients where the columns usually refer to common factors and the rows to the observed variables; elements of the matrix represent regression weights for the common factors where an observed variable is assumed to be a linear combination of the factors; for an orthogonal solution, the pattern matrix is equivalent to correlations between factors and variables.
Kaiser criterion	:	In this criterion, first we can retain only factors with eigenvalues greater than 1 when factors are calculated from correlation matrix. In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, we drop it. This criterion was proposed by Kaiser (1960), and is probably the one most widely used.
Orthogonal Factors	:	It indicates the factors that are not correlated with each other; factors obtained through orthogonal rotation.
Orthogonal Rotation	:	It refers to the operation through which a simple structure is sought under the restriction that factors be orthogonal (or uncorrelated); factors obtained through this rotation are by definition uncorrelated.
Principal Axis	:	It is a method of initial factoring in which the

Factoring		adjusted correlation matrix is decomposed hierarchically; a principal axis factor analysis with iterated commonalities leads to a least squares solutions of initial factoring.
Principal Components	**:**	It reflects linear combinations of observed variables, possessing properties such as being orthogonal to each other, and the first principal component representing the largest amount of variance in the data, the second representing the second largest and so on; often considered variants of common factors, but more accurately they are contrasted with common factors which are hypothetical.
Scree Test	**:**	It is a rule of thumb criterion for determining the number of significant factors to retain; it is based on the graph of roots (eigenvalues); claimed to be appropriate in handling disturbances due to minor (unarticulated) factors.
Varimax	**:**	It is a method of orthogonal rotation which simplifies the factor structure by maximizing the variance of a column of the pattern matrix.

13.13 SOME USEFUL BOOKS

1) Anderson, T. W. (1984); *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Second Edition.

2) Bartlett, M.S. (1954); A note on Multiplying factors for various Chi-squared Approximations, Journal of the Royal Statistical Society, 16, 296-298.

3) Johnson R.A., & Wichern D. W. (2002); *Applied Multivariate Statistical Analysis*, Pearson Education, Inc.

4) Johnson, Dallas E. (1998); *Applied Multivariate Methods for Data Analysis*, International Thomson Publishing Inc.

5) Kaiser, H.F. (1958); The Varimax criterion for analytical rotation in factor analysis, Psychometrica, 23, 187-200.

6) https://onlinecourses.science.psu.edu

7) https://en.wikipedia.org/wiki/Factor_Analysis

8) http://www.researchgate.net/file.PostFileLoader.html?id=53c50b3fd5a3f2140c8b465e&assetKey=AS%3A271747469774848%401441801053551.

13.14 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 13.1

2) See Section 13.4

3) See Section 13.4

4) See Section 13.4

Check Your Progress 2

1) See Section 13.6

2) See Section 13.7

3) Hints: How to run Factor Analysis in SPSS?

After opening the SPSS windows sheet in your computer system, the following are the steps which can be used for performing, evaluating and analyzing the factor analysis solution:

Step 1 : Feed the entries in the SPSS work sheet.

Step 2 : Go to "Analyze" in the menu available at the top of the window screen.

Step 3 : Click on "Data Reduction"

Step 4 : Click on "Factor"

Step 5 : Click on "Extraction" and select "principal axis factoring" and press "Continue"

Step 6 : Click on "OK"

13.15 EXERCISES

1) What do you understand by the Factor score? Describe the different methods for the estimation of factor score.

2) Write a short note on Factor rotation and Oblique rotation.

3) What do you understand by the methods of estimation? Describe in details, the principal component and maximum likelihood methods.

4) With suitable example, explain communality in context to factor analysis.

5) Define the following terms:

   a) Correlation matrix

   b) Specific variance

   c) Factor loadings

   d) Specific factors

UNIT 14 CANONICAL CORRELATION ANALYSIS

Structure

14.0 Objectives

14.1 Introduction

14.2 Canonical Correlation Analysis (CCA): Concept and Meaning

14.3 Assumptions of Canonical Correlation

14.4 Canonical Correlation Analysis as Generalization of the Multiple Regression Analysis

14.5 Steps and Procedure Involved in Computation of CCA Results

14.6 Illustration of CCA

14.7 Interpretation of CCA Results

14.8 Limitations of Canonical Correlation

14.9 Let Us Sum Up

14.10 Key Words

14.11 Some Useful Books

14.12 Answers or Hints to Check Your Progress

14.13 Exercises

14.0 OBJECTIVES

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

- explain the concept of canonical correlation analysis;
- state the similarity and difference between multiple regression and canonical correlation;
- discuss the steps and procedure involved in canonical correlation;
- elucidate how to interpret the results of canonical correlation;
- point out the limitations of canonical correlation analysis.

14.1 INTRODUCTION

The conventional wisdom that the economic agents are rational and are guided by self-interest in decision making is being questioned as it ignores the psychological and social factors influencing decision making process. The research findings from many disciplines like neuroscience, cognitive science, psychology, behavioral economics, sociology, anthropology etc. indicate that the decisions made by the individuals in many aspects of development (like savings, investment, energy consumption, health etc.) are influenced by social contexts, local social networks, cultural factors, social norms and shared mental models etc. (World Development Report, 2015). Hence, inter-disciplinary perspective is being recognized as research approach to analyze human behavior so as to improve the predictive power of economics. Canonical

correlation analysis is a powerful analytical tool to analyze the association between two sets of variables belonging to different disciplines. This method measures the strength of association between the two sets of variables. An attempt is made in this method to concentrate a high-dimensional relationship between two sets of variables into a few pairs of canonical variables. Hence, in this unit, we shall throw light on various issues relating to canonical correlation like concept and meaning of canonical correlation, its similarity and differences with multiple regression, procedure involved in the analysis of canonical correlation, and its advantages and limitations.

14.2 CANONICAL CORRELATION ANALYSIS (CCA): CONCEPT AND MEANING

Canonical correlation analysis is a multivariate statistical model used to study the interrelationships among sets of multiple dependent variables and multiple independent variables. This technique is distinct from the multiple regression model in the sense that multiple regression predicts a single dependent variable from a set of multiple independent variables whereas canonical correlation simultaneously predicts multiple dependent variables from multiple independent variables.

Let us understand the concept with an example. As a student of economics, you may like to know the association between economic inequality $X^*$and political instability Y^*. The economic inequality can be measured by five variables i.e. (i) the division of farmland (X_1), (ii) the gini coefficient (X_2), (iii) the percentage of tenant farmers (X_3), (iv) the gross national product (X_4), and (v) the percentage of farmers (X_5). Similarly the political instability can be measured by four variables (indicators) i.e. (i) the instability of leadership (Y_1), (ii) the level of internal group violence (Y_2), (iii) the occurrence of internal war (Y_3), (iv) stability of democracy (Y_4). These two theoretical concepts X^* and Y^* can be called two sets of variables or canonical variables. These can be shown in the following figure:

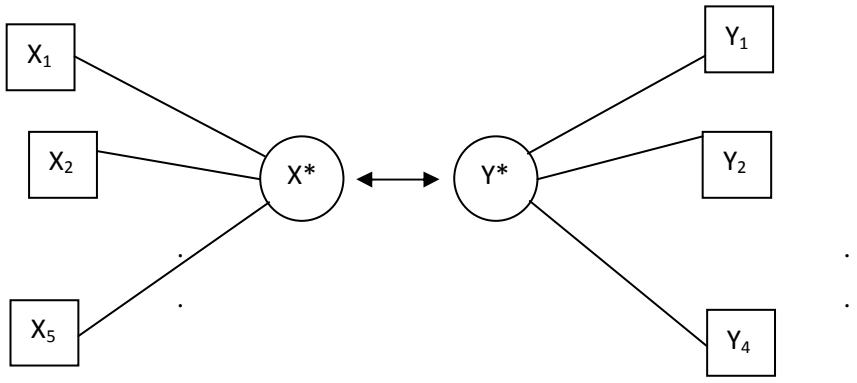


Fig. 14.1: Canonical Correlation

Source: The sage encyclopedia of Social Sciences Research Methods vol. 1 (2004) page no. 83.

The first canonical variable X^* is measured by five variables (P = 5) and can be considered as a linear combination (a weighted sum) of these X variables. The second canonical variable Y^* is a linear combination of the q = 4 indicators, Y_1 to Y_4. The double side curved arrow indicates that the question of casualty remains open.

The purpose of canonical correlation analysis is to find the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set. The idea behind this approach is first to determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the

initially selected pair, and so on. As stated in the above para, the pairs of linear combinations are called the canonical variables and their correlations are called canonical correlations.

Thus, canonical correlation aims to (i) identify the dimensions among the dependent and independent variables, and (ii) maximize the relationship between the dimensions.

In this manner, in canonical correlation, we can distinguish three types of correlations:

1) Correlation between X variables, the correlation matrix is R_{xx}.
2) Correlation between Y variables, the correlation matrix is R_{yy}.
3) Correlation between X and Y variables, the correlation matrix is $R_{xy} = R'_{yx}$.

14.3 ASSUMPTIONS OF CANONICAL CORRELATION

1) The correlation coefficient between any two variables is based on linear relationship.
2) The canonical correlation is the linear relationship between the variates.
3) The distribution of variables is normal.
4) Hetro-scedasticity, to the extent it decreases the correlation between variables.

14.4 CANONICAL CORRELATION ANALYSIS AS A GENERALIZATION OF MULTIPLE REGRESSION ANALYSIS

Multiple regression analysis is a multivariate technique which can predict the value of a single dependent variable from a linear function of a set of independent variables. But this is not always the case. There are real life problems, however, when interest may not center on a single dependent variable. Rather, the researcher may be interested in relationships between sets of multiple dependent and multiple independent variables. Canonical correlation analysis is a multivariate statistical model that facilitates the study of interrelationships among sets of multiple dependent variables and multiple independent variables. Whereas multiple regression predicts a single dependent variable from a set of multiple independent variables, canonical correlation simultaneously predicts multiple dependent variables from multiple independent variables. Therefore, canonical correlation analysis is said to be a generalization of multiple correlation used in multiple regression problems.

The coefficient of determination R^2, in regression problems is the proportion of the variability in a dependent variable that is accounted for by a set of predictor variables and $R = \sqrt{R^2}$ is called the multiple correlation coefficient. The multiple correlation coefficient can also be interpreted as a measure of the maximum correlation that is attainable between the dependent variable and any linear combination of the predictor variable. Canonical correlation places the fewest restrictions on the types of data on which it operates. Because the other techniques impose more rigid restrictions, it is generally believed that the information obtained from them is of higher quality and may be presented in a more interpretable manner. For this reason, many researchers view canonical correlation as a last effort, to be used when all other higher-level techniques have been exhausted. But in situations with multiple dependent and independent variables, canonical correlation is the most appropriate and powerful multivariate technique. It has gained acceptance in many fields and represents a useful tool for multivariate analysis, particularly while considering multiple dependent variables.

14.5 STEPS AND PROCEDURE INVOLVED IN COMPUTATION OF CCA

In 1935-36, Hotelling proposed a method, known as Canonical Correlation Analysis to investigate "linear" relationship between the two sets of variates.

James Press (2005) has expressed the whole idea of this Canonical Correlation Analysis in the following words:

"The Canonical correlation" model selects weighted sums of variables from each of the two sets to form new variables in each of the sets, so that the correlation between the new variables in "different sets" is maximized while the new variables within each set are constrained to be uncorrelated with mean zero and unit variance.

We shall adhere to Press's approach.

Let

$$\alpha: p_1 \times 1$$

and

$$\gamma: p_2 \times 1$$

be two unknown vectors to be determined such that the correlation between $\alpha'Y$ and $\gamma'Z$ be as large as possible.

So, let

$$U_1 = \alpha'Y$$

$$V_1 = \gamma'Z$$

and

$$\rho(U_1, V_1) = \text{Correlation coefficient between } U_1 \text{ and } V_1.$$

The problem of correlation coefficient now amounts to the following:

$$\text{maximize } \rho(U_1, V_1)$$

Subject to

$$\text{Var}(U_1) = \text{Var}(V_1) = 1$$

and

$$E(U_1) = E(V_1) = 0.$$

Hotelling solved this problem using the celebrated method of Lagrange Multipliers. However, we shall just state the final results. Hotelling showed that this maximization problem is equivalent to following algorithm:

Step 1: Solve for λ the equation

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0 \qquad \text{----------} \qquad \text{(A)}$$

Here, $\Sigma_{11} = \text{Var}(Y)$

$\Sigma_{22} = \text{Var}(Z)$

$\Sigma_{12} = \Sigma_{21} = \text{Var}(Y, Z)$

Let λ_1 be the largest positive root of the above equation.

Step 2: Now, solve the system of equations:

$$\begin{pmatrix} -\lambda_1\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda_1\Sigma_{22} \end{pmatrix}\begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = 0$$

for α and γ.

Mathematically, it is not so simple to solve the above system of linear equations. However, there is an equivalent formulation. To compute α and γ, we solve the pair of equations:

$$(\Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{21} - {\lambda_1}^2\Sigma_{11})\alpha = 0$$

$$(\Sigma_{21}{\Sigma_{11}}^{-1}\Sigma_{12} - {\lambda_1}^2\Sigma_{22})\gamma = 0$$

Step 3: We now call these α and γ as α_1 and γ_1 respectively; and

$$U_1 = \alpha_1'\gamma$$

and

$$V_1 = \gamma_1'Z$$

as **First Canonical Variates**.

It will turn out that $\lambda_1$will be the correlation coefficient between $U_1$and V_1. We, therefore, write

$$\lambda_1 = \rho(U_1, V_1)$$

and call this as **First Canonical Correlation**.

Step 4: We now proceed to the next iteration. We now define:

$$U_2 = \alpha_2'\gamma$$

$$V_2 = \gamma_2'Z,$$

where α_2 and γ_2 are to be determined.

$$\text{maximize } \rho(U_2, V_2)$$

subject to

$$\text{Var}\,(U_2) = \text{Var}\,(V_2) = 1$$

and

$$E\,(U_2) = E(V_2) = 0.$$

We repeat the above procedure to compute α_2 and γ_2 as solution of

$$\begin{pmatrix} -\lambda_2\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda_2\Sigma_{22} \end{pmatrix}\begin{pmatrix} \alpha_2 \\ \gamma_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

where λ_2 is the second largest positive root of the equation (A).

Step 5: We continue this procedure uptil the smallest positive root.

Step 6: The result is now to be collected in a vector format:

$$U = \begin{pmatrix} U_1 \\ U_2 \\ . \\ . \\ . \\ U_{p_1} \end{pmatrix} = (U_1 \ldots U_{p_1})'$$

and

$$V = \begin{pmatrix} V_1 \\ V_2 \\ . \\ . \\ . \\ V_{p_1} \end{pmatrix} = (V_1 \ldots V_{P1})'.$$

The elements of U and V are called **Canonical Variates** and λ_i is corresponding canonical correlation.

14.6 ILLUSTRATION OF CCA

Let us now understand the canonical correlation analysis technique with the help of an illustration.

This example is based on 416 observations collected through primary survey for the purpose of research study entitled "Assessment of Human Well-being in Delhi: Multidimensional Approach". The researcher attempted to study the relationship between economic well-being variables and overall life satisfaction variables. One of the questions raised in this study was to examine how does economic well-being influence the overall life satisfaction of the people or more specifically, to know whether economic well-being indicators are predictive of overall life satisfaction of people.

The main characteristic of canonical analysis is the investigation of the relationship between two sets of variables. One set is the predictor set or, say, analytically, the set of independent variables. The second consists of the criteria or dependent variables. In our example, the set of economic well-being variables constitutes the set of independent variables. This set of economic well-being variables consist of 5 indicators of economic well-being:

1) Annual Income
2) Movable Assets
3) Fixed Assets
4) Employment Status
5) Educational Attainment

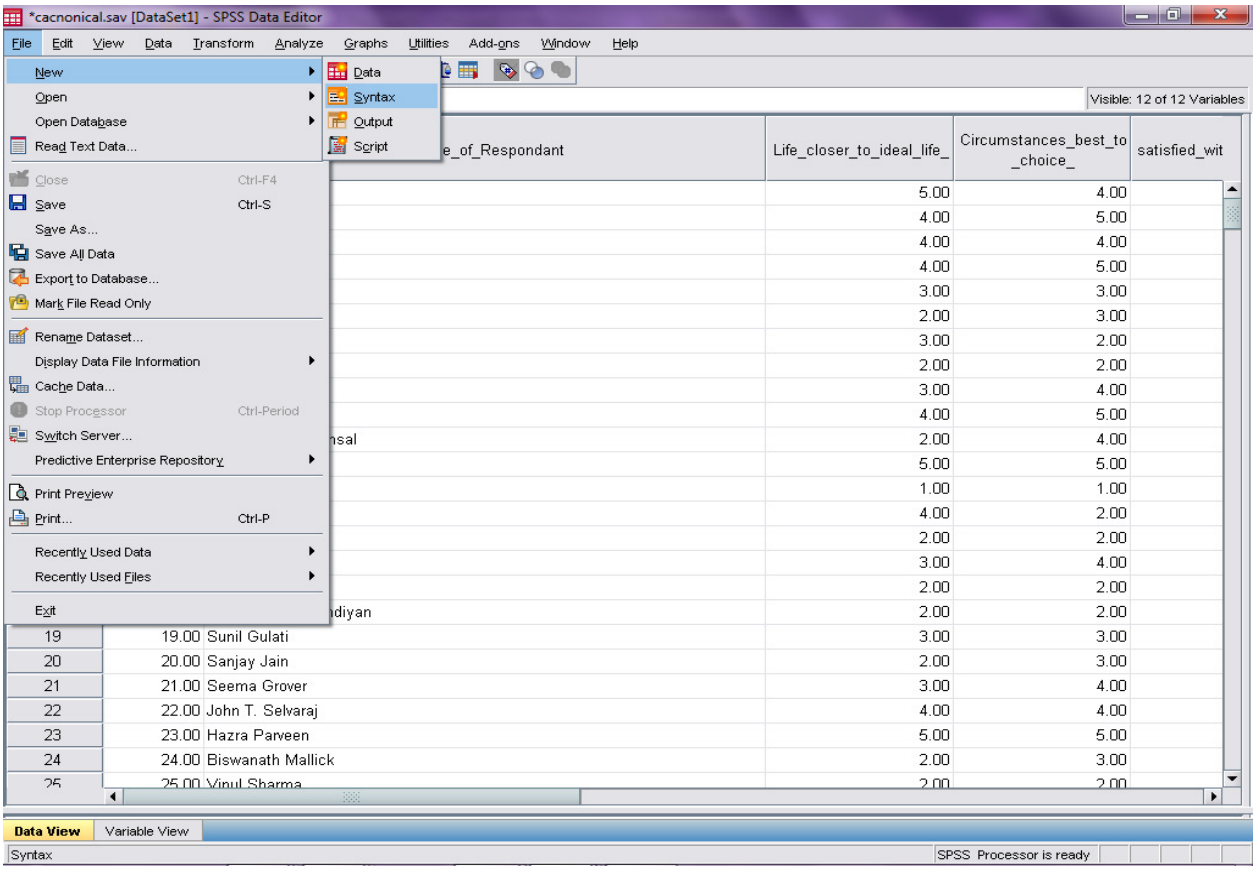
Next, the set of criteria variables constitute the overall life satisfaction indicators:

1) My life closer to my ideal life
2) Circumstances of my life are best to my choice
3) I am satisfied with my life
4) I have achieved the things in my life I aspired
5) I would not prefer to make any drastic change in my rest of life till I survive

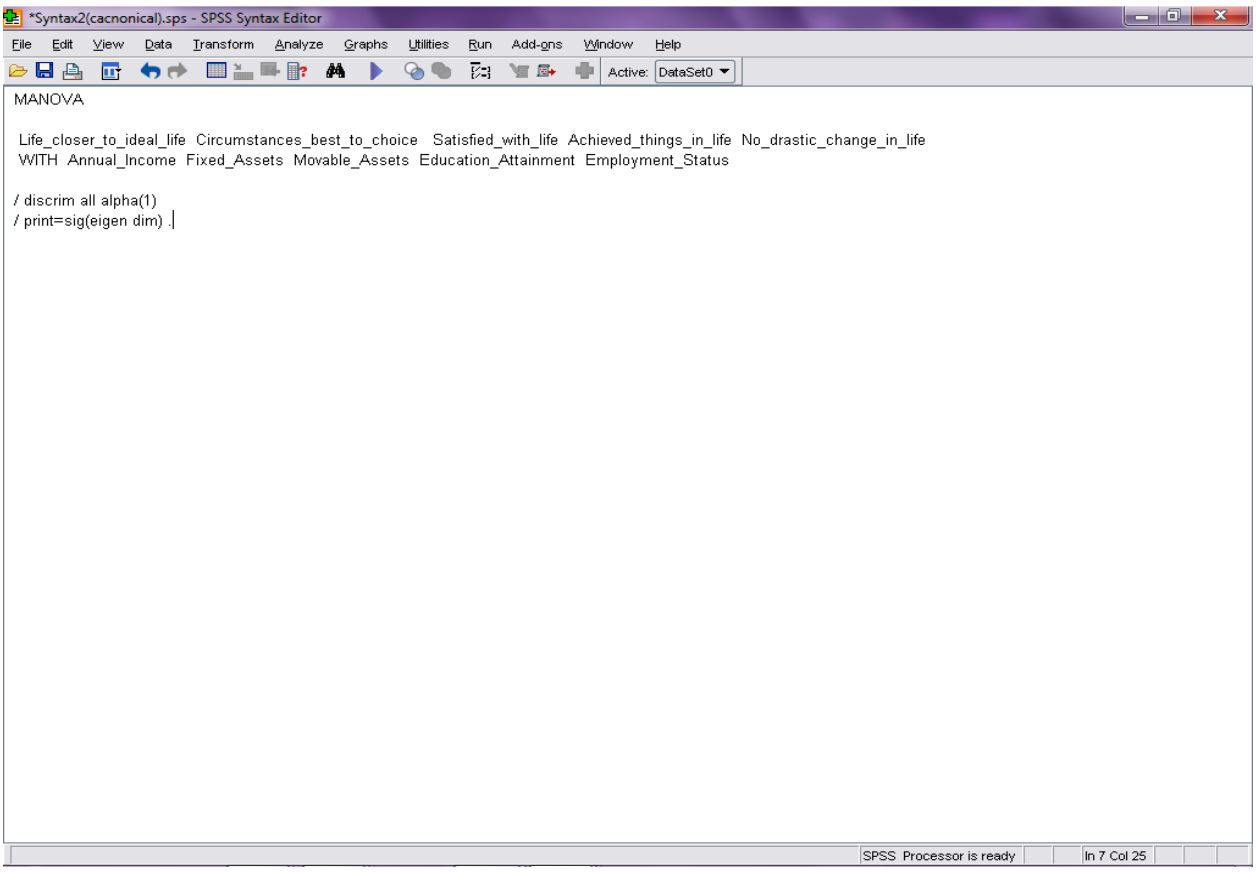
The data on these two sets of variables was collected through primary survey for the above said study.

Let us now learn to run the Canonical Correlation Analysis for our example using SPSS

Step 1: Click file, new and syntax sequence.

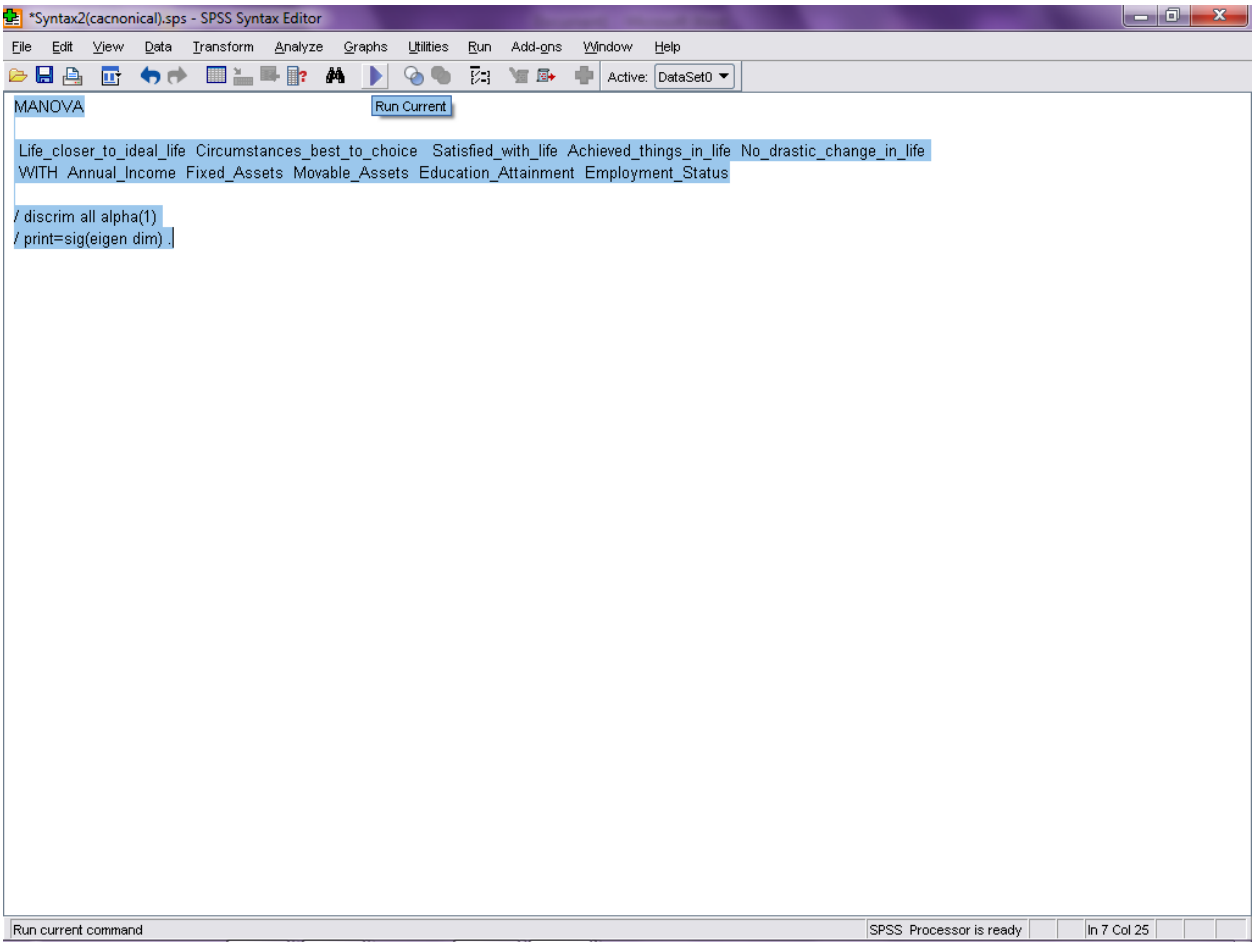


Step 2: Then type the following syntax :



Note that the criterion set of variables are listed before **WITH** and predictor variables are listed afterwords

Step 3: This command can be implemented using RUN button on toolbar menu.



The results of above illustration using SPSS has been presented in this section. The table 1 below shows an overall multivariate test of the entire model using different multivariate criteria.

Table 14.1: Multivariate Tests of Significance

Test Name	Value	Approx. F	Hypoth. DF	Error DF	Sig. of F
Pillais	.14258	2.40693	25.00	2050.00	.000
Hotellings	.15845	2.56302	25.00	2022.00	.000
Wilks	.86055	2.49153	25.00	1509.72	.000
Roys	.11708				.000

Table 14.2: Eigenvalues and Canonical Correlations

Root No.	Eigenvalue	Pct.	Cum. Pct.	Canon Cor.	Sq. Cor
1	.13260	83.69055	83.69055	.34217	.11708
2	.01720	10.85381	94.54437	.13003	.01691
3	.00698	4.40258	98.94695	.08323	.00693
4	.00167	1.05222	99.99917	.04080	.00166
5	.00000	.00083	100.00000	.00115	.00000

In this example, we are not only interested to know whether there is a relationship between the predicator and criterion variables but also wanted to know which indicators of economic well-being are more or less useful in explaining the relationship between life satisfaction and economic well-being. This is exactly where CCA comes into play. The interpretation and evaluation of the results has been discussed in next section.

14.7 INTERPRETATION OF CCA RESULTS

We now come to the conclusive stages of the Canonical Correlation Analysis. To draw the right conclusions amounts to the correct interpretation of the results obtained by conducting the CCA.

The first step is to evaluate the overall statistical significance of the full canonical model. This is done by testing the Null hypothesis. The Null hypothesis is that there is no relationship between the two set of variables. The alternative hypothesis is that the two sets of variables are related.

The interpretation of CCA is now accomplished by computing the F-statistic value using the Wilks-λ test.

We now come to the actual interpretation of the CCA as conducted in our example. Here, the value of the computed value of the F-statistic is high and significant. This is given in table 1 as 0.860. This means that we can reject the Null hypothesis. That is to say that we accept the alternative hypothesis. Thus, statistically significant relationship exists between life satisfaction and economic well-being indicators.

Let us recall, that the first root (function) is created in such a manner that the canonical correlation between the new variable is maximized and these new variables within each set are uncorrelated with zero mean and unit variance.

Let us interpret only those functions which explain reasonable amount of variance between variable sets. In our illustrations, we interpret only the first function as it explains 12% variance within the function as shown in table no. 2. All other functions explain less than 10% of variance in their functions, hence we can ignore them.

So, what we have concluded so far is that there is a statically significant relationship between our two set of variables. Further, this relationship is largely captured by the first root (function) in the canonical model. Next, we want to identify those variables which contribute significantly to explain this relationship between economic well-being and overall life satisfaction.

In multiple regression analysis, we often look at Beta weights to identify the relative contribution of one independent variable to explain dependent variable. In CCA, we look at the structure coefficient to decide which variables are useful for the model. Therefore, we examine the standardized weights and structure coefficients to interpret the first root (function). Let us underline the point that we are only concerned with the first function and will ignore other functions as they are not significant.

To understand the pattern among two set of variables, we have created table 3 showing coefficients which presents the standardized canonical function coefficients (i.e. weights) and structure coefficients for all variables. The squared structure coefficient (r_s^2) are also given, which represent the percentage of shared variance between observed variable and the new variable created from the observed variables set.

Table 14.3: Canonical Solution for Economic Well-being Indicator Predicting Overall Life Satisfaction for Function 1

Variable	Coef	r_s	r_s^2 (%)
Life closer to ideal life	.221	-.502	25.20
Circumstances of life best to my choice	-.692	-.905	81.90
Satisfied with life	-.223	-.729	53.14
Achieved things in life	-.400	-.746	55.65
Want no drastic change in life	-.054	-.427	18.23
Annual Income	-.381	-.857	73.44
Fixed Assets	.029	-.256	6.55
Movable Assets	-.210	-.696	48.44
Educational Attainment	-.668	-.895	80.10
Employment status	.219	-.286	8.17

Coef: Standized Canonical Function Coefficient, r_s = Structure Coefficient, r_s^2 = Squared Structure Coefficient

Looking at the coefficient, we find that relevant dependent variables are: the circumstances of life are best to my choice, I am satisfied with my life and I have achieved things in my life I aspired, because all these variables high squared structure coefficient which indicates the amount of variance the observed variable can contributed to new latent criterion variable. Looking at the other side of the equation on function 1 which involves predicator set, we find that Annual Income, Movable Assets and Educational Attainment variables were primary contributors. You must note that this process for interpretation of function is same as identifying the useful predictors in regression analysis with the exception that in canonical correlation analysis we have two set of equations for consideration.

Check Your Progress I

1) What do you mean by the term 'inter-disciplinary perspective'

……

……

……

2) How is CCA useful to address the inter-disciplinary nature of research questions?

……

……

……

3) State the steps involved in computation of CCA through SPSS software.

……

……

……

14.8 LIMITATIONS OF CANONICAL CORRELATION

1) The canonical correlation express the variance shared by the linear composites of the set of variables and not the variance extracted from the variables.

2) Canonical weights derived in computing canonical functions are subject to great deal of instability.

3) The interpretation of canonical variates is different due to the efforts to maximize the relationship.

4) It is difficult to identify meaningful relationship between the subset of independent and dependent variables because precise statistics is yet to be developed.

14.9 LET US SUM UP

Canonical correlation analysis is a useful and powerful technique for exploring the relationships among multiple dependent and independent variables. The technique is primarily descriptive, although it may be used for predictive purposes. In this technique, weighted sums of variables are selected from each of the two sets to form new variables in each of the sets so that the correlation between the new variables in different sets in maximized while the new variables within each set are constrained to be uncorrelated with mean zero and unit variance. Results obtained from a canonical analysis suggest answers to questions concerning the number of ways in which the two sets of multiple variables are related, the strengths of the relationships, and the nature of the relationships defined. Canonical analysis enables the students to combine into a composite measure what otherwise might be an unmanageably large number of bivariate correlations between sets of variables. It is useful for identifying overall relationships between multiple independent and dependent variables, particularly when we have little a priori knowledge about relationships among the data for two sets of variables. Essentially, we can apply canonical correlation analysis to a set of variables that appear to be significantly related.

The CCA is based on two statistical assumptions. First, the correlation coefficient between any two variables is based on a linear relationship. Second, the parent population from which the sample has been drawn is normally distributed. CCA has several advantages and limitations.

14.10 KEY WORDS

Canonical correlation : Measure of the strength of the overall relationships between the linear composites (canonical variates) for the independent and dependent variables. In effect, it represents the bivariate correlation between the two canonical variates.

Canonical correlation analysis : It is a multivariate statistical analysis that facilitates the study of interrelationships among the sets of multiple dependent variables and multiple independent variables.

Canonical cross-loadings : Correlation of each observed independent or dependent variable with the opposite canonical variate. For example, the independent variables are correlated with the dependent canonical variate. They can be interpreted like canonical loadings, but with the opposite canonical variate.

Canonical function : Relationship (correlation) between two linear composites (canonical variates). Each canonical function has two canonical variates, one for the set of dependent variables and one for the set of independent variables. The strength of the relationship is given by the canonical correlation.

Canonical loadings : Measure of the simple linear correlation between the independent variables and their respective canonical variates. These can be interpreted like factor loadings, and are also known as canonical structure correlations.

Canonical roots : Squared canonical correlations, which provide an estimate of the amount of shared variance between the respective optimally weighted canonical variates of dependent and independent variables. It is also known as eigenvalues.

Canonical variates : Linear combinations that represent the weighted sum of two or more variables and can be defined for either dependent or independent variables. It is also known as linear composites, linear compounds, and linear combinations.

Multiple regression analysis : Multiple regression analysis predicts a single dependent variable from a set of multiple independent variables.

14.11 SOME USEFUL BOOKS

1) Alissa Sherry, Robin K. Henson (2005); Conducting and Interpreting Canonical Correlation Analysis in Personality Research. Journal of Personality Assessment.
2) Hotelling, H. (1936); *Relations between two sets of variables*, Biometrika 28, 321-377
3) Johnson, R.A. & Wichern, D.W. (2002); *Applied Multivariate Statistical Analysis*, Pearson Education, Inc.
4) Johnson, Dallas E. (1998); *Applied Multivariate Methods for Data Analysis*, International Thomson Publishing Inc.
5) Michael S. Lewis-beck Alan Bryman Tim Futing Liao (2004); *The sage encyclopedia of Social Sciences Research Methods* vol. 1 page no. 83
6) Magnus Borga; Canonical Correlation. A tutotial www.cs.cmu.edu/~tom/10701-spll/sides/cca-totorial.pdf
7) S. Press James (2005): Applied Multivariate Analysis. Dover publication.inc.

14.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress I

1) See Section 14.1
2) See Section 14.6
3) See Section 14.6

14.13 EXERCISES

1) Under what circumstances would you select canonical correlation analysis instead of multiple regressions as the appropriate statistical technique?
2) Discuss in details the procedure for computation of the canonical correlation.
3) What are the limitations associated with canonical correlation analysis?

UNIT 15 CLUSTER ANALYSIS

Structure

15.0 Objectives
15.1 Introduction
15.2 Cluster Analysis: Concept and Meaning
15.3 Steps and Algorithm Involved in Cluster Analysis
15.4 Methods of Cluster Analysis
15.5 Partitioning Cluster Methods
15.6 Hierarchical Cluster Methods
15.7 Other Approaches: Two-step Cluster Analysis
15.8 Interpretation of the Results
15.9 Let Us Sum Up
15.10 Key Words
15.11 Some Useful Books
15.12 Answers or Hints to Check Your Progress Exercises
15.13 Exercises

15.0 OBJECTIVES

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

- state the concept and purpose of cluster analysis;
- list the steps to be followed in cluster analysis;
- explain the different approaches to cluster analysis; and
- to learn how to apply cluster analysis in analyzing economic problems and interpret its results.

15.1 INTRODUCTION

In social sciences, data set usually take the form of observations on unit of analysis for a set of variables. You as a student of economics may be interested to identify groups of similar objects like countries, enterprises, households on the basis of selected variables like unemployment rate of men and women in different countries, deprivation indicators of households etc. Similarly you may need to segment customers into a small number of groups for additional analysis and marketing activities. Further researcher may aim to find out spatial or temporal pattern in human developments, educational development etc. In such cases, a simple classification of units into sub-groups is required for the purpose of analysis because classes or conceptually meaningful groups (clusters) share common characteristics and play an important role in analysis and description of the world. Cluster analysis is a useful technique in such situations for data analysis. For example, cities can be grouped or clustered in

terms of their social, economic and demographic characteristics. Similarly, people can be clustered in terms of their psychological profiles or other attributes they possess. Hence, in this unit, we shall discuss various clustering methods and algorithm involved in cluster analysis. Let us begin with explaining the concept of cluster analysis.

15.2 CLUSTER ANALYSIS: CONCEPT AND MEANING

Classes or conceptually meaningful groups of objects sharing common characteristics, play an important role in how people analyze and describe the world. The creation of groups of similar objects or variables based upon measured characteristics is an important method of analysis which we call cluster analysis. This is an important method of multi-variate analysis. After collecting the data, we divide the observations into groups (clustering) and assign the labels to these groups. Thus cluster analysis is the study of techniques for automatically finding classes. Cluster analysis is a strong tool of the multivariate exploratory data analysis. It involves a great amount of techniques, methods and algorithms which can be applied in various fields of social sciences including economics.

Let us remember that Grouping, or clustering, is distinct from the classification methods. Classification pertains to known number of groups with the purpose to assign new observations to one of these groups. Cluster analysis is a more primitive technique in the sense that no assumptions are made concerning the number of groups or the group structure. Grouping is done on the basis of similarities or distances (dissimilarities). The inputs required are similarity measures or data from which similarities can be computed. The general practical application of cluster analysis is whether the investigator knows enough about the problem to distinguish "good" groupings from "bad" groupings. This can be done by enumerating all the possible groupings and select the "best" ones for further study.

Cluster analysis aims to discover natural groupings of the items (or variables). For this, we need to first develop a quantitative scale to measure the association (similarity) between objects. Since clustering involves a collection of data objects which are similar to one another within the cluster and dissimilar to objects in other clusters, a cluster of data objects can be treated as an implicit class. Due to this it is also sometimes referred to as ‘automatic classification’. In some applications, it is also known as ‘data segmentation’ because clustering divides large data sets into groups based on their similarities. Cluster analysis enables us to detect outliers. As a branch of statistics, cluster analysis has been extensively studied with the main focus on distance-based cluster analysis and has been built into statistical analysis software packages or systems such as SPSS, SAS and S-Plus. An illustration of cluster analysis using SPSS shall be discussed towards the end of the unit. It is, however, first important to understand the steps and algorithm involved in conducting cluster analysis and the main methods used under this analytic technique in research.

15.3 STEPS AND ALGORITHIM INVOLVED IN CLUSTER ANALYSIS

The following five broad steps are involved in conducting cluster analysis:

1) Measuring the relevant variables (both quantitative and categorical) that can be included,

2) Creating a (dis) similarity Matrix for an appropriate measure of (dis) similarity,

   Distances are used to measure (dis) similarity between the samples of two variables. To determine the extend of (dis) similarity we construct what is known as (dis) similarity matrix. Every entry of (dis) similarity matrix determine using the notion of distance, more formally metric. Various sort of metric can be employed each being suitable to suit the particular sort of data set.

   The metric that is usually employed is known as *minskowi* metric, the formula of which is as under:

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^k \right)^{\frac{1}{k}}$$

   where $x_i = (x_1, x_2, x_3, \ldots, x_n)$ and $y_i = (y_1, y_2, y_3, \ldots, y_n)$ are two n – dimensional data objects and k is a positive integer.

   For k = 1 and k = 2 the M*inskowi* metric becomes equal to *Manhattan* and *Euclidean* metric respectively.

3) Creating one or more clustering's via a clustering algorithm. An example for clustering algorithm:

   For the sake of convenience, we shall consider a sub matrix from the table of (dis) similarity matrix shown in table 1.

   Let us take

$$A = \begin{pmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ 0 & 1 & 0 & & & \\ 1 & 2 & 1 & 0 & & \\ 0 & 1 & 0 & 1 & 0 & \\ 16 & 17 & 16 & 9 & 16 & 0 \end{pmatrix}$$

   Now, we implement the algorithm.

   Step 1: The entries in A are squared Euclidean distances. We first take their positive square root and call the new matrix as D_1

$$D_1 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & & & & & \\ 1 & 0 & & & & \\ 0 & 1 & 0 & & & \\ 1 & \sqrt{2} & 1 & 0 & & \\ 0 & 1 & 0 & 1 & 0 & \\ 4 & \sqrt{17} & 4 & 3 & 4 & 0 \end{pmatrix}$$

We label rows as 1, 2, 3, 4, 5, 6. Taking the row 1 as a fixed variable, we now locate the variable closest to it. We observe that there are two such choices (1,3) & (1,5) as both have distance zero. We take (1,3) and put 1 & 3 in the same cluster. Now we calculate the distance of (1,3) from remaining rows with the understanding that

$$d_{ij} = |d_{ij} - d_{ji}| \quad \forall\, i,j = 1(1)5$$

$$d_{(13)2} = \min\{d_{12}, d_{32}\} = \min\{1,1\} = 1$$

$$d_{(13)4} = \min\{d_{14}, d_{34}\} = \min\{1,1\} = 1$$

$$d_{(13)5} = \min\{d_{15}, d_{35}\} = \min\{0,0\} = 0$$

$$d_{(13)6} = \min\{d_{16}, d_{36}\} = \min\{4,4\} = 4$$

Hence, now D_2 becomes

$$D_2 = \begin{matrix} 13 \\ 2 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ 1 & \sqrt{2} & 0 & & \\ 0 & 1 & 1 & 0 & \\ 4 & \sqrt{17} & 3 & 4 & 0 \end{pmatrix}$$

We observe that because 1 & 3 were clustered together, therefore, 3rd row so does the 3rd column from D_1 gets deleted.

Step 2: We now see that taking (13) as the fixed variable, the variable closest to (13) in the matrix D_2 is row 5 as their distance is zero. Hence, we cluster together (13) & 5.

Now we calculate distance of (13)5 from the remaining variable to construct matrix D_3.

Hence now D_3 becomes

$$D_3 = \begin{matrix} (13)5 \\ 2 \\ 4 \\ 6 \end{matrix} \begin{pmatrix} 0 & & & \\ 1 & 0 & & \\ 1 & \sqrt{2} & 0 & \\ \sqrt{17} - 1 & \sqrt{17} & 3 & 0 \end{pmatrix}$$

We can continue like this and it is entirely up to us when to stop depending on how many clusters we want.This method of creating clusters is applicable in both hierarchical as well as non hierarchical approach to cluster analysis.

4) providing some assessment of the obtained clustering(s), and
5) Interpreting the clustering(s) in substantive terms.

Assessment and interpretation of the obtained clustering (s) has been explained in section 15.8 with the help of an illustration.

15.4 METHODS OF CLUSTER ANALYSIS

Imagine you are going to undertake fieldwork in Uttar Pradesh next week and for doing so you wish to divide your group into 5 teams and assign 10 villages under one team each. Strategically, you would like to identify those set of 10 villages to bear somewhat similar characteristics as much possible. Suppose that the district information is also unavailable leaving manual grouping difficult. You need a clustering tool to help you in this scenario. Clustering is a process of grouping a set of objects into multiple groups or clusters so that objects having high similarity are in same cluster but having dissimilarity in other clusters. Similarities and dissimilarities are generally assessed based on attribute values describing objects and involve distance measures.

Creation of simple group structure from a complex data set require a measure of "closeness" or "similarity". Often a great deal of subjectivity is involved in the choice of a similarity measure. Important considerations include the nature of the variables (discrete, continuous, binary), scales of measurement (nominal, ordinal, interval, ratio), and subject matter knowledge. When items (units or cases) are clustered, proximity is usually indicated by some sort of distance. On the other hand, variables are usually grouped on the basis of correlation coefficients or like measures of association.

An advantage of using Cluster analysis is to represent data in graphical way as it offers several possibilities to do that. Dendrogram or other graphical solutions can be used to visually examine linkages between observations. A dendrogram is a tree-like structure commonly used to represent the process of hierarchical clustering. It shows how objects are grouped together (in an agglomerative method) or partitioned (in a divisive method) on a step-by-step basis.

Two types of clustering algorithms are generally reported, namely non-hierarchical and hierarchical. Partitioning is the most extensively used non-hierarchical method and would be discussed in detail in this unit. While partitioning method classifies objects into a specified number of groups (say k), hierarchical method does not construct a single partition with k clusters, but deals with all values of k in the interval [1,n] where n is the total number of objects. Other types of non-hierarchical clustering methods are density-based and grid-based which are beyond the scope of this unit.Non-hierarchical cluster analysis tends to be used when large data sets are involved. It is sometimes preferred because it allows subjects to move from one cluster to another (this is not possible in hierarchical cluster analysis where a subject, once assigned, cannot move to a different cluster). The two major disadvantages of non-hierarchical cluster analysis are:

1) It is often difficult to know how many clusters you are likely to have and therefore the analysis may have to be repeated several times, and
2) It can be very sensitive to the choice of initial cluster centres.

15.5 PARTITIONING CLUSTERING METHODS

As introduced earlier, a partitioning method classifies objects into say, k groups which together must satisfy the following criteria:

Each group must contain at least one object and each object must belong to one group. To put it simply, it implies that k should be less than or equal to n, where n is the total number of objects. In other words, partitioning methods conduct one-level partitioning on data sets. The basic partitioning methods typically adopt 'exclusive cluster separation' i.e. each object must belong to exactly one group. Stratified random sampling lend an interesting example of partitioning methods. Suppose a household sample size of 100 has to be collected from Delhi representing all five geographical zones of Delhi i.e. East, West, North, South and Central zones. In this case the five zones are the 5 groups and are pre-specified in number. Each zone must contain at least one household and each household in the sample must belong to one zone. Clearly in this example, the number of groups k (here 5) are less than the number of objects n (here 100).

In general, after initial partitioning, the 'iterative relocation technique' attempts to improve the partitioning by moving objects from one group to another. When it comes to judging a good partitioning, it can be inferred on the basis of whether objects in the same cluster are 'close' or related to each other whereas objects in different clusters are 'far apart'.

We shall now discuss one of the most frequently used partitioning method in social sciences and in statistical analysis and that is the **K-means method**. It is a Centroid based technique. K-means clustering is used to split the observations into K clusters/groups and test what are the main characteristics of these clusters/groups. To understand it in simple words, let us assume that we have a data set D which contains n objects in Euclidian space i.e. two dimensional space. Partitioning methods first distribute the objects in data set D into k clusters such as $X_1, X_{2,...}X_k$,wherein each X_i is an element of D and $X_i \cap X_j = \emptyset$. Thereafter an objective function is defined to assess the partitioning quality so that within a cluster there are similar objects while other clusters have dissimilar objects.

As mentioned earlier, K-means is a centroid based method which uses the centroid of a cluster to represent that cluster. Conceptually, centroid is nothing but its center point or mean. It is, many a times, defined as the mean of the objects assigned to the cluster. After calculating the mean, this methods measures the distance of objects from the centroid in the Euclidian space. The quality of a cluster is measured by the within-cluster variation, which is the sum of squared error between all objects in X_i and the centroid c_i.

How does the k-means algorithm work? Han, Kember and Pei have explained the workings in a very simple manner. The k-means algorithm defines the centroid of a cluster as the mean value of the points within the cluster.

a) First, it randomly selects k objects in D, each of which initially represent a cluster mean or center.
b) For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the Euclidian distance between the object and the cluster mean.
c) K-means algorithm thereafter improves the intra-cluster variation iteratively.
d) For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration.
e) All objects are then reassigned using updated means as the new cluster centers.

f) Iterations continue until the assignment is stable i.e. clusters formed in the current step are equal to the clusters formed in previous step.

Check Your Progress 1

1) Explain the concept and intuition of Cluster analysis. What are the steps involved in carrying out cluster analysis?

……

……

……

……

……

2) Which are the broad methods used in cluster-analysis? Explain the k-means method clearly indicating the steps involved in it.

……

……

……

……

……

15.6 HIERARCHICAL CLUSTERING METHODS

In some situations, you may like to partition data into groups at different levels as in a hierarchy. A Hierarchical clustering method works by grouping data objects into a hierarchy or 'tree' of clusters. By doing so it becomes easy to summarize data. This will enable you to visually interpret it. For example, the employees of a college may be divided on the basis of their nature of appointment: Permanent or Temporary and one may further partition them into smaller sub-groups like Professor, Associate Professor, Reader, Assistant Professor, etc. All these groups form a hierarchy. You can easily characterize or summarize the data in a hierarchy and can use it to find the average working experience or salaries of Professors and of Assistant Professors.

You may use hierarchical clustering methods to analyze the hierarchical structure in an organization or dataset or may also use to discover an underlying hierarchical structure. In the study of evolution, hierarchical clustering has been used to group animals according to their biological features which has enabled scientists to discover evolutionary paths, which are a hierarchy of species.

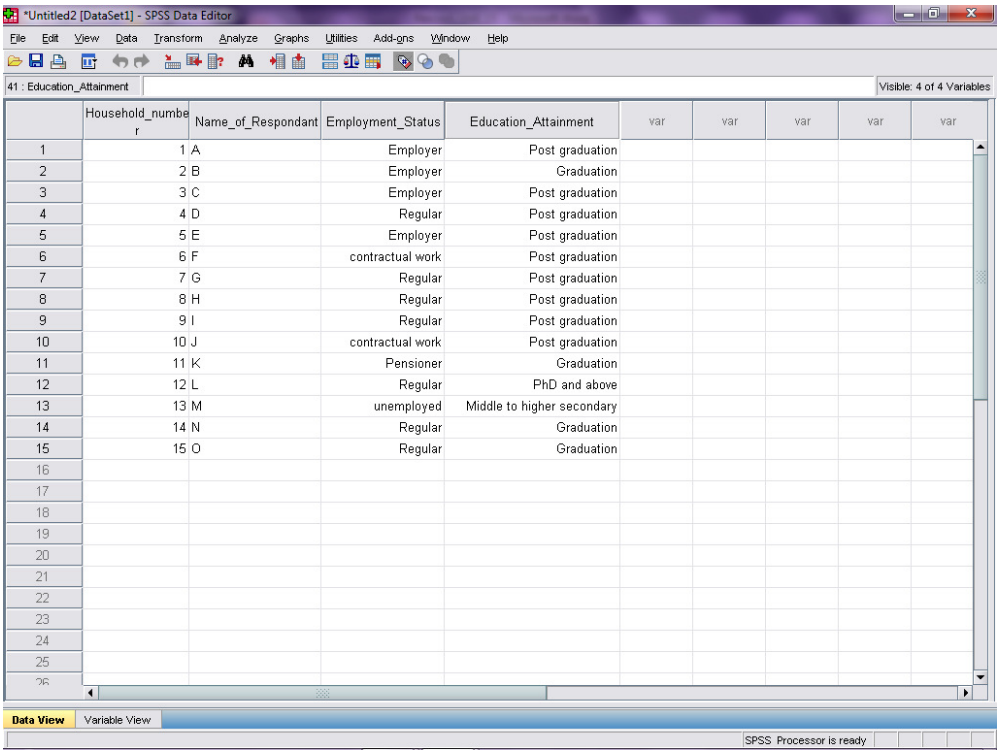
Hierarchical algorithms do not construct a single partition with k clusters, but they deal with all values of k in the interval [1, n]. Hierarchical clustering methods are of two types: Agglomerative and Divisive. Agglomerative hierarchical methods start with each unit in its own cluster and systematically merges units and clusters until all units form a single cluster. Thus, there are initially as many clusters as objects. The most similar objects are first grouped, and these initial groups are merged according to their similarities. Eventually, as the similarity decreases, all subgroups are fused into a single cluster. An agglomerative hierarchical method uses a bottom-up strategy. **Han, Kamber**

and **Pei** explain that it typically starts by letting each object form its own cluster and iteratively merges clusters into larger and larger clusters, until all objects are in a single cluster. The single cluster becomes the hierarchy's root. For merging, it finds two clusters closest to each other and combines to form one cluster. It requires at most n iterations as two clusters are merged per iteration and each cluster contains atleast one object.

Divisive hierarchical methods work in the reverse direction. Initially all objects belong to the same cluster; then, iteratively, a cluster is chosen according to a selection criterion and bi-partitioned such that the objects in one subgroup are "far from" the objects in the other. These subgroups are then further divided into dissimilar subgroups; the process continues until there are as many clusters as objects that is, until each object forms a group. Thus it employs a top-down approach. It starts by placing all objects in one cluster, which is the hierarchy's root. It then divides the root cluster into several smaller sub-clusters and repeatedly partitions them into smaller clusters.

In other words, hierarchical clustering techniques proceed by either a series of successive mergers or a series of successive divisions. In either agglomerative or divisive hierarchical clustering, a user can specify the desired number of clusters as a termination condition. The results of both agglomerative and divisive methods may be displayed in the form of a two- dimensional diagram known as dendrogram. However, this graph is generally useful for relatively small data sets only. In case of large data sets or sample size, individual objects cannot be identified. As we shall see, the dendrogram illustrates the mergers or divisions that have been made at successive levels.

Illustration: Suppose we take a sample of size 15 and wish to examine the inter-relationship of employment status and educational attainment as shown in Figure 15.1. Cluster Analysis helps in clustering objects which are similar in attributes.



	Household_number	Name_of_Respondant	Employment_Status	Education_Attainment
1	1	A	Employer	Post graduation
2	2	B	Employer	Graduation
3	3	C	Employer	Post graduation
4	4	D	Regular	Post graduation
5	5	E	Employer	Post graduation
6	6	F	contractual work	Post graduation
7	7	G	Regular	Post graduation
8	8	H	Regular	Post graduation
9	9	I	Regular	Post graduation
10	10	J	contractual work	Post graduation
11	11	K	Pensioner	Graduation
12	12	L	Regular	PhD and above
13	13	M	unemployed	Middle to higher secondary
14	14	N	Regular	Graduation
15	15	O	Regular	Graduation

Fig. 15.1: Sample data for Cluster Analysis

Source: Assessment of Human wellbeing: A Multidimensional Approach un published study undertaken by prof. Narayan Prasad

In Figure15.2, a dendrogram is constructed using SPSS which clusters sample data. A dendrogram is a branching diagram that represents the relationships of similarity among a group of entities. Each branch is called a clade. The terminal end of each clade is called a leaf. Clades can have just one leaf (these are called simplicifolious, a term from botany that means "single leafed") or they can have more than one. Two-leaved clades are bifolious, three-leaved are trifolious, and so on. There is no limit to the number of leaves in a clade. The arrangement of the clades tells us which leaves are most similar to each other. The width of the branch points indicates how similar or different they are from each other: the greater the width, the greater the difference. In this particular dendrogram, we see that chunk 6, 10 and 13 are completely separate from all the others as they are clustered through a separate leaf of the clade. We interpret its placement as indicating that the distribution in that chunk is substantially different from the distribution in the remaining chunks. By closer inspection, it can be observed that these 3 individuals are post-graduates as well as employed on either contractual basis or unemployed i.e. it reflects that higher educational attainment has not improved their job status as well as job security.

If we are trying to identify which individual segments are most similar to each other, we read the dendrogram from the left to right, identifying the first clades to join together as we move from left to right. By closer inspection we find 4 clusters which have most similar objects. The first cluster is of 2 and 15 indicating better job status after graduation, the second cluster is of 1 and 3 (both are employers and post-graduates), the 3rd cluster is of 4,8,9 and 12 which has linked post-graduates and the regularity of their employment while the 4th cluster is made up of 6 and 10 indicating contractual work despite being post-graduate. The further the clades away from the observations/objects, less is the similarity between them.

Table 15.1: Similarity between level of education and level of employment

Proximity Matrix

Observation	Squared Euclidean Distance														
	1:A	2:B	3:C	4:D	5:E	6:F	7:G	8:H	9:I	10:J	11:K	12:L	13:M	14:N	15:O
1:A	.000	1.000	.000	1.000	.000	16.000	1.000	1.000	1.000	16.000	5.000	2.000	29.000	2.000	2.000
2:B	1.000	.000	1.000	2.000	1.000	17.000	2.000	2.000	2.000	17.000	4.000	5.000	26.000	1.000	1.000
3:C	.000	1.000	.000	1.000	.000	16.000	1.000	1.000	1.000	16.000	5.000	2.000	29.000	2.000	2.000
4:D	1.000	2.000	1.000	.000	1.000	9.000	.000	.000	.000	9.000	2.000	1.000	20.000	1.000	1.000
5:E	.000	1.000	.000	1.000	.000	16.000	1.000	1.000	1.000	16.000	5.000	2.000	29.000	2.000	2.000
6:F	16.000	17.000	16.000	9.000	16.000	.000	9.000	9.000	9.000	.000	5.000	10.000	5.000	10.000	10.000
7:G	1.000	2.000	1.000	.000	1.000	9.000	.000	.000	.000	9.000	2.000	1.000	20.000	1.000	1.000
8:H	1.000	2.000	1.000	.000	1.000	9.000	.000	.000	.000	9.000	2.000	1.000	20.000	1.000	1.000
9:I	1.000	2.000	1.000	.000	1.000	9.000	.000	.000	.000	9.000	2.000	1.000	20.000	1.000	1.000
10:J	16.000	17.000	16.000	9.000	16.000	.000	9.000	9.000	9.000	.000	5.000	10.000	5.000	10.000	10.000
11:K	5.000	4.000	5.000	2.000	5.000	5.000	2.000	2.000	2.000	5.000	.000	5.000	10.000	1.000	1.000
12:L	2.000	5.000	2.000	1.000	2.000	10.000	1.000	1.000	1.000	10.000	5.000	.000	25.000	4.000	4.000
13:M	29.000	26.000	29.000	20.000	29.000	5.000	20.000	20.000	20.000	5.000	10.000	25.000	.000	17.000	17.000
14:N	2.000	1.000	2.000	1.000	2.000	10.000	1.000	1.000	1.000	10.000	1.000	4.000	17.000	.000	.000
15:O	2.000	1.000	2.000	1.000	2.000	10.000	1.000	1.000	1.000	10.000	1.000	4.000	17.000	.000	.000

Source: Assessment of Human Wellbeing: A Multidimensional Approach, a research study (unpublished) undertaken in 2015 by Prof. Narayan Prasad.

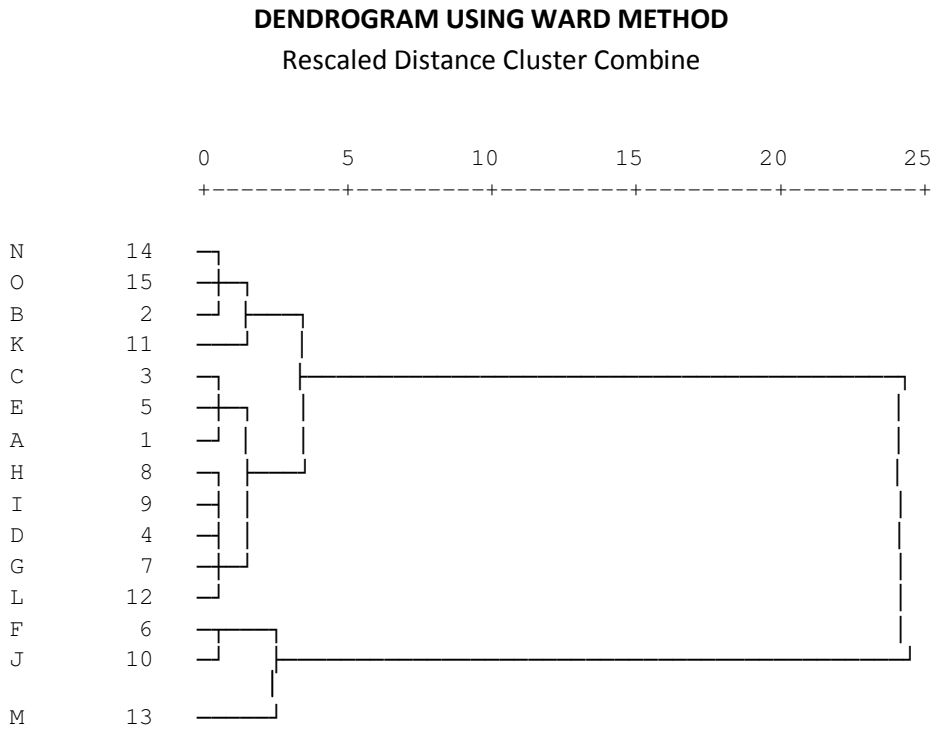


Fig.15.2: Dendrogram

15.7 OTHER APPROACHES: TWO-STEP CLUSTER ANALYSIS

Among the other widely used non-hierarchical approaches is the approach of Two-step cluster analysis. The Two Step Cluster Analysis procedure is an exploratory tool designed to reveal naturalgroupings (or clusters) within a data set that would otherwise difficult to find. The Algorithm employed by this procedure has several desirable features that differentiate it from traditional clustering techniques:

- The ability to create clusters based on both categorical and continuous variables.
- Automatic selection of the number of clusters.
- The ability to analyze large data files efficiently.

In order to handle categorical and continuous variables, the TwoStep Cluster Analysis procedure uses a likelihood distance measure which assumes that variables in the cluster model are independent. Further, each continuous variable is assumed to have a normal (Gaussian) distribution and each categorical variable is assumed to have a multinomial distribution.

The two steps of the Two Step Cluster Analysis procedure's algorithm can be summarized as follows:

Step 1 The procedure begins with the construction of a Cluster Features (CF) Tree. The tree begins by placing the first case at the root of the tree in a leaf node that contains variable information about that case. Each successive case is then added to an existing node or forms a new node, based upon its similarity to existing nodes and using the distance measure as the similarity criterion. A

node that contains multiple cases contains a summary of variable information about those cases. Thus, the CF tree provides a capsule summary of the data file.

Step 2 The leaf nodes of the CF tree are then grouped using an agglomerative clustering algorithm. The agglomerative clustering can be used to produce a range of solutions. To determine which number of clusters is "best", each of these cluster solutions is compared using Schwarz's Bayesian Criterion (BIC) or the Akaike Information Criterion (AIC) as the clustering criterion.

15.8 INTERPRETATION OF THE RESULTS

We shall now discuss this method with the help of an example based on 416 observations collected through primary survey for the purpose of research study entitled "Assessment of Human Wellbeing in Delhi :Multidimensional Approach". The researcher in this study attempted to find out the clusters on the basis of household level of income and their educational status. Four income levels have been used to classify the sample into four categories: income up to 2.5 lakhs, between 2.5 and 5 lakhs, between 5-10 lakhs and above 10 lakhs. The educational attainment has been take as : illiterate (no education), upto middle education level, between middle to higher secondary education, upto graduation, upto post-graduation and PhD and above. The hypothesis was that higher the level of education, higher will the level of income meaning thereby a positive association between level of education and the level of income. Let us examine this hypothesis by applying the two-step cluster analysis technique.

Figure 15.3 shows results of cluster 1 wherein educational attainment of post-graduation has been found to be inter-linked with income category of 10 lakhs and above. This is logically also true as higher levels of educational attainment are associated with higher income levels. In **Figure 15.4,** it can be observed that graduation is associated with income-levels of 5-10 lakhs and these two sub-groups have been clustered together.

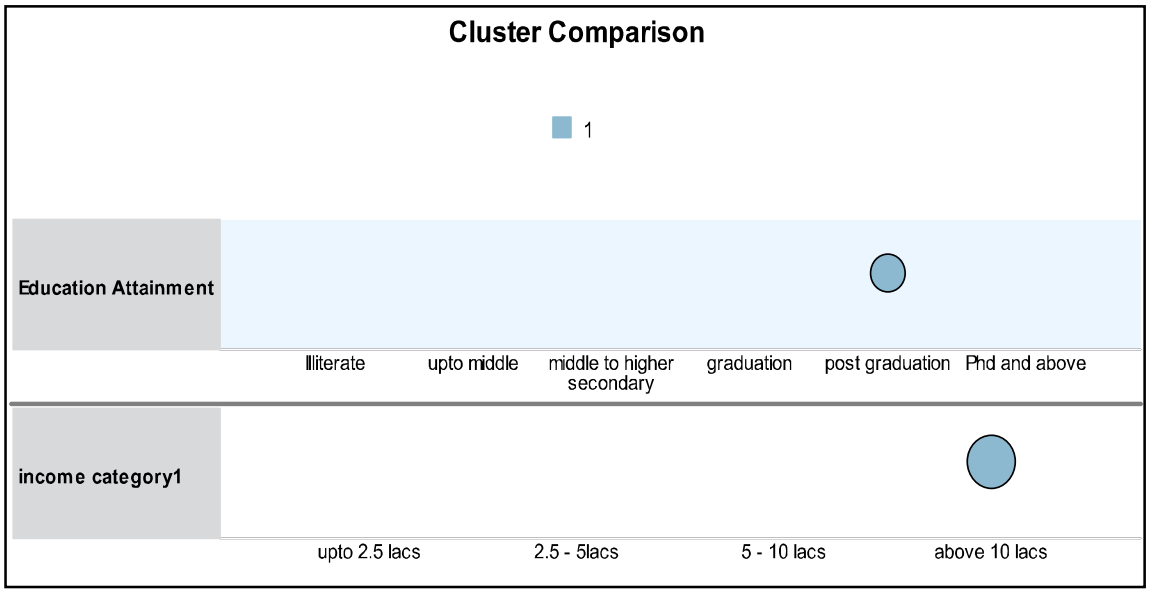


Fig.15. 3: Cluster 1

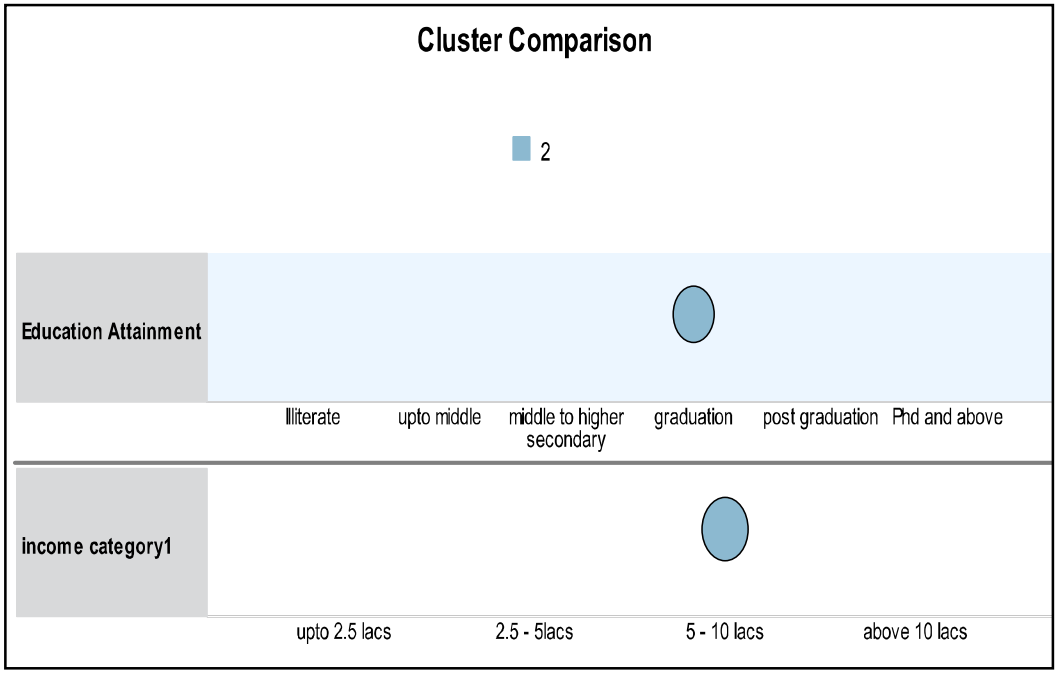


Fig.15.4: Cluster 2

Figure 15.5 finds another cluster linking illiterate individuals with the lowest income category (i.e. less than 2.5 lakhs) while **Figure 15.6** finds the fourth cluster consisting of middle to higher secondary education and income level upto 2.5 lakhs. Thus intuitively also, two-step cluster analysis has formed clusters of similar attributes.

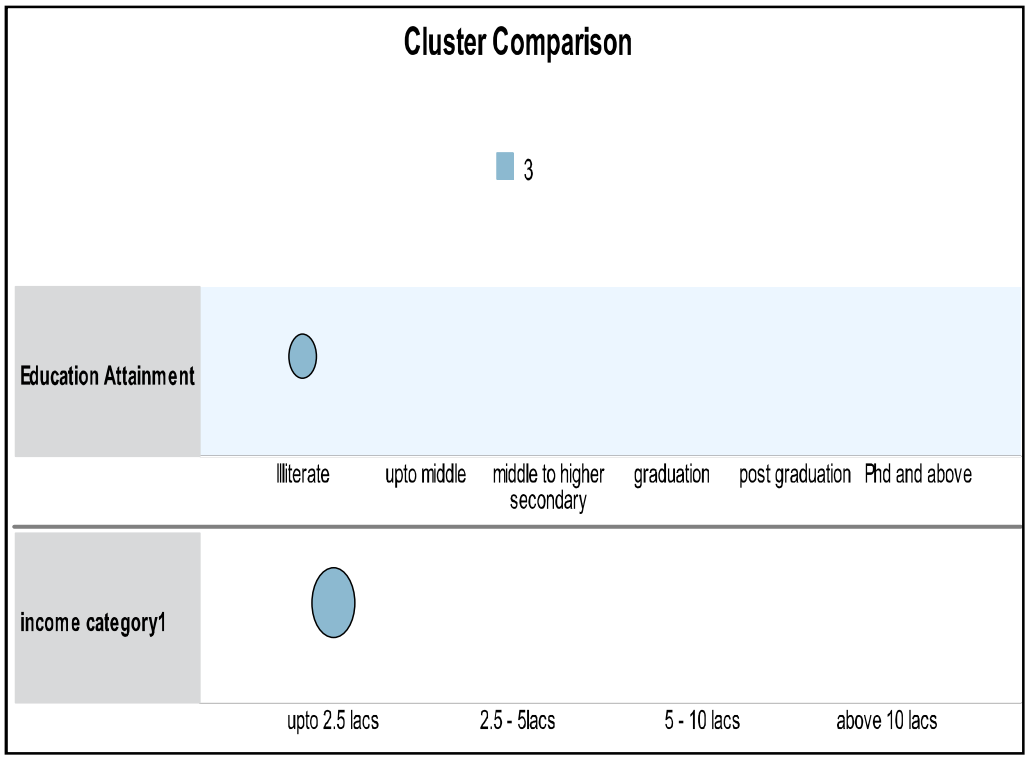


Fig.15.5: Cluster 3

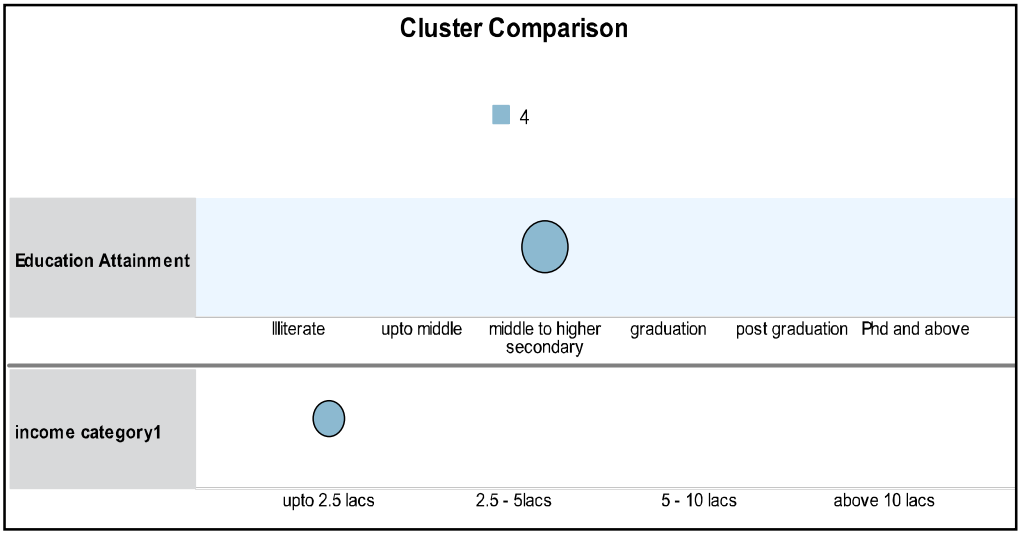


Fig.15.6: Cluster 4

Thus, by using Two step cluster analysis technique , we have separated 416 individuals into 4 broad clusters based on their income category and level of educational attainment which, in turn , has enabled us to test the hypothesis whether an association exist between the level of education and level of income. Thus the hypothesis that higher the level of education, higher the level of income has been accepted by the results of the cluster analysis.

Check Your Progress 2

1) What are the hierarchical methods used in cluster analysis? Critically explain how dendrograms help in visual interpretation of clustered data.

   ……

   ……

   ……

   ……

   ……

   ……

2) Explain the application of two-step cluster analysis in research. What are its advantages?

   ……

   ……

   ……

   ……

   ……

   ……

15.9 LET US SUM UP

The objective of cluster analysis is to assign observations to groups (clusters)so that observations within each group are similar to one another with respect to variables or attributes of interest and the groups themselves stand apart from

one another. Cluster analysis embraces a variety of methods aiming to group observations or variables into homogeneous and distinct clusters. For groupings based on variables, frequently used measures of the similarity of observations are the Euclidean. Squared Euclidean are applied to the original, standardized, or weighted variables. For groupings based on attributes, a measure of the similarity of two observations is the ratio of the number of matches (identical categories) to the number of attributes.

Hierarchical methods begin with as many clusters as there are observations and end with a single cluster containing all observations. All clusters formed by these methods are mergers of previously formed clusters. Other types of clustering methods are the hierarchical divisive (beginning with a single cluster and ending with as many clusters as there are observations) and the non-hierarchical methods. In recent years, some other clustering methods have also evolved like the two-step cluster analysis discussed above which are efficient in handling large data sets of continuous and categorical variables as well. Cluster analysis technique gives a different angle of viewing and interpreting data which may help the researcher to discover significant new relationships between the variables used in a study.

15.10 KEY WORDS

Cluster analysis : It is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Agglomerative method : In which subjects start in their own separate cluster. The two 'closest' (most similar) clusters are then combined and this is done repeatedly until all subjects are in one cluster. At the end, the optimum number of clusters is then chosen out of all cluster solutions.

Divisive method : In which all subjects start in the same cluster and the above strategy is applied in reverse until every subject is in a separate cluster. Agglomerative methods are used more often than divisive methods, so this handout will concentrate on the former rather than the latter.

Dendrogram : When carrying out a hierarchical cluster analysis, the process can be represented on a diagram known as a dendrogram. This diagram illustrates which clusters have been joined at each stage of the analysis and the distance between clusters at the time of joining.

Centroid method : Here the centroid (mean value for each variable) of each cluster is calculated and the distance between centroids is used. Clusters whose centroids are closest together are merged. This method is also fairly robust.

Ward's Method : In this method all possible pairs of clusters are combined and the sum of the squareddistances within each cluster is calculated. This is then summed over all clusters. The combination that gives the lowest sum of squares is chosen. This method tends to produce clusters of approximately equal size, which is not always desirable. It is also quite sensitive to outliers. Despite this, it is one of the most popular methods, along with the average linkage method.

K-Means Clustering : The *k*-means algorithm was popularized and refined by Hartigan (1975). The basic operation of that algorithm is relatively simple. Given a fixed number of (desired or hypothesized) *k* clusters, assign observations to those clusters so that the means across clusters (for all variables) are as different from each other as possible.

Euclidean Distance : The straight line distance between two points in a Cartesian coordinate system. The Euclidean distance can be determined using the Pythagorean Theorem. In two dimensions, the Euclidean distance is $[(x_1-x_2)^2 + (y_1-y_2)^2]^{0.5}$. Usually, the points represent samples and the axes of the Cartesian coordinate system represent the abundances of species.

Exploratory Analysis : A general term for an analysis in which the chief objectives is to find pattern in the data. Often, exploratory analysis conflicts with hypothesis testing. For example, stepwise regression is permissible in exploratory analysis, but can cause serious problems if you are interested in testing hypotheses.

15.11 SOME USEFUL BOOKS

1) Data Mining, Concepts and Techniques: Jiawei Han, Micheline Kamber, Jian Pei, 3rd Edition, Morgan Kaufmann, 2012.

2) Cluster Analysis of Economic Data, Hana Rezankova, Statistika, 94 (1), Czech Republic, 2014

3) http://wheatoncollege.edu/lexomics/files/2012/08/How-to-Read-a-Dendrogram-Web-Ready.pdf

4) https://en.wikipedia.org/wiki/Cluster_Analysis

5) https://arifkamarbafadal.files.wordpress.com/2011/09/ebook-038-tutorial-spss-two-step-cluster-analysis.pdf

15.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 15.2 and 15.3

2) See Section 15.4 and 15.5

Check Your Progress 2

1) See Section 15.6

2) See Section 15.7

15.13 EXERCISES

1) For what purposes and situations would you choose to undertake cluster analysis?

2) Explain in your own words briefly what hierarchical cluster analysis does. Check your answer against the material above.

3) Explain the difference between hierarchical methods and non-hierarchical methods of clustering citing their advantages and disadvantages.

4) Is clustering same as classification? What is the essential difference between classification and clustering?

UNIT 16 CORRESPONDENCE ANALYSIS

Structure

16.0 Objectives
16.1 Introduction
16.2 Correspondence Analysis: Concept and Its Features
16.3 Steps and Algorithm Involved in Correspondence Analysis Technique
16.4 Basic Concepts and Definitions
16.4.1 Primitive Matrix
16.4.2 Profiles
16.4.3 Masses
16.4.4 Correspondence Matrix
16.4.5 Augmented Correspondence Matrix
16.4.6 Inertia
16.4.7 Distances
16.5 Reduction of Dimensionality
16.6 Biplots
16.7 Interpretation of the Results of Correspondence Analysis
16.8 Multiple Correspondence Analysis
16.9 Let Us Sum Up
16.10 Key Words
16.11 Some Useful Books
16.12 Answers or Hints to Check Your Progress
16.13 Exercises

16.0 OBJECTIVES

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

- state the concept of Correspondence Analysis (CA) as a special case of principal component analysis;
- discuss the special features and applications of CA technique;
- explain the computational algorithm of Correspondence Analysis;
- elucidate the process of interpretation of Correspondence Analysis results; and
- learn how to apply Correspondence Analysis technique in conducting research in social sciences.

16.1 INTRODUCTION

Understanding data is the ultimate problem of statistics and no one technique/method can satisfactorily answer all our curiosities regarding any particular set of data. Given a set of data, we often formulate hypothesis regarding it. This formulation is motivated by our information regarding the variables of the data. We then apply various statistical tools to test our hypothesis.

In such traditional methods of hypothesis testing, we verify a priori hypothesis, that is, the hypothesis pre-assumed regarding the relations between variables involved in the data. However, these methods naturally suffer from a limitation. The pre-assumed hypotheses may not be the only possibility. There can be other hidden relationships among the variables of the data. How do we find them?

You have studied in Unit 1 and 8 that from measurement scale point of view, data can be of four types – nominal, ordinal, interval and ratio scale. Principal Component Analysis (PCA) technique is used to identify the relative importance of factors or components responsible for variance in the dependent variable. PCA, however, pre-supposes tables consisting of continuous measurement (i.e. interval or ordinal scale of data). If a situation arises when one is interested to know association between two or more categorized variables, search for an alternative analytic technique is needed. For example, you may be interested to examine how the level of education is related to the nature of employment (i.e. regular, casual, contract etc.). Similarly how the quality of employment is associated with different levels of income? In such situation, you need a data analytic technique which can help you in detecting structural relationships among the categorized variables. Correspondence Analysis (CA) is analytic technique which helps to examine such structural relationship. This technique is very useful to analyze simple two way and multi-way tables containing some measure of correspondence between the rows and columns. Hence, in this unit, we shall throw light on various concepts used in correspondence analysis, its pre-requisites, its computation algorithm, interpretation of its results, etc. Let us begin with stating the meaning of correspondence analysis.

16.2 CORRESPONDENCE ANALYSIS: CONCEPT AND ITS FEATURES

Correspondence Analysis (CA) is a descriptive exploratory technique of multivariate statistical analysis, allowing to define the nature and structure of the relationship between qualitative variables measured in nominal and ordinal scales. CA is a way of seeing the association in a two way – cross tabulation rather than measuring it. In general, CA as a method is used to decomposes the overall chi-square statistics by defining a system with a smaller number of dimensions, in which the deviation from the expected values are presented. In brief, CA may be defined as a special case of Principal Component Analysis of the rows and columns of a table especially applicable to a cross tabulation. The primary goal of CA is to transform a table of numerical information into a graphical display, in which each row and each column is depicted as a point.

The important features of correspondence analysis are:

1) CA provides the multivariate treatment of the data through simultaneous consideration of multiple categorical variables.
2) This technique can reveal relationships that would not be detected in a series of pair wise comparison of variable.
3) It also provides graphical display of row and column points in biplots to dectect structural relationships among the variable categories and objects.

4) Except rectangular data matrix with non-negative entries, its data requirement is highly flexible.

16.3 STEPS AND ALGORITHM INVOLVED IN CORRESPONDENCE ANALYSIS TECHNIQUE

CA technique involves the following steps and algorithm:

1) Construction of the primitive matrix.
2) Calculation of the row profile and the column profile.
3) Computation of the correspondence matrix.
4) Obtain total row mass & total column mass.
5) Construction of the augmented correspondence matrix. Sometimes for the sake of simplicity, it is also named as correspondence matrix.
6) Obtain the chi-square distance & inertia.
7) Reduction of the dimension of the data.
8) Obtain the biplots.

16.4 BASIC CONCEPTS AND DEFINITIONS

Let us understand the basic concepts and definitions of correspondence analysis with the help of an illustration. We are taking up a worked out example to understand some of the points involved in the correspondence analysis technique.

Let us consider 2 qualitative variables: Educational Status and Employment Status. Correspondence analysis technique will help us to examine the association between these two qualitative variables. For the purpose of our analysis, we have divided employment status into six categories: Employer, regular, pensioner/earning income from assets, temporary, contractual/own account worker and unemployed. Similarly, educational attainment of the persons have been categorized as illiterate, upto medium education, between medium and higher education, upto graduation, uptopost graduation or any technical degree and Ph.D and above.

Before examining the association between these two categorical variables, let us first understand the basic terms and definitions involved in the correspondence analysis technique.

16.4.1 Primitive Matrix

Primitive Matrix is the information which has been collected and classified in the tabular form. It is a cross tabulation of row and column variables. Table no. 16.1 offers a good example of Primitive Matrix. The information about employment status and educational attainment with their sub categories in rows and columns respectively has been provided in this table.

Table 16.1: Correspondence Between Educational Attainment and Employment Status

Primitive Matrix ($A = a_{ij}$)

	Employment_Status						
Educational Attainment	unemployed	Contractual	Temporary	Pensioner	Regular	Employer	Active Margin
Illiterate	22	16	1	1	0	1	41
Upto Middle	27	7	0	1	7	2	44
Middle to Higher Secondary	69	8	3	2	4	12	98
Graduation	51	2	4	4	22	20	103
Post grad/Technical degree	13	11	7	0	46	14	91
PhD & Above	1	3	4	0	28	3	39
Active Margin	183	47	19	8	107	52	416

Source: Assessment of Human Wellbeing: A Multidimensional Approach, a research study (unpublished) undertaken in 2015 by Prof. Narayan Prasad.

16.4.2 Profiles

Profiles display the characteristics of a particular category which is determined by their respective frequencies. As shown in the table no. 16.2a, row profile describes the educational status while the column profile displays the employment status.

Row profiles give the frequency of each row point such that summation for row will be equal to 1. Row profile are calculated by taking individual row point divided by their corresponding row total (active margin). That is,

$$\text{row profile for the } i^{th} \text{ row} =: \frac{\text{every entry of the } i^{th} \text{ row of A}}{\text{sum of the entries of the } i^{th} \text{ row of A}}$$

$$= \frac{a_{ij}}{\sum_{j=1} a_{ij}}$$

Column profiles are calculated in similar manner.

$$\text{Column profile for the } i^{th} \text{column} =: \frac{\text{every entry of the } i^{th} \text{ column of A}}{\text{sum of the entries of the } i^{th} \text{ column of A}}$$

$$= \frac{a_{ij}}{\sum_{i=1} a_{ij}}$$

The only difference is that in this case, each column point is divided by their corresponding column total (active margin) as shown in table 16.2b.

Table 16.2a: Correspondence Between Educational Attainment and Employment Status

Row Profiles

Educational Attainment	Employment_Status						
	unemployed	Contractual	Temporary	Pensioner	Regular	Employer	Active Margin
Illiterate	.537	.390	.024	.024	.000	.024	1.000
Upto Middle	.614	.159	.000	.023	.159	.045	1.000
Middle to Higher Secondary	.704	.082	.031	.020	.041	.122	1.000
Graduation	.495	.019	.039	.039	.214	.194	1.000
Post grad/Technical degree	.143	.121	.077	.000	.505	.154	1.000
PhD & Above	.026	.077	.103	.000	.718	.077	1.000
Mass	.440	.113	.046	.019	.257	.125	

Table 16.2b: Correspondence Between Educational Attainment and Employment Status

Column Profiles

Educational Attainment	Employment_Status						
	unemployed	Contractual	Temporary	Pensioner	Regular	Employer	Mass
Illiterate	.120	.340	.053	.125	.000	.019	.099
Upto Middle	.148	.149	.000	.125	.065	.038	.106
Middle to Higher Secondary	.377	.170	.158	.250	.037	.231	.236
Graduation	.279	.043	.211	.500	.206	.385	.248
Post grad/Technical degree	.071	.234	.368	.000	.430	.269	.219
PhD & Above	.005	.064	.211	.000	.262	.058	.094
Active Margin	1.000	1.000	1.000	1.000	1.000	1.000	

16.4.3 Masses

The mass of the i^{th} row is the marginal frequency of i^{th}row divided by grand total. Marginal frequency of the i^{th} is the sum of the entries of the i^{th} row.Mass for each row and column points are depicted in table no. 16.2a & 16.2b respectively.Grand total in this particular illustration is total sample size which is 416. Similarly mass for j^{th} column is marginal frequency of j^{th} column divided by grand total.

16.4.4 Correspondence Matrix

Correspondence matrix is defined as the original matrix divided by grand total. The correspondence matrix shows how value of mass is distributed across the cells. The row total and the column total of the correspondence matrix are the row mass and column mass respectively as shown in table 16.3.

Table 16.3: Correspondence Between Educational Attainment and Employment Status

Correspondence Matrix

Educational Attainment	Employment_Status					
	unemployed	Contractual	Temporary	Pensioner	Regular	Employer
Illiterate	0.053	0.038	0.002	0.002	0.000	0.002
Upto Middle	0.065	0.017	0.000	0.002	0.017	0.005
Middle to Higher Secondary	0.166	0.019	0.007	0.005	0.010	0.029
Graduation	0.123	0.005	0.010	0.010	0.053	0.048
Post grad/Technical degree	0.031	0.026	0.017	0.000	0.111	0.034
PhD & Above	0.002	0.007	0.010	0.000	0.067	0.007

16.4.5 Augmented Correspondence Matrix

Having computed the correspondence matrix, we now append this correspondence matrix by adding the column of row masses and the row of column masses in the table 16.4a.

Table 16.4 a: Correspondence Between Educational Attainment and Employment Status

Augmented Correspondence Matrix

Educational Attainment	Employment_Status						
	unemployed	Contractual	Temporary	Pensioner	Regular	Employer	Row Mass
Illiterate	0.053	0.038	0.002	0.002	0.000	0.002	0.099
Upto Middle	0.065	0.017	0.000	0.002	0.017	0.005	0.106
Middle to Higher Secondary	0.166	0.019	0.007	0.005	0.010	0.029	0.236
Graduation	0.123	0.005	0.010	0.010	0.053	0.048	0.248
Post grad/Technical degree	0.031	0.026	0.017	0.000	0.111	0.034	0.219
PhD & Above	0.002	0.007	0.010	0.000	0.067	0.007	0.094
Column Mass	0.440	0.113	0.046	0.019	0.257	0.125	1.000

16.4.6 Inertia

The term inertia, in the correspondence analysis explains the degree of variation. The value of inertia is high when there is large deviation of row and column profiles from their averages. The method to compute the total inertia is as follows:

$$\text{Total inertia} = \sum_i \sum_j \frac{\left(p_{ij} - r_i c_j\right)^2}{r_i c_j}$$

Where p_{ij} = (i, j)th entry of correspondence matrix

R_i = ith row profile

C_j = jth column profile

The inertia in table no. 16.3 gives the total variance explained by each dimension in the model. In this illustration value of inertia is0.480. This value of inertia indicates that in this model knowing something about education attainment explains around 48% of something about employment status and vice versa. This association is highly significant as indicated by chi square statistic.

16.4.7 Distance

Independence and association between the two variables is measured by chi square. Chi square distances are weighted Euclidean distance between the profile points. Weights here refers to weights of dimension and not to the weights of profile points. Chi square is calculated as follows:

Formula to compute the chi-square distance:

Suppose the correspondence matrix of the primitive matrix is mxn. Then, we define:

Chi-square distance between r^{th} and s^{th} rows is:=

$$d(r,s) = \sqrt{\sum_{j=1}^{n} \frac{1}{a_j}\left(\frac{a_{rj}}{a_r} - \frac{a_{sj}}{a_s}\right)^2}$$

wherea_j = marginal frequence of j^{th} column

$= \sum_{i=1}^{m} a_{ij}$

a_r = marginal frequenceofr^{th} column

$= \sum_{j=1}^{n} a_{rj}$

a_s = marginal frequence of s^{th} column

$= \sum_{j=1}^{n} a_{sj}$

16.5 REDUCTION OF DIMENSIONALITY

Another way of looking at correspondence analysis is to consider it as a method for decomposing the overall inertia by identifying a small number of dimensions in which the deviations from the expected values can be represented. This is similar to the goal of factor analysis, where the total variance is decomposed, so as to arrive at lower – dimensional representation of variables that allows one to reconstruct most of the variance/covariance matrix of variables.

Notice the tables 16.4 given below. The educational attainment column shows only 5 categories. Also notice that employment status has been condensed into two categories only – employed and unemployed. This reduction of dimensionalities is not arbitrary. It is a result of classification of sub categories in distances from same point of reference. This will further help in pictorial presentation of correspondence between the two variables – namely education attainment and employment status. Correspondence analysis aims to decompose the overall inertia so as to arrive at a lower dimensional space. Only those dimensions are included in the model which can easily be

interpreted. That is why the value of inertia does not always add up to 1 as all dimensions are not included in the model.

Table 16.4: Correspondence Between Educational Attainment and Employment Status

Summary

Dimension	Singular Value	Inertia	Chi Square	Sig.	Proportion of Inertia		Confidence Singular Value	
					Accounted for	Cumulative	Standard Deviation	Correlation 2
1	.594	.353			.735	.735	.035	.009
2	.329	.108			.226	.960	.054	
3	.113	.013			.026	.987		
4	.069	.005			.010	.996		
5	.041	.002			.004	1.000		
Total		.480	199.660	.000[a]	1.000	1.000		

a. 25 degrees of freedom

16.6 BIPLOTS

The biplot is concerned with data reconstruction in a joint map of the rows and columns, rather than distance reconstruction. In the simple case of a two-way table one can think of reconstructing different variants of the table depending on the way we think of the table: either as a set of rows, or a set of columns, or just a two-way table of entries where rows and columns are symmetric entities.

A biplot is a graphical representation of the information in an n x p data matrix. The bi- refers to the two kinds of information contained in a data matrix. The information in the rows pertains to samples or sampling units and that in the columns pertains to variables. When there are only two variables, scatter plots can represent the information on both the sampling units and the variables in a single diagram. This permits the visual inspection of the position of one sampling unit relative to another and the relative importance of each of the two variables to the position of any unit.

With several variables, one can construct a matrix array of scatter plots, but there is no one single plot of the sampling units. On the other hand, a two-dimensional plot of the sampling units can be obtained by graphing the first two principal components. The idea behind biplots is to add the information about the variables to the principal component graph.

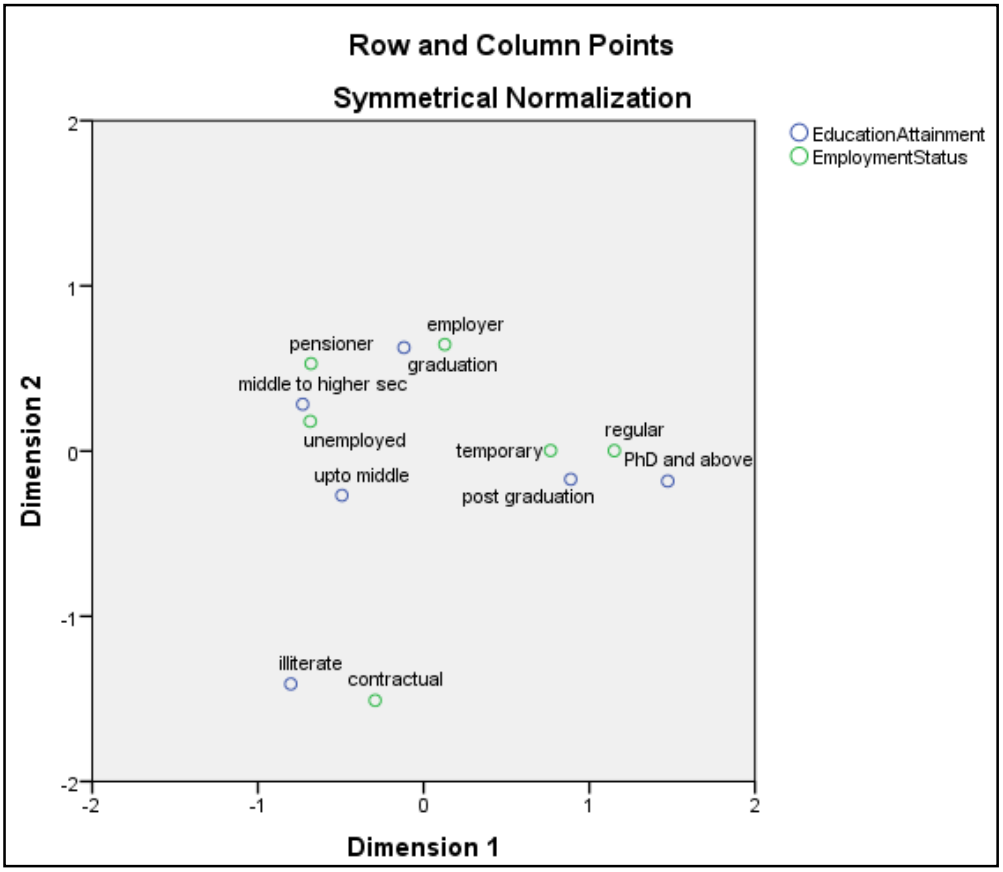


Fig.16.1: Biplots showing correspondence between Educational Attainment and Employment Status

Data has been graphically represented by the biplot in two dimensional space as shown in figure 1. However, there is one caution in this particular model that this model explains only 48% of employment status based on educational attainment. From the above biplot some general trends can be seen. Two points closer to each other shows greater degree of association between them as compared to others. This plot is offering us significant insight into direction of correspondence between the level of educational attainment and employment status. For example, we see that illiterate individuals are more likely to be employed on the contractual basis or they are mostly engaged in low level of own account work. Those individuals who have studied upto middle school are mostly found to be unemployed. Those with Ph.D and above education has higher possibilities of regular employment.The status of employment can be seen improving for individuals who have completed graduation and post-graduation as they are one who provides employment opportunities to others also.Therefore , this plot clearly reflects a greater degree of association between educational attainment and employment status which is also significant as confirmed by the value of chi-square.

16.7 INTERPRETATION OF THE RESULTS OF CORRESPONDENCE ANALYSIS

The interpretation of the results of correspondence analysis comprises the interpretation of numerical results and factor graphics yielded by CA. The former implies selection of significant axes and significant points.

We are taking up a worked out example for Correspondence Analysis to illustrate some of the points involved in the interpretation of the results of the correspondence analysis. Table 16.1 has information on 416 persons, their educational status and their employment status. The data originated from the study entitled 'Assessment of Human Wellbeing: A Multidimensional Approach' based on a sample survey conducted in Delhi in 2015. Through this technique, an attempt was made to find out correspondence between educational attainment and employment status among the respondents selected for the study. Educational status is subdivided into 6 categories, ranging from illiterate to Ph.D and above. The employment status varies from outright unemployed to contractual, temporary, regular, pensioner and also where respondents are employers rather than employees.

The table 16.4 shows summary of our calculation. The first column shows the five dimension – notice that as we are looking at distance from some dimension – the number of dimension declines from 6 to 5. The singular value and inertia are reported in column 2 to 3 respectively. Column 6 shows proportion of inertia accounted for by the first dimension which is highest at 0-735 of the total. As we move through dimension 2 onwards, inertia falls rather sharply. Notice the similarity with principal components analysis – there too, the first component explained largest variation. The column 7 shows cumulative proportion of inertia accounted for.

The table 16.5 gives an overview of row points. Column 2 shows mass & column 3 & 4 respectively show scores in dimension 1 & 2. Column 5 records inertia of respective categories in column 1. Column 6 & 7 record contribution of a point to inertia of dimension 1 & 2 respectively, while column 8 & 9 describe contribution of two dimension to inertia of the point concerned last column show total inertia.

Table 16.5: Correspondence Between Educational Attainment and Employment Status

Overview Row Points[a]

EducationAttainment	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
illiterate	.099	-1.534	.948	.179	.350	.321	.860	.137	.997
upto middle	.106	-.771	.275	.046	.095	.029	.912	.048	.960
middle to higher secondary	.236	-.458	-.503	.050	.074	.216	.654	.328	.982
graduation	.248	.108	-.511	.022	.004	.235	.085	.798	.883
post graduation	.219	.975	.295	.146	.313	.069	.941	.036	.977
PhD and above	.094	1.074	.619	.085	.163	.130	.840	.116	.956
Active Total	1.000			.528	1.000	1.000			

The table 16.6 sums up similar information for column points.

Table 16.6

Overview Column Points[a]

EmploymentStatus	Mass	Score in Dimension		Inertia	Contribution				
		1	2		Of Point to Inertia of Dimension		Of Dimension to Inertia of Point		
					1	2	1	2	Total
unemployed	.440	-.684	.180	.129	.346	.043	.948	.036	.984
contractual	.113	-.292	-1.510	.091	.016	.782	.063	.926	.989
temporary	.046	.766	.003	.019	.045	.000	.858	.000	.858
pensioner	.019	-.679	.529	.011	.015	.016	.472	.158	.630
regular	.257	1.151	.002	.204	.574	.000	.992	.000	.992
employer	.125	.128	.645	.026	.003	.158	.046	.658	.705
Active Total	1.000			.480	1.000	1.000			

a. Symmetrical normalization

The figure 16.1 presents symmetrical normalization of the row and the column points in form of a picture in two dimension plane.

We must note that in our table 16.2a as we move up the educational attainment ladder, initially unemployment proportion is rising from 53.7% for illiterates to 61.4% for middle school educated and then 70% for middle to higher secondary level of education. For graduates, the proportion of unemployed come down to 49.5%.For post graduates, it is merely 14.3% and for highest educational category, it is barely 2.6%. Even, this data alone is offering us significant insight into direction of correspondence between the levels of educational attainments and employment status. The sub-categories of employment – contractual, temporary, regular, employer are shown separately and the data further tends to support the hypothesis that higher level of education attainment leads to higher possibilities of regular employment. However, the sub-category of 'graduates' also performs credibly as employers, while Ph.D and above get more regular employment.

Our correspondence analysis tends to provide us 'concrete' evidence in support of our tentative 'conclusion' from inspection of data described above.

16.8 MULTIPLE CORRESPONDENCE ANALYSIS

In a two-way contingency table, simple correspondence analysis provides association between two categorical variables. On the other hand, multiple correspondence analysis, an extension of the simple correspondence analysis, tackles the more general problem of associations among a set of more than two categorical variables. Generalization to more than two variables is neither obvious nor well-defined. In other areas of multivariate analysis, such as regression and log linear modelling, the situation is less complicated: for example, the transition from the regression of a response variable on a single predictor variable to the case of several predictors is quite straightforward.

The main problem is that the notion of association between two categorical variables is already a complex concept and there are several ways to generalize this concept to more than two variables.

Many different approaches exist in literature to define multiple correspondence analysis. Important among them are two main approaches: first, the correlation between sets of variables, known as canonical correlation, and second, the geometric approach, which is directly linked to data visualization, and which has many similarities to Pearson-style principal component analysis. In the explanation of each approach, one can consider the case of two variables only and then describe possible generalizations to more than two variables.

Check Your Progress 1

1) What is correspondence matrix?

……

……

……

……

……

2) A sample of n= 416 people is given in the following table showing the annual income and level of education:

Table:Income level and level of educational attainment

Educational Attainment	**Annual Income**				
	Upto 2.50 Lakh	**2.50 lakh to 5.00 lakh**	**5.00 lakh to 10.00 lakh**	**10.00 lakh to 20.00 lakh**	**Above 20.00 lakh**
Illiterate	40	1	0	0	0
Up to Middle	28	3	8	3	2
Middle to higher secondary	44	18	23	11	2
Graduation	26	18	25	27	7
Post Graduation	2	6	25	31	27
Ph.D& Above	1	2	6	18	12

Compute the value of chi-square statistics and inertia from the above table. Also interpret the results.

……..

……..

……..

……..

……..

16.9 LET US SUM UP

Correspondence analysis is a data analytic technique that helps to detect structural relationships among the categorized variables. It allows to define the nature and structure of the relationship between qualitative variables measured in nominal and ordinal scales. This technique can be considered a special case of principal component analysis of the rows and columns of a table especially applicable to a cross tabulation. Determination of the correspondence matrix, row and column profiles and masses, calculation of the distances between the rows and columns, presentation of row and column profiles, determination of the average row and column profiles, reducing the dimension of spaces, and plotting the correspondence map in terms of biplots are the various steps involved in correspondence analysis technique.

Correspondence analysis facilitates the visualization of the associations between the rows and columns of a contingency table. This technique can be applied across the desciplines ranging from social sciences (including economics, psychology etc.) to health sciences, biometry etc. It is a versatile method of data analysis in the situations where exploratory or more indepth data analysis of categorical data is required.

By virtue of being primarily a graphical technique designed to represent associations in a low-dimensional space, correspondence analysis can be regarded as a scaling method.This technique is also viewed as a complement to other methods such as biplots and multidimensional scaling. Correspondence analysis also has links to principal component analysis and canonical correlation analysis.

16.10 KEY WORDS

Correspondence Analysis : It is a multivariate statistical technique by Hirschfeld and later developed by Jean-Paul Benzecri. It is conceptually similar to principal component analysis, but applies to categorical rather than continuous data. In a similar manner to principal component analysis, it provides a means of displaying or summarizing a set of data in two-dimensional graphical form.

Inertia : The inertia of the whole table is a function of the Chi-square statistic (χ^2).

If $\chi^2 = \sum_i \sum_j \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$

where $O_{ij}$is the count of row *i*and column j of the contingency table, $E_{ij}$is the expected value under the assumption of row-by-column independence, and N is the total table count, then the total inertia of the table is given by

Total Inertia = $\frac{\chi^2}{n}$

The inertia value reported is the proportion of the total inertia that is due to this profile. Another way to interpret the inertia is that it is the weighted average of

the Chi-square distances between the row profiles and their average profile.

Centroid : The (weighted) mean of a multivariate data set which can be represented by a vector. For many ordination techniques, the centroid is a vector of zeros (that is, the scores are centered and standardized). In a direct gradient analysis, a categorical variable is often best represented by a centroid in the ordination diagram.

Categorical Variable : A variable that is represented by several different types; for example: lake/river/stream, farm/pasture/ unmanaged, pitfall trap/fence trap/direct sighting. For most multivariate analyses, categorical variables must be converted to *k*-1 dummy variables (where *k* = the number of categories).

Classification : The act of putting things in groups. Most commonly in community ecology, the "things" are samples or communities. Classification can be completely subjective, or it can be objective and computer-assisted (even if arbitrary).

Correlation : A method which determines the strength of the relationship between variables, and/or a means to test whether the relationship is stronger than expected due to the null hypothesis. Usually, we are interested in the relationship between two variables, *x* and *y*. The coefficient of correlation *r* is one measure of the strength of the relationship.

Correlation Coefficient : Usually abbreviated by r, and a number which reflects the strength of the relationship between two variables. It varies between -1 (for a perfect negative relationship) to +1 (for a perfect positive relationship). If variables are standardized to have zero mean and a unit standard deviation, then *r* will also be the slope of the relationship. The value r^2 is known as the coefficient of determination; it varies between 0 and 1. The coefficient of determination is loosely interpreted as "the proportion of variance in *y* which can be explained by *x*".

Multiple Correspondence Analysis(MCA) : Multiple correspondence analysis applied to the general problem of associations among a set of more than two categorical variables. We shall see that the generalization to more than two variables is neither obvious nor well-defined. In other areas of multivariate analysis, such as regression and log linear modelling, the situation is less complicated. The main problem is that the notion of association between two categorical variables is already a complex concept and there are several ways to generalize this concept to more than two variables.

Singular Value Decomposition : The Singular Value Decomposition is the fundamental mathematical result for Correspondence Analysis, as it is for other dimension reduction techniques such as principal component analysis, canonical correlation analysis etc. This matrix decomposition expresses any rectangular

matrix as a product of three matrices of simple structure, i.e. $D_r^{-1/2} P D_c^{-1/2}$. The columns of the matrices $D_r^{-1/2}$ and $D_c^{-1/2}$ are the left and right singular vectors respectively, and the positive values of the diagonal of P, in descending order are the singular values. The SVD is related to the more well-known eigenvalue- eigenvector decomposition of a square symmetric matrix. If a generalized form of the SVD were defined, where the singular vectors are normalized with weighting by the masses, then the CA solution can also be easily obtained.

16.11 SOME USEFUL BOOKS

1) Greenacre, M. J. (1984); *Theory and applications of Correspondence analysis*, London Academic Press.
2) Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Chichester, UK: Wiley.
3) Johnson R.A., &WichernD. W. (2002); *Applied Multivariate Statistical Analysis*, Pearson Education, Inc.
4) Manly, B.F.J. (2005), *Multivariate Statistical Methods*: A Primer, Third edition, Chapman and Hall.
5) Rencher, A.C. (2002), *Methods of Multivariate Analysis*, Second edition, Wiley.
6) https://onlinecourses.science.psu.edu
7) https://en.wikipedia.org/wiki/Factor_Analysis
8) Young, F.W, and R.M Hamer (1987); Multidimensional scaling: History, Theory and Applications, Lawrence Associates Publishers.
9) http://www.unesco.org/webworld/idams/advguide/Chapt6_5.htm
10) Laura Doey and Jessica Kurta (2011); Correspondence Analysis applied to psychological research, Tutorials in Quantitative Methods for Psychology, 2011, Vol. 7(1), Page 5-14.
11) Michael S. Lewis-Beck Alan Bryman Tim Futing Liao (Ed.) (2004); The Sage Encyclopedia of Social Science Research Methods Vol. 1, Page 203-205.

16.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS

Check Your Progress 1

1) See Section 16.4.6
2) The value of chi square is 219.816 and the value of inertia 0.528.

16.13 EXERCISES

1) What is Correspondence Analysis? Why do we use Correspondence Analysis?
2) What do you understand by multidimensional scaling? How it differ from correspondence analysis.

3) A sample of 901 individuals was cross classified according to three categories of income and four categories of job satisfaction. The results are given in the following table:

Table: Income and Job Satisfaction

Income	Job Satisfaction			
	Very Dissatisfied	Somewhat Dissatisfied	Moderately Satisfied	Very Satisfied
< $ 25,000	42	62	184	207
< $ 25,000 - $ 50,000	13	28	81	113
>$ 50,000	7	18	54	92

Perform a correspondence analysis of these data. Also interpret the results.

UNIT 17 STRUCTURAL EQUATION MODELING

Structure

17.0 Objectives
17.1 Introduction
17.2 History of Structural Equation Modelling (SEM)
17.3 Why do we Conduct Structural Equation Modelling?
17.4 Assumptions of SEM
17.5 Similarities between Traditional Statistical Methods and SEM
17.6 Differences between Traditional Statistical Methods and SEM
17.7 Concepts and Terminology used in SEM
17.7.1 Path Diagrams
17.7.2 Classical SEM Notations
17.8 SEM Models Spepecification
17.8.1 Based on Reflective Indicators
17.8.2 Based on Formative Indicators
17.8.3 Based on on both Reflective and Formative Indicators
17.9 Issues in SEM Tecnique
17.10.1 Sample Size and Power
17.10.2 Missing Data
17.10.3 Multivariate Normality and Outliers
17.10 Steps in SEM
17.10.1 Initial Model Conceptualization
17.10.2 Model Estimation
17.10.3 Model Evaluation
17.10.4 Model Modification
17.10.5 Reporting the Results
17.11 An Example of SEM
17.12 Software Programs for SEM
17.12.1 LISREL
17.12.2 EQS
17.12.3 Mplus
17.12.4 Amos
17.12.5 Mx
17.12.6 Others
17.13 Advantages and Disadvantages of SEM
17.13.1 Advantages of SEM
17.13.2 Disadvantages of SEM
17.14 Let Us Sum Up
17.15 Key Words
17.16 References

17.0 OBJECTIVES

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

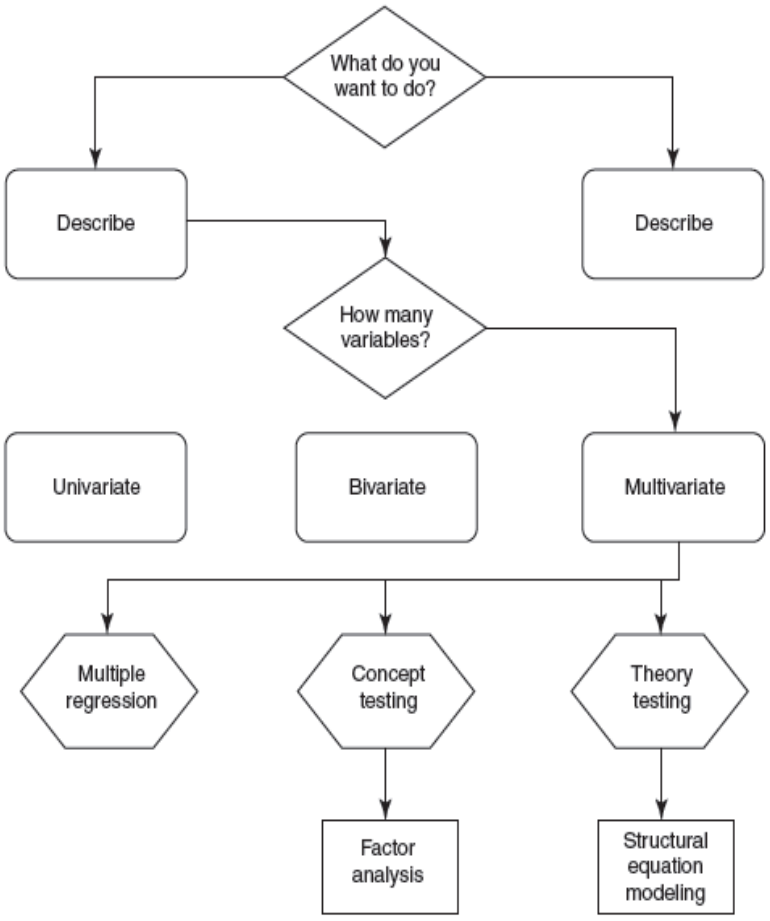
- State the meaning and importance of Structural Equation Modeling (SEM);
- Describe the basic concepts and terminology of SEM;
- Explain why do we conduct the SEM;
-

- Discuss the various models of SEM;
- Learn the steps involved in SEM;
- Know the different softwares used in SEM; and
- Discuss the advantages and disadvantages of SEM.

17.1 INTRODUCTION

SEM techniques are considered today to be a major component of applied multivariate statistical analysis and are used by education researchers, economists, marketing researchers, medical researchers, and variety of other social and behavioural scientists. Although the statistical theory that underlies the techniques appeared decades ago, a considerable number of years passed before SEM received widespread attention it hold today. There are two reasons for the recent attention: (1) the availability of specialized SEM programme and (2) publication of several introductory and advanced text books on SEM.

Structural equation modelling (SEM) does not designate a single statistical technique but instead refers to a family of related procedures. SEM is a collection of statistical techniques that allow a set of relations between one or more independent variables (IVs), either continuous or discrete, and one or more dependent variables (DVs), either continuous or discrete, to be examined. Both IVs and DVs can be either measured variables (directly observed), or latent variables (unobserved, not directly observed).



SEM is an attempt to model causal relations between variables by including all variables that are known to have some involvement in the process of interest. It can be viewed as a combination of factor analysis and regression or path

analysis. The interest in SEM is often on theoretical construct, which are represented by the latent factors. The relationship between theoretical construct are represented by regression or path coefficients between the factors.

The SEM implies a structure for the covariance between the observed variables which provides the alternative names like Covariance structural model. SEM provides a very general and convenient framework for statistical analysis that includes several traditional multivariate procedure for example, factor analysis, regression analysis, and canonical correlation as special case. Structural Equation Models are often visualized by a graphical path diagram. The statistical model is usually represented in a set of matrix equations.

Structural equation models, also called simultaneous equation models, refers to multi equation systems that include continuous **latent variables** each representing a **concept or construct**, multiple **indicators** of a concept or construct that may be continuous, ordinal, dichotomous or censored, errors of measurement and errors in equations. One may also view it as an interrelated system of regression equations where some of the variables (latent or observable) have multiple indicators and where measurement error is taken into account when estimating relationships. From a different point of view, these are factor analysis models in which factor loadings are restricted to zero or some other constants, and the researcher allows factors to influence each other, directly and indirectly. The most general form of the structural equation model includes Analysis of Variance, Analysis of Covariance, Multiple Linear Regression, Multivariate Multiple Regression, Recursive and Non-recursive Simultaneous Equations, Path Analysis, Confirmatory Factor Analysis and many other procedures as special cases.

Structural equation modeling (SEM) uses various types of models to depict relationships among observed variables, with the same basic goal of providing a quantitative test of a theoretical model hypothesized by the researcher. More specifically, various theoretical models can be tested in SEM that hypothesize how sets of variables define constructs and how these constructs are related to each other. For example, an educational researcher might hypothesize that a student's home environment influences her later achievement in school. A marketing researcher may hypothesize that consumer trust in a corporation leads to increased product sales for that corporation. A health care professional might believe that a good diet and regular exercise reduce the risk of a heart attack.

In each example, the researcher believes, based on theory and empirical research, sets of variables define the constructs that are hypothesized to be related in a certain way. The goal of SEM analysis is to determine the extent to which the theoretical model is supported by sample data. If the sample data support the theoretical model, then more complex theoretical models can be hypothesized. If the sample data do not support the theoretical model, then either the original model can be modified and tested, or other theoretical models need to be developed and tested. Consequently, SEM tests theoretical models using the scientific method of hypothesis testing to advance our understanding of the complex relationships among constructs.

SEM has been defined by different researchers in different ways.

According to **Holey (1995)** it is a comprehensive statistical approach to testing hypotheses about relations among observed and latent variables.

Rigdon, (1998): SEM is a methodology for representing, estimating, and testing a theoretical network of (mostly) linear relations between variables.

MacCallum & Austin (2000): it tests hypothesized patterns of directional and non directional relationships among a set of observed (measured) and unobserved (latent) variables.

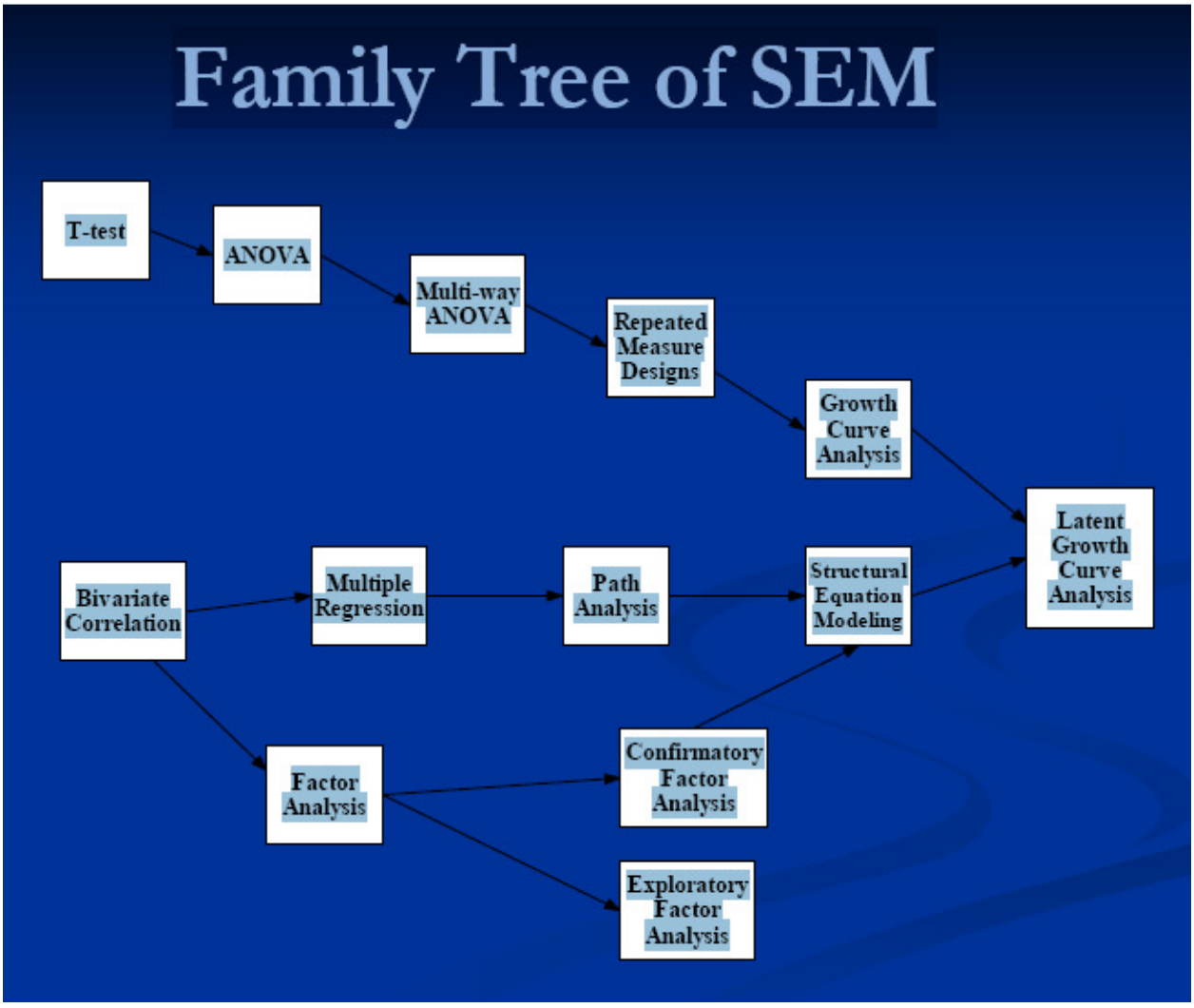
So, the term "Structural Equation Model" (SEM) refers to a comprehensive statistical methodology for testing and estimating causal relations using a combination of cross-sectional statistical data and qualitative causal assumptions. Unlike the usual multivariate linear regression model, the response variable in one regression equation in a structural equation model may appear as a predictor in another equation. Indeed, variables in a structural equation model may influence one-another reciprocally, either directly or through other variables.

Two **goals** in SEM are:

1) to understand the patterns of correlation/covariance among a set of variables and

2) to explain as much of their variance as possible with the model specified (Kline, 1998).

This is the family tree of SEM



17.2 HISTORY OF STRUCTURAL EQUATION MODELLING (SEM)

To discuss the history of structural equation modelling, it is better to explain the chronological order of following four models: **regression, path, confirmatory factor, and structural equation models.**

The first model involves linear regression models that use a correlation coefficient and the least squares criterion to compute regression weights. Regression models were made possible because Karl Pearson created a formula for the correlation coefficient in 1896 that provides an index for the relationship between two variables (Pearson, 1938). The regression model permits the prediction of dependent observed variable scores (Y scores), given a linear weighting of a set of independent observed scores (X scores) that minimizes the sum of squared residual error values. Regression analysis provides a test of a theoretical model that may be useful for prediction.

Some years later, Charles Spearman (1904, 1927) used the correlation coefficient to determine which items correlated or went together to create the factor model. His basic idea was that if a set of items correlated or went together, individual responses to the set of items could be summed to yield a score that would measure, define, or infer a construct. Spearman was the first to use the term *factor analysis* in defining a two-factor construct for a theory of intelligence. D.N. Lawley and L.L. Thurstone in 1940 further developed applications of factor models, and proposed instruments (sets of items) that yielded observed scores from which constructs could be inferred. The term *confirmatory factor analysis* (CFA) is used today based in part on earlier work by Howe (1955), Anderson and Rubin (1956), and Lawley (1958). The CFA method was more fully developed by Karl Jöreskog in the 1960s to test whether a set of items defined a construct.

Factor analysis has been used for over 100 years to create measurement instruments in many academic disciplines, while today CFA is used to test the existence of these theoretical constructs.

Sewell Wright (1918, 1921, 1934), a biologist, developed the third type of model, a path model. Path models use correlation coefficients and regression analysis to model more complex relationships among observed variables. The first applications of path analysis dealt with models of animal behavior. Unfortunately, path analysis was largely overlooked until econometricians reconsidered it in the 1950s as a form of simultaneous equation modeling (e.g., H. Wold) and sociologists rediscovered it in the 1960s (e.g., O. D. Duncan and H. M. Blalock). In many respects, path analysis involves solving a set of simultaneous regression equations that theoretically establish the relationship among the observed variables in the path model.

The final model type is structural equation modeling (SEM). SEM models essentially combine path models and confirmatory factor models, that is, SEM models incorporate both latent and observed variables. The early development of SEM models was due to Karl Jöreskog (1969, 1973), Ward Keesling (1972), and David Wiley (1973). This approach was initially known as the JKW model, but became known as the linear structural relations model (LISREL) with the development of the first software program, LISREL, in 1973. Since then, many SEM articles have been published; for example, Shumow and

Lomax (2002) tested a theoretical model of parental efficacy for adolescent students. For the overall sample, neighbourhood quality predicted parental efficacy, which predicted parental involvement and monitoring, both of which predicted academic and social-emotional adjustment.

Jöreskog and van Thillo originally developed the LISREL software program at the Educational Testing Service (ETS). The first publicly available version, LISREL III, was released in 1976. Later in 1993, LISREL8 was released; it introduced the SIMPLIS (SIMPLE LISREL) command language in which equations are written using variable names. In 1999, the first interactive version of LISREL was released. LISREL8 introduced the dialog box interface using pull-down menus and point-and-click features to develop models, and the path diagram mode, a drawing program to develop models.. The field of structural equation modelling across all disciplines has expanded since 1994.

17.3 WHY DO WE CONDUCT STRUCTURAL EQUATION MODELLING?

Why is structural equation modelling popular? There are at **least four major reasons** for the popularity of SEM.

The first reason suggests that researchers are becoming more aware of the need to use multiple observed variables to better understand their area of scientific inquiry. Basic statistical methods only utilize a limited number of variables, which are not capable of dealing with the sophisticated theories being developed. The use of a small number of variables to understand complex phenomena is limiting. In contrast, structural equation modelling permits complex phenomena to be statistically modelled and tested.

The second reason involves the greater recognition given to the validity and reliability of observed scores from measurement instruments. Specifically, measurement error has become a major issue in many disciplines, but measurement error and statistical analysis of data have been treated separately. Structural equation modelling techniques explicitly take measurement error into account when statistically analyzing data.

The third reason pertains the ability to analyze more advanced theoretical SEM models. For example, group differences in theoretical models can be assessed through multiple-group SEM models. In addition, analyzing educational data collected at more than one level—for example, school districts, schools, and teachers with student data—is now possible using multilevel SEM modelling. These advanced SEM models and techniques have provided many researchers with an increased capability to analyze sophisticated theoretical models of complex phenomena, thus requiring less reliance on basic statistical methods.

Finally, SEM software programs have become increasingly user friendly. For example, until 1993 LISREL users had to input the program syntax for their models using Greek and matrix notation. Today, most SEM software programs are Windows-based therefore, the SEM software programs are now easier to use and contain features similar to other Windows-based software packages.

17.4 ASSUMPTIONS OF SEM

Four basic assumptions must be met for SEM to be appropriate. These are:

1) The relationship between the coefficients and the error term must be linear.
2) The residuals must have a mean zero, be independent, be normally distributed, and have variances that are uniform across the variable.
3) Variables in SEM should be continuous, interval level data. This means SEM is not appropriate for censored data.
4) No specification error. If necessary variables are omitted or unnecessary variables are included in the model, there will be measurement error and the measurement model will not be accurate. Variables included in the model must have acceptable level of kurtosis.

17.5 SIMILARITIES BETWEEN TRADITIONAL STATISTICAL METHODS AND SEM

SEM is similar to traditional methods like correlation, regression and analysis of variance in many ways. First, both traditional methods and SEM are based on linear statistical models. Second, statistical tests associated with both methods are valid if certain assumptions are met. Traditional methods assume a normal distribution and SEM assumes multivariate normality. Third, neither approach offers a test of causality.

17.6 DIFFERENCES BETWEEN TRADITIONAL METHODS AND SEM

Traditional approaches differ from the SEM approach in several areas. **First**, SEM is a highly flexible and comprehensive methodology. This methodology is appropriate for investigating achievement, economic trends, health issues, family and peer dynamics, self-concept, exercise, self-efficacy, depression, psychotherapy, and other phenomenon.

Second, traditional methods specify a default model whereas SEM requires formal specification of a model to be estimated and tested. SEM offers no default model and places few limitations on what types of relations can be specified. SEM model specification requires researchers to support hypothesis with theory or research and specify relations a priori.

Third, SEM is a multivariate technique incorporating observed (measured) and unobserved variables (latent constructs) while traditional techniques analyze only measured variables. Multiple, related equations are solved simultaneously to determine parameter estimates with SEM methodology.

Fourth, SEM allows researchers to recognize the imperfect nature of their measures. SEM explicitly specifies error while traditional methods assume that measurement occurs without error.

Fifth, traditional analysis provides straightforward significance tests to determine group differences, relationships between variables, or the amount of variance explained. SEM provides no straightforward tests to determine model fit. Instead, the best strategy for evaluating model fit is to examine multiple tests (e.g., chi-square, Comparative Fit Index (CFI), Bentler-Bonett Non-normed Fit Index (NNFI), Root Mean Squared Error of Approximation (RMSEA)).

Sixth, SEM resolves problems of multicollinearity. Multiple measures are required to describe a latent construct (unobserved variable). Multicollinearity cannot occur because unobserved variables represent distinct latent constructs.

Finally, a graphical language provides a convenient and powerful way to present complex relationships in SEM. Model specification involves formulating statements about a set of variables. A diagram, a pictorial representation of a model, is transformed into a set of equations. The set of equations are solved simultaneously to test model fit and estimate parameters.

17.7 CONCEPTS AND TERMINOLOGY USED IN SEM

The key elements of essentially all structural equation models are their parameters (often referred to as model parameters or unknown parameters). Model parameters reflect those aspects of a model that are typically unknown to the researcher, at least at the beginning of the analyses, yet are of potential interest to him or her.

Parameter is a generic term referring to a characteristic of a population, such as mean or variance on a given variable, which is of relevance in a particular study.

A ***model*** is a statistical statement about the relations among variables.

A ***path diagram*** is a pictorial representation of a model.

A ***measured variable*** (MV) is a variables that is directly measured whereas a latent variable (LV) is a construct that is not directly or exactly measured.

A ***latent variable (LV)*** could be defined as whatever its multiple indicators have in common with each other. LVs defined in this way are equivalent to common factors in factor analysis and can be viewed as being free from error of measurement.

Direct effect is a directional relation between two variables, e.g., independent and dependent variables.

Indirect effect is the effect of an independent variable on a dependent variable through one or more intervening or mediating variables.

Specification is formulating a statement about a set of parameters and stating a model.

Fit indices indicate the degree to which a pattern of fixed and free parameters specified in the model is consistent with the pattern of variances and covariances from a set of observed data. Examples of fit indices are chi-square, CFI, NNFI, and RMSEA.

Components of a general structural equation model are the measurement model and the structural model.

The ***measurement model*** prescribes latent variables, e.g., confirmatory factor analysis. The ***structural model*** prescribes relations between latent variables and observed variables that are not indicators of latent variables.

Identification involves the study of conditions to obtain a single, unique solution for each and every free parameter specified in the model from the observed data.

The purpose of ***estimation*** is to obtain numerical values for the unknown (free) parameters.

Exogenous variable—Independent variables not presumed to be caused by variables in the model.

Endogenous variables— variables presumed to be caused by other variables in the model.

Recursive models assume that all causal effects are represented as unidirectional and no disturbance correlations among endogenous variables with direct effects between them.

Non-recursive models are those with feedback loops Model Specification—Formally stating a model via statements about a set of parameters.

Model Identification—Can a single unique value for each and every free parameter be obtained from the observed data.

Evaluation of Fit—Assessment of the extent to which the overall model fits or provides a reasonable estimate of the observed data.

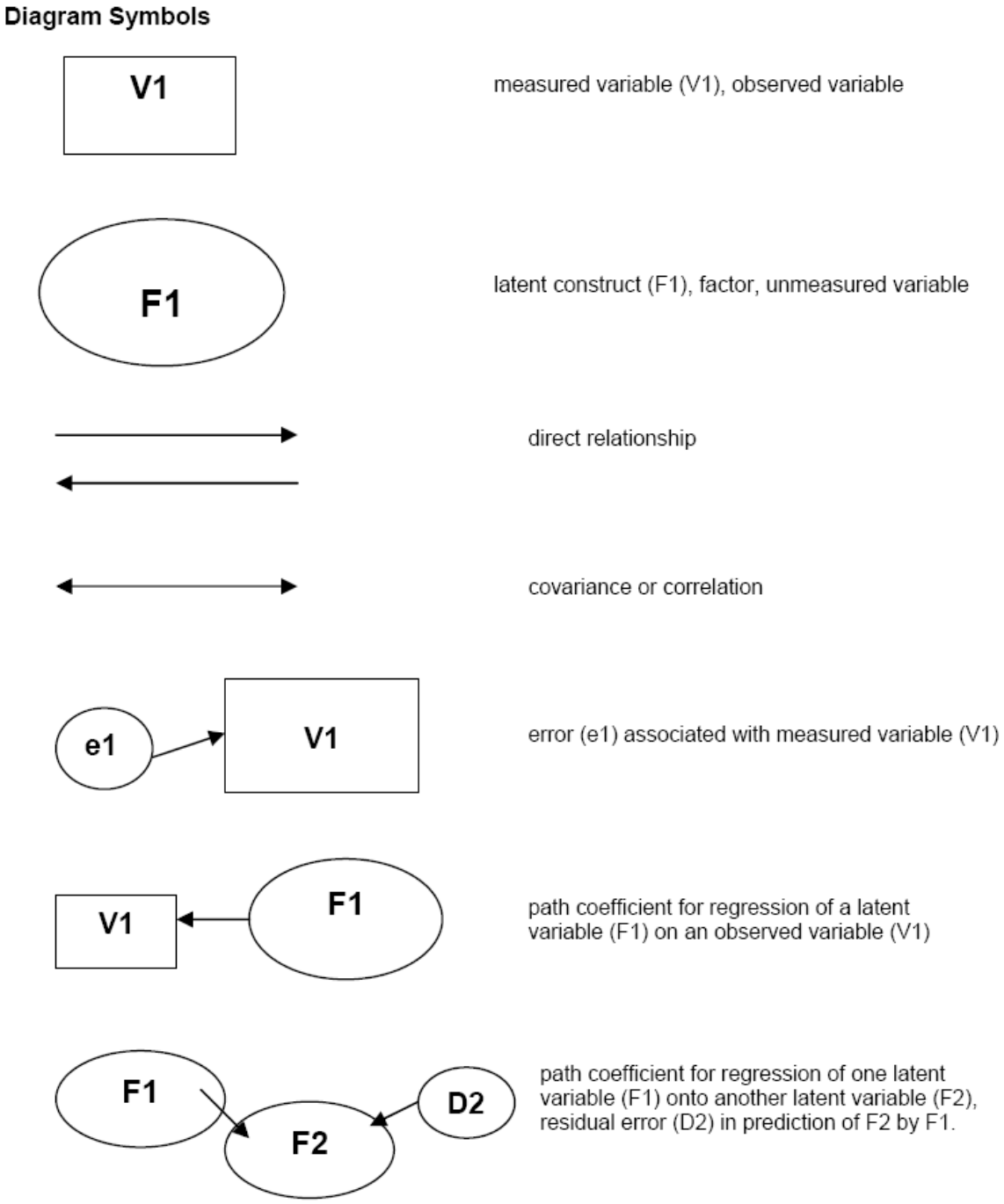
Model Modification—adjusting a specified and estimated model by freeing or fixing new parameters.

17.7.1 Path Diagrams

One of the easiest ways to communicate a structural equation model is to draw a diagram of it, referred to as *path diagram*, using special graphical notation. A path diagram is a form of graphical representation of a model under consideration. Such a diagram is equivalent to a set of equations defining a model (in addition to distributional and related assumptions), and is typically used as an alternative way of presenting a model pictorially. Path diagrams not only enhance the understanding of structural equation models and their communication among researchers with various backgrounds, but also substantially contribute to the creation of correct command files to fit and test models with specialized programs.

For constructing the path diagrams, widely accepted set of conventions for the graphical display of structural equation models can be used. According to these conventions, unobserved variables, also known as latent variables or constructs are represented as circles or ovals and measured variables also known as observed variables, indicators or manifest variables are represented by squares or rectangles. Lines indicate relations between variables; lack of a line connecting variables implies that no direct relationship has been hypothesized. Lines have either one or two arrows. A line with one arrow represents a hypothesized direct relationship between two variables. The variable with the arrow pointing to it is the DV. A line with an arrow at both ends indicates a covariance between the two variables with no implied direction of effect. Latent variables in SEM generally correspond to hypothetical constructs or factors, which are explanatory variables, supposed to reflect a range that is not directly observable. An example is the construct of intelligence. There is no single, definitive measure of intelligence. Instead, researchers use different types of observed variables, such as tasks of verbal reasoning or memory capacity, to assess various facets of intelligence. Latent variables in SEM can represent a wide range of phenomena. For example, constructs about attributes of people (e.g., intelligence, neuroticism), higher-level units of analysis (e.g.,

groups, geographic regions), or measures, such as method effects (e.g., self-report, observational), can all be represented as latent variables in SEM. An observed variable used as an indirect measure of a construct is referred to as an indicator. The explicit distinction between factors and indicators in SEM allows one to test a wide variety of hypotheses about measurement.



1.7.2 Classical SEM Notations

These are the c "classical" SEM notation that needs some consideration. ξ (ksi) represents a latent construct associated with observed xi indicators, η (eta) stands for a latent construct associated with observed yi indicators, the error terms δ (delta) and ε (epsilon) are associated with observed xi and yi indicators, respectively. ζ (zeta) is the error term associated with the formative construct. λij represents factor loading in the i-th observed indicator that is explained by

the j-th latent construct. γij represents weight in the i-th observed indicator that is explained by the j-th latent construct.

Table 17.1: Summary of abbreviations and descriptions used in the SEM study

Symbol	Name	Description
ξ	ksi	A latent construct associated with observed x_i indicators
η	eta	A latent construct associated with observed y_i indicators
δ_i	delta	The error term associated with observed x_i indicators
ε_i	epsilon	The error term associated with observed y_i indicators
ζ	zeta	The error term associated with formative construct
λ_{ij}	lambda	Factor loading in the *i-th* observed indicator that is explained by the *j-th* latent construct
γ_{ij}	gamma	Weight in the *i-th* observed indicator that is explained by the *j-th* latent construct
x_i		An indicator associated with exogenous construct, i.e. vector of observed exogenous variable
y_i		An indicator associated with endogenous construct, i.e. vector of observed endogenous variable.

17.8 SEM MODELS SPEPECIFICATION

There are three diffeerent SEM models based on different specifications.

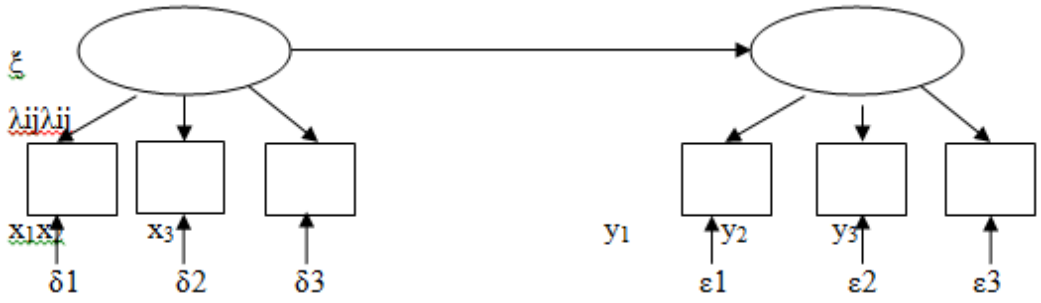
17.8.1 Based on Reflective Indicators

The following figure depicts the "classical" SEM case where the model is specified in the reflective mode known as Type A model specification. The case illustrates a path diagram between the two latent constructs (ξ – exogenous and η – endogenous), with three indicators per construct (xi and yi). This case can be represented by equations 1 and 2:

1) $xi = \lambda ij\xi + \delta i$

2) $yi = \lambda ij\eta + \varepsilon i$

This specification assumes that the error term is unrelated to the latent variable COV(η, εi) = 0, and independent COV(εi, εj) = 0, for i ≠ j and expected.



Reflective indicators

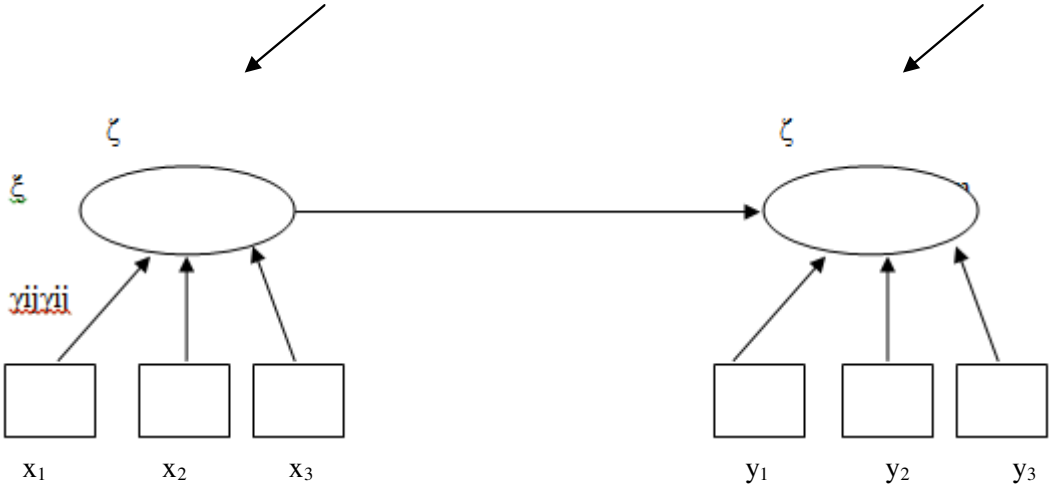
17.8.2 Based on Formative Indicators

Type B model specification, is known as a formative (Fornell&Bookstein 1982; cf. Edwards 2001) or causal indicator (Bollen and Lennox 1991),

because the direction of causality goes from the indicators (measures) to the construct and the error term is estimated at the construct level. This type of model specification can be represented by equations 3 and 4:

3) $\xi = \gamma ijxi + \zeta$

4) $\eta = \gamma ijyi + \zeta$



Formative Indicators

This specification assumes that the indicators and error term are not related, i.e. COV(yi, ζ) = 0, and E(ζ) = 0 . Formative indicators were introduced for the first time by Curtis and Jackson (1962) and extended by Blalock (1964). This type of model specification assumes that the indicators have an influence on (or that they cause) a latent construct. In other words, the indicators as a group "jointly determine the conceptual and empirical meaning of the construct.

The type B model specification would give better explanatory power, in comparison to the type A model specification, if the goal is the explanation of unobserved variance in the constructs. A model is identified if model parameters have only one group of values that create the covariance matrix. In order to resolve the problem of indeterminacy that is related to the construct-level error term, the formative-indicator construct must be associated with unrelated reflective constructs. This can be achieved if the formative construct emits paths to:

i) at least two unrelated reflective indicators;

ii) at least two unrelated reflective constructs; and

iii) one reflective indicator that is associated with a formative construct and one reflective construct.

From an empirical point of view, the latent construct captures

i) the common variance among indicators in the type A model specification; and

ii) the total variance among its indicators in the type B model specification, covering the whole conceptual domain as an entity (cf. Cenfetelli & Bassellier 2009; MacKenzie et al. 2005).

Reflective indicators are expected to be interchangeable and have a common theme. Interchangeability, in the reflective context, means that omission of an indicator will not alter the meaning of the construct. In other words, reflective measures should be unidimensional and they should represent the common theme of the construct. Formative indicators are not expected to be interchangeable, because each measure describes a different aspect of the construct's common theme and dropping an indicator will influence the essence of the latent

variable. The behavior of measures of the construct with regards to the same antecedents and consequences is an important criterion for the assessment of the construct-indicator relationship. Reflective indicators are interchangeable, which means that measures are affected by the construct, and they must have the same antecedents and consequences. The formative constructs are affected by the measures, thus are not necessarily interchangeable, and each measure can represent a different theme. For the formative indicators, it is not necessary to have the same antecedents and consequences.

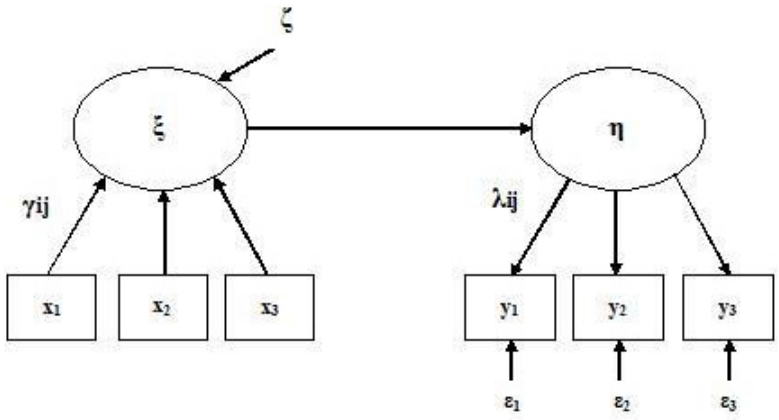
Internal consistency is implied in the reflective indicators because correlation between variables is must. High correlations among the reflective indicators are essential, because they represent the same underlying theoretical concept. This means that all of the items are measuring the same phenomenon within the latent construct. On the contrary, within the formative indicators, internal consistency is not implied because the researcher does not expect high correlations among the measures. Because formative measures are not required to be correlated, validity of construct should not be assessed by internal consistency and reliability as with the reflective measures, but with other means such as nomological and/or criterion-related validity.

17.8.3 Based on both Reflective and Formative Indicators

The researcher can create a model that uses both formative and reflective indicators. It is possible for a structural model to have one type of latent construct at the first-order (latent construct) level and a different type of latent construct at the second-order level (Fornell&Bookstein 1982; MacKenzie et al. 2005; Diamantopoulos & Siguaw 2006; Wetzels et al. 2009). In other words, the researcher can combine different latent constructs to form a hybrid model (Edwards & Bagozzi 2000; McDonald 1996; Tenenhaus et al. 2005). Development of this model type depends on the underlying causality between the constructs and indicators, as well as the nature of the theoretical concept. The researcher should model exogenous constructs in the formative mode and all endogenous constructs in the reflective mode,

i) if one intends to explain variance in the unobservable constructs (Fornell & Bookstein 1982; cf. Wetzels et al. 2009); and
ii) in case of weak theoretical background (Wold 1980).

Conducting a VBSEM approach in this model, using a PLS algorithm, is equal to redundancy analysis (Fornell et al. 1988; cf. Chin 1998b), because the mean variance in the endogenous construct is predicted by the linear outputs of the exogenous constructs.



17.9 ISSUES IN SEM TECHNIQUE

17.9.1 Sample Size and Power

SEM is based on covariance. Covariance is less stable when estimated from small samples. Therefore, large sample sizes are needed for SEM analyses. Parameter estimates and chi-square tests of fit are also very sensitive to sample size; therefore, SEM is a large sample technique. However, it may be possible to estimate small models with fewer participants if variables are highly reliable. In addition, although SEM is a large sample technique and test statistics are affected by small samples, promising work has been done by Bentler and Yuan (1999) who developed test statistics for small samples sizes.

17.9.2 Missing Data

Problems of missing data are repeatedly exaggerated in SEM due to the large number of measured variables employed. The researcher who relies on using complete cases only is often left with an inadequate number of complete cases to estimate a model. Therefore missing data attribution is particularly important in many SEM models. When there is evidence that the data are **missing at random** (MAR; missing data on a variable may depend on other variables in the dataset but the missing data does not depend on the variable itself) or **missing completely at random** (MCAR; missing data is unrelated to the variable missing data or the variables in the dataset), the **EM** (expectation maximization) algorithm to obtain **maximum likelihood** (ML) estimates, is appropriate method of imputing missing data. Some of the software packages now include procedures for estimating missing data, including the EM algorithm. EQS 6.1 (Bentler, 2004) produces the EMbased maximum likelihood solution automatically, based on the Jamshidian–Bentler (Jamshidian & Bentler, 1999) computations.

LISREL and AMOS also produce EM-based maximum likelihood estimates. but, if the data are not normally distributed, maximum likelihood test statistics— including those based on the EM algorithm—may be quite inaccurate.Multiple imputationis is another option for treatment of missing data.

17.9.2 Multivariate Normality and Outliers

The most commonly employed techniques for estimating models assume multivariate normality in SEM. It is always useful to examine both univariate and multivariate normality indexes to assess normality. Univariate distributions can be examined for outliers and skewness and kurtosis. Multivariate distributions are examined for normality and multivariate outliers. Multivariate normality can be evaluated through the use of Mardia's (1970) coefficient and multivariate outliers can be evaluated through evaluation of Mahalanobis distance. Mardia's (1970) coefficient evaluates multivariate normality through evaluation of multivariate kurtosis. Mardia's coefficient can be converted to a normalized score (equivalent to az score), often normalized coefficients greater than 3.00 are indicative of non-normality. Mahalanobis distance is the distance between a case and the centroid (the multivariate mean with that data point removed). Mahalanobis distance is distributed as a chi-square with degrees of freedom equal to the number of measured variables used to calculate the centroid. Therefore, a multivariate outlier can be defined as a case that is associated with a Mahalanobis distance greater than a critical distance specified typically by a $p < .001$.

17.10 STEPS IN SEM

Many contemporary treatments introduce SEM not just as a statistical technique but as a process involving several stages:

1) Model specification,
2) Model estimation,
3) Data-model fit assessment, and
4) Potential model modification,
5) Reporting the Results.

17.10.1 Initial Model Specification

The first step in SEM is model conceptualization or specification. This stage consists of gaining knowledge of the underlying theory on which the particular model will be based. This stage begins with stating the hypothesis in both diagrams and equation form then statistically identifying the model and then evaluating the statistical assumptions that underlie the model.

In this phase the model is specified through a diagram. The relations in the diagram are directly translated into equations and the model is then estimated. One method of model specification is the Bentler–Weeks method (Bentler& Weeks, 1980). In this method every variable in the model, latent or measured, is either an IV or a DV. The parameters to be estimated are the (a) regression coefficients, and (b) the variances and the covariances of the independent variables in the model (Bentler, 2001). In the Bentler–Weeks model only independent variables have variances and covariances as parameters of the model.

A normally tricky and often confusing topic in SEM is identification. In SEM, a model is specified, parameters (variances and covariances of IVs and regression coefficients) for the model are estimated using sample data, and the parameters are used to produce the estimated population covariance matrix. However, only models that are identified can be estimated. A model is said to be identified if there is a unique numerical solution for each of the parameters in the model. The first step consists of counting the number of data points and the number of parameters that we want to estimate. The data in SEM are the variances and covariances in the sample covariance matrix. The number of data points is the number of non redundant sample variances and covariances, where p equals the number of measured variables. The number of parameters is established by adding together the number of regression coefficients, variances, and covariances that are to be estimated. There should be more data points than parameters to be estimated for the estimation of model.

Hypothesized models are said to be over identified if there are more data than parameters to be estimated. The model is said to be just identified if, there are the same numbers of data points as parameters to be estimated. In this case, the estimated parameters perfectly reproduce the sample covariance matrix and the chi-square test statistic and degrees of freedom are equal to zero, hypotheses about the adequacy of the model cannot be tested. However, hypotheses about specific paths in the model can be tested. The model is said to be under identified if there are fewer data points than parameters to be estimated, and parameters cannot be estimated. The number of parameters needs to be reduced

by fixing, constraining, or deleting some of them. A parameter may be fixed by setting it to a specific value or constrained by setting the parameter equal to another parameter. Examining the measurement portion of the model is the second step in determining model identifiability. The measurement part of the model deals with the relationship between the measured indicators and the factors.

Factors, are hypothetical constructs and consist of common variance, as such they have no intrinsic scale and therefore need to be scaled. Measured variables have scales, for example, income may be measured in dollars or weight in pounds. To establish the scale of a factor, the variance for the factor is fixed to 1.00, or the regression coefficient from the factor to one of the measured variables is fixed to 1.00. Fixing the regression coefficient to 1 gives the factor the same variance as the common variance portion of the measured variable. If the factor is an IV, either alternative can be applied. If the factor is a DV the regression coefficient is set to 1.00. Not setting the scale of a latent variable is easily the most common error made when first identifying a model. Next, to establish the identifiability of the measurement portion of the model the number of factors and measured variables are examined. If there is only one factor, the model may be identified if the factor has at least three indicators with nonzero loading and the errors (residuals) are uncorrelated with one another. If there are two or more factors, again consider the number of indicators for each factor.

If each factor has three or more indicators, the model may be identified if errors associated with the indicators are not correlated, each indicator loads on only one factor and the factors are allowed to co-vary. If there are only two indicators for a factor, the model may be identified if there are no correlated errors, each indicator loads on only one factor, and none of the covariances among factors is equal to zero.

17.10.2 Model Estimation

The next step in SEM is to evaluate and estimate population parameters. The purpose of estimation is to minimize the difference between the structured and unstructured estimated population covariance matrices.

$$F = (\mathbf{s} - s(Q))\mathbf{W}(\mathbf{s} - s(Q)),$$

To accomplish this goal a function, F, is minimized where, **s** is the vector of data (the observed sample covariance matrix stacked into a vector); s is the vector of the estimated population covariance matrix (again, stacked into a vector) and(Q) indicates that s is derived from the parameters (the regression coefficients, variances and covariances) of the model. **W** is the matrix that weights the squared differences between the sample and estimated population covariance matrix.

In EFA the observed and reproduced correlation matrices are compared. This idea is extended in SEM to include a statistical test of the differences between the estimated structured and unstructured population covariance matrices. If the weight matrix, **W**, is chosen correctly, at the minimum with the optimal, F multiplied by (N – 1) yields a chi-square test statistic.

There are many different estimation techniques in SEM, these techniques vary by the choice of **W**.

- **Maximum likelihood** (ML) is usually the default method in most programs because it yields the most precise (smallest variance) estimates when the data are normal**.**
- **Generalized least squares** (GLS) has the same optimal properties as ML under normality.
- The **asymptotically distribution free** (ADF) method has no distributional assumptions and hence is most general, but it is impractical with many variables and inaccurate without very large sample sizes.

Satorra and Bentler (1988, 1994, and 2001) and Satorra (2000) also developed an adjustment for non-normality that can be applied to the ML, GLS, or EDT chi-square test statistics. Briefly, the Satorra–Bentler scaled $\chi 2$ is a correction to the $\chi 2$ test statistic (Satorra & Bentler, 2001). EQS also corrects the standard errors for parameter estimates to adjust for the extent of non-normality (Bentler & Dijkastra, 1985).

Sample size and plausibility of the normality and independent assumptions must be considered while choosing the suitable estimation technique. ML, the Scaled ML, or GLS estimators may be good choices with medium (over 120) to large samples and evidence of the plausibility of the normality assumptions. ML estimation is currently the most frequently used estimation method in SEM. In medium (over 120) to large samples the scaled ML test statistic is a good choice with non-normality or suspected dependence among factors and errors. In small samples (60 to 120) the Yuan–Bentler test statistic appears best. The test statistic based on the ADF estimator (without adjustment) is inappropriate choice under all conditions unless the sample size is very large (> 2,500).

17.10.3 Model Evaluation

A central issue addressed by SEM is how to assess the fit between observed data and the hypothesized model ideally operationalized as an evaluation of the degree of discrepancy between the true population covariance matrix and that implied by the model's structural and non structural parameters. As the population parameter values are rarely identified, the difference between an observed, sample-based covariance matrix and that implicit by parameter estimates must serve to approximate the population discrepancy. The observed data will fit the model perfectly for a just identified model. The system of equations expressing each model parameter as a function of the observed (co)variances is uniquely solvable; thus, the sample estimate of the model-implied covariance matrix will, by default, equal the sample estimate of the population covariance matrix. However, if a model is over identified, it is unlikely that these two matrices are equal as the system of equations (expressing model parameters as functions of observed variances and covariances) is solvable in more than a single way.

Abiding by a universal desire for parsimony, over identified models tend to be of more substantive interest than just identified ones because they characterize simpler potential explanations of the observed associations. While data-model fit for such models was initially conceived as a formal statistical test of the discrepancy between the true and model-implied covariance matrices, such a test now is often viewed as overly strict given its power to detect even trivial deviations of a proposed model from reality. Hence, many alternative

assessment strategies have emerged and continue to be developed. Data-model fit indices for such assessment can be categorized roughly into three broad classes:

- **Absolute indices** assess the overall discrepancy between observed and implied covariance matrices; fit improves as more parameters are added to the model and degrees of freedom decrease: for example, the standardized root mean square residual (SRMR), the chi-square test(recommended to be reported mostly for its historical significance), and the goodness-of-fit index (GFI).
- **Parsimonious indices** assess the overall discrepancy between observed and implied covariance matrices while taking into consideration a model's complexity; fit improves as more parameters are added to the model, as long as those parameters are making a useful contribution: for example, the root mean square error of approximation(RMSEA) with its associated confidence interval, the Akaike information criterion (AIC) for fit comparisons across non-nested models, and the adjusted goodness-of-fit index (AGFI).
- **Incremental indices** assess absolute or parsimonious fit relative to a baseline model, usually the null model (a model that specifies no relations among measured variables): for example, the comparative fit index (CFI), the normed fit index (NFI), and the non-normed fit index (NNFI).

SEM software programs routinely report a handful of goodness-of-fit indices. Some of these indices work better than others under certain conditions. It is generally recommended that multiple indices be considered simultaneously when overall model fit is evaluated. For instance, Hu and Bentler (1999) proposed a 2-index strategy that is, reporting SRMR along with one of the fit indices (e.g., RNI, CFI, or RMSEA). The authors also suggested the following criteria for an indication of good model-data fit using those indices: RNI (or CFI) $\geq$ .95, SRMR $\leq$ .08, and RMSEA $\leq$ .06. Despite the sample size sensitivity problem with the chi-square test, reporting the model chi-square value with its degrees of freedom in addition to the other fit indices is recommended. If, after considering several indices, data model fit is deemed acceptable (and judged best compared to competing models, if applicable),the model is reserved as rational, and individual parameters may be interpreted. If, however, evidence suggests unacceptable data-model fit, the next and often final stage in the SEM process is considered: modifying the model to improve fit in hopes of also improving the model's correspondence to reality.

17.10.4 Model Modification

Any hypothesized model is only an approximation to reality; the question remains is one of degree of that misspecification. With regard to external specification errors—when irrelevant variables are included in the model or substantively important ones remain left out—remediation can only occur by respecifying the model based on more relevant theory. On the other hand, internal specification errors—when unimportant paths among variables were included or when important paths were omitted—can potentially be diagnosed and remedied using Wald statistics (predicted increase in chi-square if a previously estimated parameter were fixed to some known value, e.g., zero) and Lagrange multiplier statistics (also referred to as modification indices; estimated decrease in chi-square if a previously fixed parameter were to be estimated). All are asymptotically equivalent under the null hypothesis but

approach model modification differently. As these tests' recommendations are directly motivated by the data and not by theoretical considerations, any resulting respecifications must be viewed as exploratory in nature and might not lead to a model that resembles reality any more closely than the one(s) initially conceptualized.

Chi-square difference test

If models are nested, that is, models are subsets of each other, the $\chi 2$ value for the larger model is subtracted from the $\chi 2$ value for the smaller, nested model and the difference, also $\chi 2$, is evaluated with degrees of freedom equal to the difference between the degrees of freedom in the two models. When the data are normally distributed, the chi-squares can simply be subtracted. However, when the data are non normal and the Satorra–Bentler scaled chi-square is employed, an adjustment is required (Satorra, 2000; Satorra &Bentler, 2001) so that the S–B $\chi 2$ difference test is distributed as a chi-square.

Chi-square difference tests are popular methods for comparison of nested models especially when, as in this example, there are two nested a priori models of interest; however, a potential problem arises when sample sizes are small. Because of the relationship between sample size and $\chi 2$, it is hard to detect a difference between models with small sample sizes.

Lagrange Multiplier (LM) test

The LM test also compares nested models but requires estimation of only one model. The LM test asks would the model be improved if one or more of the parameters in the model that are currently fixed were estimated. Or, equivalently, what parameters should be added to the model to improve the fit of the model?

There are many approaches to using the LM tests in model modifications. It is possible and often desirable to look only at certain parts of the model for possible change although it is also possible to examine the entire model for potential additions.

The LM test can be examined either univariately or multivariately. There is a danger in examining only the results of univariate LM tests because overlapping variance between parameter estimates may make several parameters appear as if their addition would significantly improve the model. All significant parameters are candidates for inclusion by the results of univariate LM tests, but the multivariate LM test identifies the single parameter that would lead to the largest drop in the model $\chi 2$ and calculates the expected change in $\chi 2$ if this parameter was added. After this variance is removed, the next parameter that accounts for the largest drop in model $\chi 2$ is assessed, similarly. After a few candidates for parameter additions are identified, it is best to add these parameters to the model and repeat the process with a new LM test, if necessary.

Wald test

While the LM test asks which parameters, if any, should be added to a model, the Wald test asks which, if any, could be deleted. Are there any parameters that are currently being estimated that could instead be fixed to zero? Or, equivalently, which parameters are not necessary in the model? The Wald test is analogous to backward deletion of variables in stepwise regression, in which one seeks a non significant change in R^2 when variables are left out. However, if the goal had been the development of a parsimonious model, the Wald test could have been examined to evaluate deletion of unnecessary parameters.

Finally, no matter how complex the model under study, it is always conceivable that other variables or causal assumptions that the researcher had failed to consider can account for the observed relationships. To enhance the theoretical value of their investigations, researchers are well advised to consider also models that are almost equivalent to the one reported. These are models that fit the analyzed data set almost as well as the reported model and that may offer interesting alternative theoretical interpretations.

Comparison of Nested Models

In addition to evaluating the overall model fit and specific parameter estimates, it is also possible to statistically compare nested models to one another. Nested models are models that are subsets of one another. When theories can be specified as nested hypotheses, each model might represent a different theory. These nested models are statistically compared, thus providing a strong test for competing models. Some of the common variance among the items may be due to wording and the general domain area (e.g., health problems, relationships problems), not solely due to the underlying constructs. We can compare our model to a model that also includes paths that account for the variance explained by the general domain or wording of the item. The model with the added paths to account for this variability would be considered the full model. To test this hypothesis, the chi-square from the model with paths added to account for domain and wording would be subtracted from the chi-square for the nested model in that does not account for common domains and wording among the items. The corresponding degrees of freedom for these two models would also be subtracted. Given nested models and normally distributed data, the difference between two chi-squares is a chi-square with degrees of freedom equal to the difference in degrees of freedom between the two models. The significance of the chi-square difference test can then be assessed in the usual manner. If the difference is significant, the fuller model that includes the extra paths is needed to explain the data. If the difference is not significant, the nested model, which is more parsimonious than the fuller model, would be accepted as the preferred model.

17.10.5 Reporting the Results

The final step is to precisely and completely describe the analysis in written reports. How authors structure the results section obviously is dictated by the particular model(s) and research questions under study. It is the researcher's responsibility to provide access to data in order to facilitate verification of the obtained results: If moment-level data were analyzed, a covariance matrix (or correlation matrix with standard deviations) should be offered in a table or appendix; if raw data were used, information on how to obtain access should be provided.

Results from both the measurement and structural phases should be presented. Judging the overall quality of a hypothesized model usually is presented early in the results section or overidentified models. Given that available data-model fit indices can lead to inconsistent conclusions, researchers should consider fit results from different classes so readers can arrive at a more complete picture regarding an model's acceptability. Also, a comparison of fit across multiple, a priori specified alternative models can help in weighing the relative merits of preferring one model over others. As illustrated, when competing models are nested, a formal chi-square difference test is available to judge if a more restrictive-but also more parsimonious-model can elucidate the observed data equally well, without a considerable loss in data-model fit (alternative models that

are not nested have traditionally been compared only descriptively—relative evaluations of values are recommended, with smaller values indicating better fit—but recent methodological developments suggest statistical approaches as well).

If post hoc model modifications are performed following unacceptable data-model fit from either the measurement or structural phase of the analysis, authors should give a thorough account of the nature and reasons (both statistical and theoretical)for the respecification(s), including summary results from Lagrange multiplier tests and revised final fit results. If data-model fit has been assessed and believed to be satisfactory, with or without respecification ,more detailed results are offered, usually in the form of individual unstandardized and standardized parameter estimates for each structural equation of interest, together with associated standard errors and/or test statistics and coefficients of determination. When latent variables are part of a model, estimates of their construct reliability should be presented, with values ideally falling above .70 or .80.

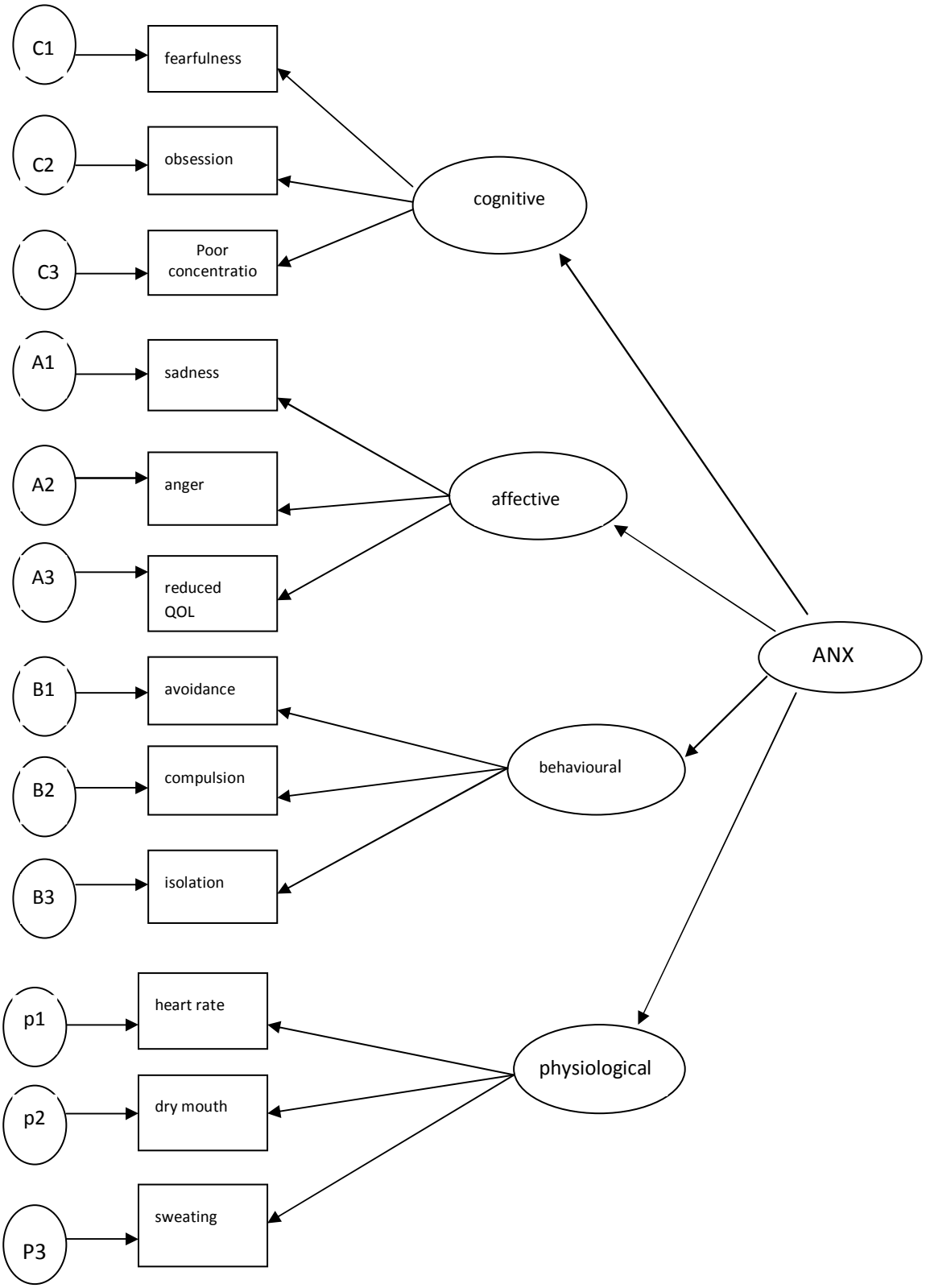
Researcher should provide a logic of what implications the results from the SEM analysis have on the theory or theories that gave rise to the initial model(s). Claims that a well-fitting model was "confirmed" or that a particular theory was proven to be "true," especially after post hoc respecifications, should be avoided. Such statements are greatly deceptive given that alternative, structurally different, yet mathematically equivalent models always exist that would produce identical data-model fit results and thus would explain the data equally well. At most, a model with acceptable fit may be interpreted as *one* reasonable explanation for the associations observed in the data. From this perspective, a SEM analysis should be evaluated from a disconfirmatory, rather than a confirmatory perspective. Based on unacceptable data model fit results, theories can be shown to be false but not proven to be true by acceptable data-model fit.

If evidence of data-model misfit was presented and a model was modified based on statistical results from Lagrange multiplier tests, readers must be made aware of potential model overfitting and the capitalization on chance. Statistically rather than theoretically based respecifications are purely exploratory and might say little about the true model underlying the data. While some model modifications seem appropriate and theoretically justifiable (usually, minor respecifications of the measurement portion are more easily defensible than those in the structural portion of a model), they only address internal specification errors and should be cross-validated with data from new and independent samples. Finally, the interpretation of individual parameter estimates can involve explicit causal language, as long as this is done from within the context of the particular causal theory proposed and the possibility/probability of alternative explanations is raised unequivocally. Though some might disagree, we think that explicit causal statements are more honest than implicit ones and are more useful in articulating a study's practical implications; after all, is not causality the ultimate aim of science.

In the end, SEM is a powerful disconfirmatory tool at the researcher's disposal for testing and interpreting theoretically derived causal hypotheses from within an a priori specified causal system of observed and/or latent variables. However, we urge authors to resist the apparently still popular belief that the main goal of SEM is to achieve satisfactory data-model fit results; rather, it is to get one step closer to the "truth." If it is true that a proposed model does not reflect reality, then reaching a conclusion of misfit between data and model should be a desirable goal, not one to be avoided by careless respecifications until satisfactory levels of fit are achieved.

17.11 AN EXAMPLE OF SEM

For example, we take some fictitious data in which anxiety was measured. We performed an confirmatory factor analysis in which we defined anxiety as consisting of four components, namely, affective, behavioral, cognitive and physiological symptoms. Thus, we specified a model which looked like as the figure below:



The next step is to estimate the parameters of the model. The software performs this step and provides final results of parameters that produce best fit to data possible.

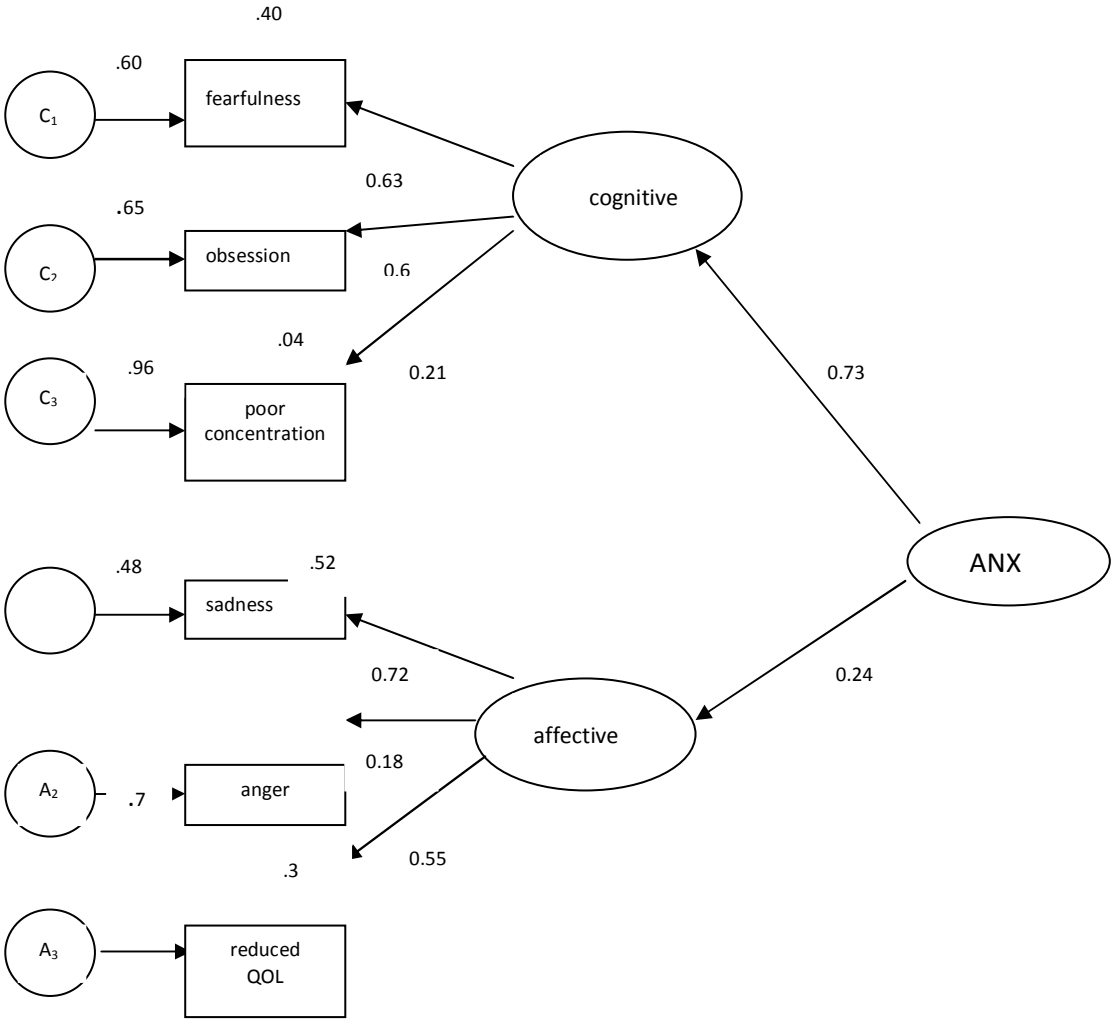


Figure 2 shows only 2 of the 4 components of anxiety. Above each arrow from a latent variable to a variable is a number, called the path coefficient. This can range from –1.0 to 1.0, with higher numbers (positive or negative) showing a stronger association. As can be seen, the variable "obsessions" doesn't fit very well with the cognitive trait; and the variable "anger" doesn't seem to go with the affective trait. The numbers over the variable names are the SMRs, which are the squared values of the path coefficients; they are interpreted in the same way as R2 multiple regression-in terms of how much of the variance in one variable is explained by, or are in common with, the other variable. Finally, the numbers over the arrows between the error terms and the variables are the variances of the errors. You'll note that the sum of the SMR plus the error variance for each variable is 1.0; that is, all the variance of a variable is divided between that shared with the latent variable and error. The cognitive domain is correlated more highly with the latent trait of anxiety than is the affective realm, as reflected in their respective path coefficients.

The next step is model specification stage. Although no mathematics is involved, it is probably the most difficult—and most important—part. It is the most difficult because it requires the most thought and understanding of the

theoretical model of the purported influences. No computer program can help us at this stage, only our knowledge of the field. It is the most important step because everything depends on how well we specify the model. The computer programs may help us in determining whether some variables aren't important, we can play with different paths to see whether they improve the model or not. However, the primary cause of poorly fitting is the omission of crucial variables.

The next step is relatively easy. The main focus is on the paths—from the latent variable to the measured variables and from that latent variable to the next one in the path. Once we have cleaned up the model by pruning noncontributory paths, we can examine the fit of the overall model. The most common index of how well the data match the model is RMSEA. it sees how much the data deviate from the model. Values over 0.10 are considered to be a bad fit, those less than 0.08 reflect a reasonable fit, and values less than 0.05 indicate a good fit.

Given is the table showing RMSEA of the model:

RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Independence model	.100	.087	.115	.000

PCLOSE = .000 for the **Independence model**. Under the hypothesis of "close fit" (i.e., that RMSEA is no greater than .05 in the population), the probability of getting a sample RMSEA as large as .100 is .000.

17.12 SOFTWARE PROGRAMS FOR SEM

Most SEM analyses are conducted using one of the specialized SEM software programs. However, there are many options, and the choice is not always easy. Below is a list of the commonly used programs for SEM. Special features of each program are briefly discussed. It is important to note that this list of programs and their associated features is by no means comprehensive. This is a rapidly changing area and new features are regularly added to the programs.

17.12.1 LISREL

LISREL (linear structural relationships) is one of the earliest SEM programs and perhaps the most frequently referenced program in SEM articles. Its version 9.1 has three components: PRELIS, SIMPLIS, and LISREL. PRELIS (pre-LISREL) is used in the data preparation stage when raw data are available. Its main functions include checking distributional assumptions, such as univariate and multivariate normality, imputing data for missing observations, and calculating summary statistics, such as Pearson covariances for continuous variables, polychoricor polyserial correlations for categorical variables, means, or asymptotic covariance matrix of variances and covariances (required for asymptotically distribution-free estimator or Satorra and Bentler's scaled chi-square and robust standard errors). PRELIS can be used as a stand-alone program or in conjunction with other programs.

Summary statistics or raw data can be read by SIMPLIS or LISREL for the estimation of SEM models. The LISREL syntax requires understanding of

matrix notation while the SIMPLIS syntax is equation-based and uses variable names defined by users. Both LISREL and SIMPLIS syntax can be built through interactive LISREL by entering information for the model construction wizards. Alternatively, syntax can be built by drawing the models on the Path Diagram screen.

LISREL 9.1 allows the analysis of multilevel models for hierarchical data in addition to the core models. A free student version of the program, which has the same features as the full version but limits the number of observed variables to12, is available from the web site of Scientific SoftwareInternational,Inc. (http://www.ssicentral.com). This web site also offers a list of illustrative examples of LISREL's basic and new features.

17.12.2 EQS

Version 6.1 of EQS (Equations) provides many general statistical functions including descriptive statistics, *t*-test, ANOVA, multiple regression, nonparametric statistical analysis, and EFA. Various data exploration plots, such as scatter plot, histogram, and matrix plot are readily available in EQS for users to gain intuitive insights into modelling problems. Similar to LISREL, EQS allows different ways of writing syntax for model specification.

The program can generate syntax through the available templates under the "Build_EQS" menu, which prompts the user to enter information regarding the model and data for analysis, or through the Diagrammer, which allows the user to draw the model. Unlike LISREL, however, data screening (information about missing pattern and distribution of observed variables) and model estimation are performed in one run in EQS when raw data are available. Model-based imputation that relies on a predictive distribution of the missing data is also available in EQS. Moreover, EQS generates a number of alternative model chi-square statistics for non normal or categorical data when raw data are available. The program can also estimate multilevel models for hierarchical data. Visit http://www.mvsoft.com for a comprehensive list of EQS's basic functions and notable features.

17.12.3 Mplus

Version 7.3 of the Mplus program includes a Base program and three add-on modules. The Mplus Base program can analyze almost all single-level models that can be estimated by other SEM programs. Unlike LISREL or EQS, Mplus version 7.33 is mostly syntax-driven and does not produce model diagrams. Users can interact with the Mplus Base program through a language generator wizard, which prompts users to enter data information and select the estimation and output options. Mplus then converts the information into its program-specific syntax. However, users have to supply the model specification in Mplus language themselves.

Mplus Base also offers a robust option for non-normal data and a special full-information maximum likelihood estimation method for missing data (see footnote 4). With the add-on modules, Mplus can analyze multilevel models and models with latent categorical variables, such as latent class and latent profile analysis. The modelling of latent categorical variables in Mplus is so far unrivalled by other programs. The official web site of Mplus (http://www.statmodel.com) offers a comprehensive list of resources including

basic features of the program, illustrative examples, online training courses, and a discussion forum for users.

17.12.4 Amos

Amos (Analysis of Moment Structure) version 22 is distributed with IBM SPSS. It has two components: Amos Graphics and Amos Basic. Similar to the LISREL Path Diagram and SIMPLIS syntax, respectively, Amos Graphics permits the specification of models by diagram drawing whereas Amos Basic allows the specification from equation statements. A notable feature of Amos is its capability for producing bootstrapped standard error estimates and confidence intervals for parameter estimates.

An alternative full-information maximum likelihood estimation method for missing data is also available in Amos. The program is available at **http://www.smallwaters.com** or **http://www.spss.com/amos/.**

17.12.5 Mx

Mx (Matrix) version 6 is a free program downloadable from http://www.vcu.edu/mx/. The Mx Graph version is for Microsoft Windows users. Users can provide model and data information through the Mx programming language. Alternatively, models can be drawn in the drawing editor of the Mx Graph version and submitted for analysis. Mx Graph can calculate confidence intervals and statistical power for parameter estimates. Like Amos and Mplus, a special form of full-information maximum likelihood estimation is available for missing data in Mx.

17.12.6 Others

In addition to SPSS, several other general statistical software packages offer built-in routines or procedures that are designed for SEM analyses. They include the CALIS (covariance analysis and linear structural equations) procedure of SAS (SAS Institute Inc., 2000; http://www.sas.com/), the RAMONA (Reticular Action Model Or Near Approximation) module of SYSTAT (Systat Software, Inc., 2002;http://www.systat.com/), and SEPATH (structural equation modeling and path analysis) of Statistica (StatSoft, Inc.,2003; http://www.statsoft.com/products/advanced.html).

17.13 ADVANTAGES AND DISADVANTAGES OF SEM

17.13.1 Advantages of SEM

1) SEM allows researcher to identify direct and indirect effects. Direct effects are principally the relationship between a dependent variable we are interested in explaining and an independent variable we think is causing or related to the dependent variable. Thus this establishes links directly from a cause to an effect. An indirect effect occurs when one variable goes through another variable on the way to some dependent or independent variable.
2) It is possible to have multiple indicators of a concept. With multiple regression analysis we only evaluate the effect of individual independent variables on the dependent variable. It creates farily a simple path model

while SEM allows for the estimation of the combined effects of independent variable into concepts/ construct.

3) SEM includes measurement error into the model. A problem for path analyses using regression as its underlying analysis not adequately control for measurement error. SEM analyses do account for measurement error, therefore providing a better understanding of how good the theoretical model predicts actual behaviour.

17.13.2 Disadvantages of SEM

1) Most of the procedures that have been suggested involve nonstandard and complex model specifications that are challenging for the average user and thus susceptible to error. Indeed, errors have even been noted in the specifications developed by SEM specialists.
2) Because products of normally distributed observed and latent variables are themselves not normally distributed, standard errors and estimates of fit might not be accurate. This problem will be more severe to the extent that the latent exogenous variables used to form the product term are highly correlated.
3) If the latent variables that denote main effects are not normally distributed, the parameter estimates yielded by several procedures are not consistent.
4) Most of the methods proposed in the literature are applicable to a restricted class of measurement models.
5) Although a number of alternative procedures and options have been proposed, the selection of an optimal approach is made difficult by the absence of any one study or set of studies that compares all the viable alternatives that have been proposed to date under a variety of conditions.
6) Some of the more promising approaches are not easily available in conventional SEM software.

17.14 LET US SUM UP

SEM techniques are considered today to be a major component of applied multivariate statistical analysis and are used by education researchers, economists, marketing researchers, medical researchers, and variety of other social and behavioural scientists. SEM is a collection of statistical techniques that allow a set of relations between one or more independent variables(IVs), and one or more dependent variables (DVs). SEM is an attempt to model causal relations between variables by including all variables that are known to have some involvement in the process of interest. It can be viewed as a combination of factor analysis and regression or path analysis. Structural equation modeling (SEM) uses various types of models to depict relationships among observed variables, with the same basic goal of providing a quantitative test of a theoretical model hypothesized by the researcher. Two goals in SEM are (1) to understand the patterns of correlation/covariance among a set of variables and (2) to explain as much of their variance as possible with the model specified.

To discuss the history of structural equation modeling, it is better to explain the chronological order of following four models : regression, path, confirmatory factor, and structural equation models.

One of the easiest ways to communicate a structural equation model is to draw a diagram of it, referred to as *path diagram*, using special graphical notation. A path diagram is a form of graphical representation of a model under consideration. Such a diagram is equivalent to a set of equations defining a model (in addition to distributional and related assumptions), and is typically used as an alternative way of presenting a model pictorially. Path diagrams not only enhance the understanding of structural equation models and their communication among researchers with various backgrounds, but also substantially contribute to the creation of correct command files to fit and test models with specialized programs. Five steps are involed in SEM: Initial model conceptualization, Model estimation, Data-model fit assessment, Potential model modification and Reporting the Results. Most SEM analyses are conducted using one of the specialized SEM software programs. These include LISREL, EQS, Mplus, Amos and Mx. The SEM has certain advantages and disadvantages also.

17.15 KEY WORDS

SEM	**:**	Collection of statistical techniques that allow a set of relations between one or more independent variables (IVs), and one or more dependent variables (DVs).
Path Diagrams	**:**	Pictorial representation of a model.
Observed variables	**:**	Variables that are directly measured.
Unobserved variable	**:**	Latent variable in terms of multiple indicators that are in common with each other.
Residual or error terms	**:**	Can be associated with either observed variables or factors specified as outcome (dependent) variables.
Measurement model	**:**	The part of the model that relates the measured variables to the factors.
ξ (ksi)	**:**	Latent construct associated with observed x_i indicators.
η (eta)	**:**	Latent construct associated with observed y_i indicators.
δ (delta)	**:**	Error term associated with observed x_i indicators.
ε (epsilon)	**:**	Error term associated with observed x_i, y_i indicators.
ζ (zeta)	**:**	Error term associated with the formative construct.
Λij	**:**	Factor loading in the i-th observed indicator that is explained by the j-th latent construct.
Γij	**:**	Weight in the i-th observed indicator that is explained by the j-th latent construct.

Reflective indicator	**:**	The construct is the cause of the observed measures, so a variation in the construct leads to a variation in all its measures.
Formative or causal indicator	**:**	The direction of causality goes from the indicators (measures) to the construct and the error term is estimated at the construct level.
Model specification	**:**	Involves using all available relevant theory, research, and information to construct the theoretical model.
Model identification	**:**	A model is said to be identified if there is a unique numerical solution for each of the parameters in the model.
Underidentified	**:**	If one or more parameters cannot solved uniquely on the basis of the covariance matrix **S** (more unknown parameters than equations).
Just-identified	**:**	There is just enough information in the covariance matrix **S** to solve the parameters of the model (equally number of equations as unknown parameters).
Over-identified	**:**	More than one way to estimate the unknown parameters (more equations than unknowns).
Model estimation	**:**	Minimize the difference between the structured and unstructured estimated population covariance matrices.
Model evaluation	**:**	Assess the fit between observed data and the hypothesized model ideally operationalized as an evaluation of the degree of discrepancy between the true population covariance matrix and that implied by the model's structural and non structural parameters.
Model modification	**:**	Respecifying the model based on more relevant theory.
Nested models	**:**	Models that are subsets of one another.

17.16 REFERENCES

Anderson, J.C., & Gerbing, D.W. (1988). *Structural Equation Modeling in Practice: A Review and Recommended Two-Step approach. Psychological Bulletin, 103(3),* 411-423.

Bentler, P.M. (1990). *Comparative fit indexes in structural models. Psychology Bulletin*, *107*, 256–259.

Bentler, P.M. (2001). *EQS 6 structural equations program manual*. Encino,CA: Multivariate Software.

Bentler, P. M., & Dijkstra, T. (1985). *Efficient estimation via linearization instructural models*. In P. R. Krishnaiah (Ed.), *Multivariate analysis 6* (pp.9–42). Amsterdam: North-Holland.

Bentler, P. M., & Raykov, T. (2000). *On measures of explained variance innonrecursive structural equation models*. *Journal of Applied Psychology*, *85*, 125–131.

Bentler, P. M., &Weeks, D. G. (1980). *Linear structural equation with latent variables*. *Psychometrika, 45,* 289–308.

Bentler, P. M., & Yuan, K.-H. (1999). *Structural equation modeling withsmall samples: Test statistics*. *Multivariate Behavioral Research, 34,*181–197.

Fox, J. (2006). *Structural equation modelling with the sem package* in R. *Structural equation modelling, 13(3),* 465-486.

Hu, L., & Bentler, P.M. (1998). *Fit indices in covariance structural equation modeling: Sensitivity to underparameterized model misspecification*. *Psychological Methods, 3,* 424–453.

Hu, L., &Bentler, P. M. (1999). *Cut off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives*. *Structural Equation Modeling, 6,* 1–55.

Kline RB. (1998). *Principles and Practice of Structural Equation Modeling*. New York: Guilford

Lei, P., & Wu, Q. (2007). *Introduction to Structural Equation Modeling: Issues and Practical Considerations*. *Instructional topics in educational measurement, fall,* 33-43.

Paxton, P., Curran, P.J., Bollen K.A., Kirby, J.,Chen, F.(2001). *Monte carlo experiments: design And implementation*. *Structural Equation Modeling*8:288–312

Satorra, A. (2000). *Scaled and adjusted restricted tests in multi-sample analysisof moment structures*. In D. D. H. Heijmans, D. S. G. Pollock, & A.Satorra (Eds.), *Innovations in multivariate statistical analysis: Afestschrift for Heinz Neudecker* (pp. 233–247). Dordrecht, The Netherlands: Kluwer Academic.

Satorra, A., & Bentler, P. M. (1988). *Scaling corrections for chi-square statistics in covariance structure analysis*. *Proceedings of the American Statistical Association,* 308–313.

Satorra, A., &Bentler, P. M. (1994). *Corrections to test statistics and standard errors in covariance structure analysis*. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Thousand Oaks, CA: Sage.

Satorra, A., &Bentler, P. M. (2001). *A scaled difference chi-square test statisticfor moment structure analysis*. *Psychometrika, 66,* 507–514.

Tomarken, A. J., & Waller, N. G. (2005). *Structural equation modeling: strengths, limitations, and misconceptions*. *Annual review of clinical psychology, 1,* 31-65.

Ullman, J. B., (2006). *Structural Equation Modeling: Reviewing the Basics and Moving Forward. Journal of personality assessment, 87(1)*, 35–50.

Valluzi, J.L., Larson, S.L., Miller, G.E. (2003). *Indications and Limitations of Structural Equation Modeling in Complex Surveys: Implications for an Application in the Medical Expenditure Panel Survey (MEPS)*. Joint Statistical Meetings - Section on Survey Research Methods.

Indira Gandhi
National Open University
School of Social Sciences

MEC-109
Research Methods in Economics

Block

5

QUALITATIVE METHODS

UNIT 18
Participatory Method **5**

UNIT 19
Content Analysis **21**

UNIT 20
Action Research **37**

Expert Committee

Prof. D.N. Reddy
Rtd. Professor of Economics
University of Hyderabad, Hyderabad

Prof. Harishankar Asthana
Professor of Psychology
Banaras Hindu University, Varanasi

Prof. Chandan Mukherjee
Professor and Dean
School of Development Studies
Ambedkar University, New Delhi

Prof. V.R. Panchmukhi
Rtd. Professor of Economics
Bombay University and Former
Chairman ICSSR, New Delhi

Prof. Achal Kumar Gaur
Professor of Economics
Faculty of Social Sciences
Banaras Hindu University, Varanasi

Prof. P.K. Chaubey
Professor, Indian Institute of Public
Administration, New Delhi

Shri S.S. Suryanarayana
Rtd. Joint Advisor
Planning Commission, New Delhi

Prof. Romar Korea
Professor of Economics
University of Mumbai
Mumbai

Dr. Manish Gupta
Sr. Economist
National Institute of Public
Finance and Policy
New Delhi

Prof. Anjila Gupta
Professor of Economics
IGNOU, New Delhi

Prof. Narayan Prasad **(Convenor)**
Professor of Economics
IGNOU, New Delhi

Prof. K. Barik
Professor of Economics
IGNOU, New Delhi

Dr. B.S. Prakash
Associate Professor in Economics
IGNOU, New Delhi

Shri Saugato Sen
Associate Professor in Economics
IGNOU, New Delhi

Course Coordinator and Editor: Prof. Narayan Prasad

Block Preparation Team

Units	Resource Person	IGNOU Faculty (Format, Language and Conent Editing)
18	Prof. Shalina Mehta Professor Department of Anthropology Panjab University Chandigarh	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi
19-20	Prof. H.S. Asthana Professor Department of Psychology Banaras Hindu University Varanasi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi

Print Production

Mr. Manjit Singh
Section Officer (Pub.)
SOSS, IGNOU, New Delhi

October, 2015

© *Indira Gandhi National Open University, 2015*

ISBN-978-81-266-

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi

Printed at :

BLOCK 5 QUALITATIVE METHODS

As explained in Block 1 of this course, the major difference between the quantitative approach and qualitative approach is not the type of data used or preferred but is much broader and deeper. The quantitative research is associated with positivist and post-positivist paradigm whereas qualitative research is associated with interpretative paradigm and critical theory paradigm.

Notwithstanding with the debate about quantitative research vs. qualitative research, a researcher may use either quantitative methods or qualitative methods or mixed methods blending the techniques of the two. In Block 3 and 4 we have covered various quantitative methods of data analysis. In this block you will find the qualitative methods exclusively associated with interpretative paradigm and critical theory paradigm. This block comprises of 3 units.

Unit 18 on **Participatory Method** throwing light on the limitations of the methods to collect quantitative data either through primary survey or secondary sources covers the various tools and techniques used in collection and analysis of qualitative data under participatory method and its advantages.

Unit 19 on **Content Analysis** deals with the conceptual and methodological issues related to content analysis as a research method to undertake research studies in social sciences in general and in economics in particular.

Unit 20 entitled **Action Research** throws light on Action Research as an offshoot to emancipatory method associated with critical theory paradigm wherein knowledge is generated in the process of knowing through doing rather through conceptualization and theorizing. This unit covers the various models of action research and the different steps involved in conducting action research.

UNIT 18 PARTICIPATORY METHOD

Structure

18.0 Objectives

18.1 Introduction

18.2 What is Participatory Research?

18.3 Methods of Participatory Research: Observation Method

- 18.3.1 Participant Observation
- 18.3.2 Non-Participant Observation
- 18.3.3 Quasi-participant Observation

18.4 Focused Interview

18.5 Oral Histories

18.6 Life History

18.7 Case Study Method

18.8 Narratives

18.9 Focus Group Discussion

18.10 Grounded Theory

- 18.10.1 Data Sources and Sampling
- 18.10.2 Subjectivity
- 18.10.3 Transcription

18.11 Analysis of Qualitative Data

18.12 Report Writing

18.13 Criticism of Participatory Methods

18.14 Advantages of Participatory Research

18.15 Let Us Sum Up

18.16 Key Words

18.17 Some Useful Books/Readings

18.18 Answers or Hints to Check Your Progress Exercises

18.0 OBJECTIVES

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

- state the concept of participatory research;
- state the distinction between data collection and data generation;
- explain the various tools of participatory research;
- discuss the grounded theory as a method of inductive theorising;
- identify the various steps involved in the analysis of qualitative data;
- pinpoint the major criticisms against participatory methods; and
- draw the lessons for participatory researchers.

18.1 INTRODUCTION

As discussed in Unit 6, traditionally two approaches – quantitative research associated with post positivist paradigm and qualitative research associated with interpretative paradigm have been followed in undertaking research in social sciences. The researchers working in the area of economics (both in micro and macro) generally rely upon the inferences drawn from the large pool of data. The data for these studies are collected either by the data compiling agencies or by the individuals hired for data collection without involvement of individual researchers in the field. The results of such quantitative research are insensitive to context and ignore the people's voices or their views about which studies have been undertaken. Instead researcher's opinion is imposed on what people say and want. The collection of quantitative data through sampling survey and its analysis involving several steps is time consuming. In the situation of natural calamities and disasters like tsunami, we cannot wait for the outcome of quantitative research. Further, quantitative research has not been very effective in evaluation of poverty alleviation schemes and programmes. In such situation, alternative method i.e. participatory method is increasingly being used in social sciences to generate and analyse the data.

Hence, in this unit, we shall take up the various issues related to participatory methods like concept of participatory research, the various tools to collect qualitative data, analysis of qualitative data, advantages, limitations and important lessons for participatory research etc. Let us begin with explanation of the concept of participatory research.

18.2 WHAT IS PARTICIPATORY RESEARCH?

The fundamental premise of participatory research involves the direct participation of researcher in the process of data generation. Important distinction is made here between data collection and data generation. When data is collected by an agency on large sample populations, standardizations are made with the assumption that all those being interviewed will understand and respond to the questions in the manner in which the primary researcher has conceptualized it. In the field situation this may or may not be the case. Each field researcher asking those questions may convey a different meaning and the respondent may give answer that may not fit into any of the standard categories, but the field researcher will reduce the answer and fit it in any of the given categories in a structured schedule. Hence, the results obtained may not necessarily reflect the market sentiments or the opinion of the people involved in the study. It is for this reason that this process of procuring data using survey method and questionnaire is called data collection.

In participatory research, the primary researcher is always in contact with the respondent and has a face-to-face interaction. If the researcher thinks that the respondent has not understood the query, he has the freedom to change the language or reconstruct his probe question or collect information from other indirect source. In this approach called **data generation**, the researcher has the flexibility to generate multiple answers to a single query and then use his/her interpretative skills to draw inference or meaning out of it to arrive at a generalization. In participatory method of data generation, the process of data

collection and analysis proceeds simultaneously, making it more reliable and presenting plurality of responses and possibilities. It is this flexibility and its ability to generate reliable generalizations that participatory methods of research have acquired importance in research methodology being used by present day economists.

Let us first try and understand what participatory research is and how and where it began? Methods of participatory research first find mention in anthropological empirical research commonly known as ethnography. Ethnography is defined as a method used to generate qualitative data from an insider's perspective known as **EMIC view**. In simple terms it implies that the people about whom we are writing, must tell their stories and the researcher simply interprets them as "thick description" (Geertz, 1973). People's voices are critical to research instead of researcher's imposition of his opinion on what people say and want.

18.3 METHODS OF PARTICIPATORY RESEARCH: OBSERVATION METHOD

Bronislaw Malinowski is regarded as the founding father of this methodology. He described this method as "participant observation". He stayed for more than three years among the "Trobriand Islanders" and described their lifeway's in a classic ethnographic work titled *Argonauts of Western Pacific* first published in 1922. Methods of participatory research in the last ninety years or more have been defined and redefined in many different ways. Malinowski's method of participant observations demands the researcher's stay in the field area for prolonged periods and become a part of the society that he/she is researching. Later researchers realized that it was always not possible to stay for extended duration (from three months to three years and in some cases even longer) in the field and some problems can require immediate and urgent interventions. Another critique of participant observation is that an interviewer is always an outsider and even when he lives at a particular field site for long periods, there is no guarantee that the community will view him as an insider and share every aspect of their private lives with the researcher.

Nonetheless, none of these limitations influenced the importance of using observations as one of the most important tools of participatory researches. Balki (2009:206) calls it as "qualitative method par excellence". Bryman (1998, 47-9) puts forward a case for method of observation by saying that it is not a single method but can combine several ways of doing observations. Given the significance of method of observation, the techniques went in for some modifications. There are three important ways of doing observations that are now recognized: (i) Participant Observations, (ii) Quasi-Participant Observations, (iii) Non-Participant Observations.

18.3.1 Participant Observations

Uncontrolled Participant observation is defined by Goode & Hatt (1981: 121) as a procedure in which "the investigator can so disguise himself as to be accepted as a member of the group". When researchers live at a field site for long duration, the assumption is that they are regarded as member of the community or the group. A researcher wanting to study consumer behaviour may not disclose his identity as a researcher and pretend to be a consumer, sharing their experiences

and concerns. While doing participant research, researchers often assume insider identity. Sometimes researchers take neutral positions without disclosing the real purpose of their being there.

Advantages of Participant Observation Method

Researcher can record all the information relevant for his work without disturbing the community and intruding upfront into their privacy. Instead of sitting in a corner in the market, an organization or any other work or field situation writing notes, the researcher participates in these activities. In doing so, he collects relevant information without disturbing the normal activities and records his observations in a work diary. Thereafter, he makes a mental note of it and returns to his/her workplace and records these observations immediately. The community or people may not be aware that they are being researched. This helps obtain unhindered and unbiased information.

Participant observation helps to obtain detailed information without any subject bias. Many researchers regard it as a better tool for data gathering vis-à-vis questionnaire and in-depth interviews.

Drawbacks of Participant Observation Method

One of the major drawbacks of participant observation is its inability to develop procedures for standardization. Researcher doing participant observation acquires a specific position in the community and observes from this vantage point. His recordings are personal or individual specific. Another observer observing the same situation may not be able to view it with the same perspective.

In participant observation method, the researcher may get emotionally involved with the issues and tends to loose his/her objectivity. This often happens on issues of dowry, bride burning, female foeticide, student agitations, price rise, public-private partnership, trade union activities etc.

To overcome some of these drawbacks, researchers proposed method of non-participant observations.

18.3.2 Non-Participant Observation

The qualitative researchers innovate in several ways to do non-participant observations. Many of them would participate in some activities and observe others from the outside. Some others would use video- recordings and analyse these recordings later for generalizations. While doing so, it was possible to take into account people's gestures and non-verbal movements for detailed scrutiny. However, critics of the method thought that as one is not fully participating in this situation, he/she is not in a position to attribute right meaning to these observations.

Sometimes, researchers use hidden cameras and record individual observations separately. But there are several problems with this method of observations too. Many regard use of hidden camera for research unethical. Participating in some activities and not in others amounted to what researchers later called as Quasi-participant observation.

18.3.3 Quasi- Participant Observation

Most researchers these days prefer to use quasi-participant method of observations. In this method, the researcher can follow the dual entry system. For instance, if you are trying to examine labour management relations, it is imperative that you obtain permission from the management and also establish rapport with the trade union leaders to have access to the union meetings. Many a times, one has to conduct interviews in different locations. In one such study in Punjab and Uttar Pradesh, we found that the workers preferred to talk to the researcher outside the factory. The researcher would talk to the mangers and supervisors in the factory and had lunch in the workers canteen. Appointments were fixed with the workers while having lunch and then researcher went to their houses, often spent hours there and after completing the interviews with the supervisors and mangers hired a small room in the same slum area, where most of the workers lived and shared their daily experiences.

In addition to the important research tool of observations, qualitative researchers often use Focused interview, in-depth interviews, narratives and oral /Life case histories, Focus group discussion/interviews', and content analysis. These tools have been discussed in next sections.

Check Your Progress 1

1) What is the distinction between data collection and data generation?

.....

.....

.....

.....

2) What do you mean by the term EMIC view?

.....

.....

.....

.....

3) How is participant observation method different from non-participant observation method?

.....

.....

.....

.....

18.4 FOCUSED INTERVIEW

In a focused interview, the researcher and the respondent are engaged in face-to-face or one to one conversation in which the researcher poses questions and the respondent provides detailed or precise answers. Talking to people or a specified group of people selected randomly, using an interview schedule, interview guide or a structured questionnaire is standard method of data collection. For qualitative

research, focused interviews are often unstructured and questions are invariably open ended. This gives the respondent liberty to provide extensive answers.

It is important to remember that interview is "fundamentally a process of social interaction" and involves developing a social relationship. In an interview situation it is not only the interviewer but also the interviewee who makes an opinion about the researcher. Respondent's responses can be determined by what he thinks what the researcher may expect from him and also the amount of trust and faith he invests in the researcher. We will repeatedly stress the importance of confidentiality and researcher's ability to convey the same as hallmarks of participatory research.

18.5 ORAL HISTORIES

Uninterrupted long interviews giving space to the respondents to express his opinion or experiences at length is the preferred method by several qualitative researchers. The interviewee is encouraged to share his life experiences, to express his opinion, to recollect and recount their antecedents or lives of their contemporaries, "and to discuss their perceptions of the processes involved and the changes they have seen"(Balki, 2010:207). Collecting oral histories or narratives requires skill of being a good listener. Many a times, you may assume that the information provided by the respondent is not directly related to the area of your focused interest. But that subsidiary information can provide certain reference for a better understanding of the problem, issues or question that you are exploring. In oral histories, one may pose probing questions in the beginning and then occasionally intervene to ask a lead question for seeking any detailed answer to specific query.

18.6 LIFE HISTORY

Another important tool that ethnographic and qualitative researchers have in their kit is called the life history method. In this method life histories of individuals are reconstructed to understand the historical events of that period. Many a times researchers would use diaries and autobiographies to construct chronology of events of that period. Recording experiences of individuals facilitates understanding of how and what happened in certain circumstances. Life history approach is largely dependent on the ability of the respondent to recall and share that memory with the researcher. With the help of the respondent, researcher gradually determines the social processes. In the context of economics, rise and fall of a particular model is determined by prevalent social processes of that period, e.g. popularity and weakening of models of conservative and liberal economics in the last fifty years. It also helps to understand different perspectives based on gender and class. While recollecting their life history, many a times respondents may fall silent. This silence also provides data as it tells the researcher, how that particular event may have caused anxiety, anguish or an intense emotional experience to the respondent.

18.7 CASE STUDY METHOD

To study firms and organization's behaviour, Business economists use the case study method. The researcher can focus on one company or organization and

collect a detailed case history of the organization. Yin (1984:23) defines case study research method "as an empirical inquiry that investigates a contemporary phenomenon within its real life context; when the boundaries between phenomenon and context are not clearly evident and in which multiple sources of evidence are used". In an organization, there are multiple players and a detailed case study will take into account perspective of each stakeholder there. You can use more than one method to collect evidence from different stakeholders e.g. the supervisors, workers etc. A schedule or a questionnaire can be used to talk to the workers. The interview method can be used, to collect history of the organization and organogram depicting position and functions of each member of the organization. Content analysis can be done by examining available records of the organization. This provides a holistic picture of the organization. Case study as an approach is **holistic**, **inductive** and **idiographic**. It takes complete account of the organization from all perspectives; theorizing moves from general to specific and individual cases are located in their particular context. Case study can take cross-sectional or longitudinal data.

The researcher must remember that generalizations drawn for one organization are not necessarily applicable to the other. Each organization is a distinct unit. Six important steps are involved in conducting a study by using case study method:

- Determine and define the research question
- Select the cases and determine data gathering and analysis techniques
- Prepare to collect the data
- Collect data in the field
- Evaluate and analyse the data
- Prepare the report

18.8 NARRATIVES

Social science research in anthropology, sociology, political science, history and now economics is moving from the positivist position of research to narrative approach. The positivist position also popularly known as scientific approach was the tradition in which researchers from 50's to 80's were generally trained. The standardized tools of survey research, questionnaire approach or even interviews with a standardized set of questions was expected to produce if not same, similar responses. Narrative research emphasizes that this is not true. Different respondents may interpret each question asked in the same words differently. In structured interviews both meaning and plurality of responses are lost. Researcher-using method of narrative understands that there can be standardized questions but there are no 'standardized meanings'.

18.9 FOCUS GROUP DISCUSSION

One of the methods of qualitative research that found ready acceptance by economists was that of focus group. It is frequently used in market surveys for quick appraisal of response to an economic programme, for evaluation studies and for media and health research. In focus group discussion, a small group of six to eight (market researchers prefer groups of 10-12 participants in each group) targeted audiences is gathered at a suitable location. They are encouraged to

have a frank discussion on a given topic and a recorder sitting outside the group will either audio record or take notes on the discussion. The researcher acts as a prompter posing questions on the subject of inquiry. Kruger (1988: 18) defines focus group as a "carefully planned discussion designed to obtain perceptions on a defined area of interest in a permissive, non-threatening environment".

Bloor et al. (2001) cautions that 'focus groups are the method of choice only when the purpose of the research is 'to study group norms, group meanings and group processes' (cf. Barbour, 2008:133)

Steps for focus group

- Topic for focus group should be of general interest to all the participants in the group.
- Number of participants can vary depending on the requirements of your research. Sampling is the key to data generation.
- 'Stimulus material' for initiating the discussion should be brief and encourage a free flowing conversation. Use of video clips, leaflets, cartoons, brief recollection of event or topic being discussed etc.
- Moderator must mange the discussion in a manner that without sounding intrusive, long extensive talks are avoided.
- Barbour (2008) suggests 'anticipate the analysis through attentive moderating'.
- Locate themes visible in the form of repetitive patterning of the responses.
- Second stage sampling could be considered to further explore any emerging theories, paradigms or even trends from the data.

These are some of the tools that participatory researchers frequently use to gather primary data. In this brief discussion, there are two more important steps that a participatory researcher must remember. First is **Grounded Theory** and second is **Data Sources and Sampling** procedure. We shall discuss these steps in next sections.

Check Your Progress 2

1) Make a distinction between focused interview and focused group discussion.

.....

.....

.....

.....

2) State the uses of Case Study Method.

.....

.....

.....

.....

3) Identify the steps involved in focus group discussion.

.....

.....

.....

.....

18.10 GROUNDED THEORY

Quantitative researchers are familiar with the procedure of preparing a research design that lists various steps of research including hypothesis and sample selection. Average participatory researcher will first go to the field, explore various research questions that he/she may have problematized and engage in casual conversations with the people on the ground. A research design is generally not prepared before empirically exploring the field situation. The research problem is then structured keeping in view the perspective, people's priorities and that of the researcher.

Glaser and Strauss (1967) proposed grounded theory as a method of inductive theorizing. In their opinion, theories are not to be tested but to be generated as research proceeds. They were of the opinion that "'good theory' is systematically discovered from and verified with the data of social research". Researchers further clarified 'generating a theory from data means that most hypotheses and concepts not only come from the data, but are systematically worked out in relation to the data during the course of research' (cf. Blaikie 2010:141). In this kind of analysis, comparative analysis is continuously generated along with theory building. Two kinds of theories are generated namely *substantive* and *formal*. Substantive theories are located in a specific context and are related to a specific process. Formal theories are generated at a higher level of generalization and can apply to number of substantive areas. Blaikie sums it brilliantly "research conducted from a grounded theory point of view is not a pre-planned linear process of testing hypotheses, but rather an evolving process in which what has been 'discovered' at any point will determine what happens next. An understanding of any phenomenon is seen as a developing process involving the collection of variety of data, by a variety of methods'. In simple terms, it suggests that you do not have to go to the field with a pre-decided hypothesis with the intent of verifying it but evolve your theory from the empirical evidence that was collected.

18.10.1 Data Sources and Sampling

Three types of data sources are recognised:

Primary Data – Data collected by researcher or researchers working on the project and they are responsible for collecting/generating, analysis and reporting. Majority of participatory researchers work with primary data.

Secondary Data – Data that is already collected by some other researchers, agencies, individuals (raw data) and is available in the public domain e.g. Census, National Sample Survey, NCAER Data, Economic Survey Data etc.

Tertiary Data – Analysed data either by the researcher or other researchers, institutions, agencies and its results are available for comparisons or reinterpretation.

Data for research can come from number of sources: (i) Individuals, (ii) Populations.

Samples from populations

Once data source is identified and it happens to be a population, it is not always possible to cover the entire population. In these situations a representative sample is identified. There are several methods of sampling. Probability and nonprobability sampling are commonly used. A sample is expected to contain one or more elements of the population/universe of study. Sampling frame using probability sampling is able to ensure elements of representation but that may not be possible in nonprobability sampling. Probability samples are selected using simple random and systematic random sampling. Non-probability sampling methods frequently used by participatory researchers are:

- **Accidental or Convenience sampling** – interviewing any individual, any where at any time without keeping research population in focus. This method is to be used only when no other method is feasible. This is often used in market research for quick feedback or for consumer surveys.
- **Quota sampling** – frequently used after selecting certain criterion to be explored in the research study. Quota is taken to represent the population and problem being researched. For doing an evaluation study for a product to be launched target the population e.g. for the launch of a new video game-target audience can be adolescents in the age group of 15-20 that come to a video store enquiring about new video games.
- **Judgemental or purposive sampling** – is another popular non-probability sampling method used for studying previously identified sub-set of population e.g. to analyse models for successful organizational management, a sample is selected from organizations that have shown successful management background.
- **Snowball method** – as the term suggests, this method of sampling works on the principle of networks, chain referral or reputational sampling. Researcher identifies an individual or a key respondent and he/ she may lead him to others having similar interests – the focus of the research. Internet networks, members of mac-book or Mercedes clubs etc. are few illustrative examples.

Participatory researchers are often criticized for using these sampling procedures arguing that these are not necessarily representative, as the sample is not objectively selected. One must remember that qualitative participatory research is invariably resource intensive and requires smaller samples as compared to large quantitative surveys.

18.10.2 Subjectivity

Qualitative researchers are not shy of admitting that there is always an element of subjectivity in research. Hollway and Jefferson (2000:3) draw attention to this important element of research: " As researchers, we cannot be detached but must examine our subjective involvement because it will help us to shape the way in which we interpret interview data".

18.10.3 Transcription

One of the important components of qualitative method of research or participatory approaches is rewriting the data procured from the field. The data is usually recorded in memory or sometimes on paper, and if circumstances permit it can be recorded using an audio or wherever possible a video recording. The data has to be transcribed for the purpose of analysis. This is one of the most difficult steps that many qualitative researchers have to learn to deal with.

18.11 ANALYSIS OF QUALITATIVE DATA

The findings collected with the help of qualitative tools involve a continuous process of data generation and analysis. 'It is iterative rather than being linear' (Barbour, 2008:189). Unlike quantitative methods in which first data is collected and subjected to statistical tool to obtain results and analysis follows the outcomes, in qualitative data, we do not generate results but detail 'findings'. While talking to people either in focused or in-depth interviews or in focus group discussions and case studies, people provide the answers through the process of reconstruction of these narratives. Qualitative data analysis uses '**constant comparative method**'. It involves constant comparing and contrasting of notes meaning thereby that the researcher must focus on 'who is saying what and in what context' (ibid: 217).

Steps for doing analysis of qualitative data:

- Compilations of field notes and observations recorded in the field diary.
- Some scholars prefer to have complete verbatim account of data transcribed but others opt to use 'indexed recordings and notes' (ibid: 192).
- Interrogation of the data and diligence shown by the researcher is the key to producing good transcripts and reliable generalizations.
- Codes and themes have to be developed to capture meaning.
- Grounded theory that uses concepts developed in the field by the respondent is an important way of theorizing and analysing in qualitative research.
- Both computer and Manual analysis can be done.
- NUD*IST, ATLAS/TI are some of the Computer software used for qualitative analysis.

18.12 REPORT WRITING

Richardson (2000: 923) refers to writing as "a *method of inquiry*, a way of finding out about yourself and your topic". A "personal tale of what went on in the backstage of doing research" (Ellis & Bochner, 2000: 741). Writing research differs in style depending on whether one is writing for a project agency for a journal or a dissertation/ thesis. Standard format comprises of **introduction/ background, methods, sample, results, discussion, and conclusion/ recommendations**. In qualitative research writing discussion is a challenge because the data collected is iterative or repetitive. Many qualitative researchers

prefer to use subtitle 'findings' to combine discussion and conclusions. As young researcher, you can always evolve your own style of writing but remember that when you are writing for an indexed journal, you are expected to conform to the writing style and referencing style of that particular journal which is always given at the back of the journal as instructions to the authors.

18.13 CRITICISM OF PARTICIPATORY METHODS

Opinions on the efficacy of participatory methods are widely divided. There is a growing demand for bringing in participatory methods of research in disciplines like economics, political science, management studies and law because standard methods of quantitative research have not yielded the desired results. There are others opposing reliance on participatory research arguing that there is every possibility of personal bias impacting the objectivity of research. It is restricted to a very small area and generalizations arrived on the basis of data generated by it may not be applicable to other areas. The methods of participatory research are slow. Many others call these studies impressionistic and unreliable and call it a 'soft' method of doing research.

To overcome some of the limitations of participatory research mentioned above, many researchers prefer to use a mixed methods approach. Let us remember that the details on mixed methods research have already been covered in Unit 6.

18.14 ADVANTAGES OF PARTICIPATORY RESEARCH

One of the important advantages of qualitative method of research is that it is "reflexive, flexible and iterative" (Cornwell & Jewkes: 1668; Chambers 1992). Diane Watt (2007:82) rightly points out that "since the researcher is the primary instrument of data collection and analysis, reflexivity is deemed essential"– this is a position supported by almost all researchers using participatory methods. Reflexive means that the researcher can take stock of his/her own reservations, prejudices and resolve them before recording responses obtained from the field. In simpler terms, it implies that in participatory researche, we deal with other human beings that look like us but may have differently acquired social values. The answer that they may give to an enquiry posed by us may be very different from the response that we expect or believe is the correct answer. Reflexivity in participatory research methods helps to distinguish qualitative research methods from quantitative methods of research. It teaches us to delineate **subjectivity** from **objectivity**. One of the strongest planks of **quantitative** researchers and **scientific (positivist)** methods of research was their claim that the results derived by these methods are 'objective'. Qualitative researchers accepted the charge of being 'subjectively objective' arguing that when a researcher interacts in the field with his respondents, he always carries a sense of subjectivity with him/her but in quantitative researches it is not admitted and dismissed by counting in sampling error. It was thus considered imperative to admit that bias by the qualitative researchers. They accepted that one must recognise it in terms of one's **reflexivity**. Qualitative/Participatory researchers have recognised the limitations and persistence practice that is required in recognising, admitting and becoming aware of this bias, particularly while writing their research report (Borochowitz, 2005; Denzin and Lincoln, 2005; Harding, 1991).

We discussed 'reflexivity' and 'flexibility' given in Cornwell and Jewkes statement describing advantages of participatory research. We now examine the third attribute of participatory research 'iterative' that means 'ability to repeat'– this can be inferred in several ways. One can repeatedly seek answers to the same query by enquiring about it in different ways to establish the validity of the response. Detailed narratives generated from the respondents may have responses that are repetitive of the statements made by other respondents and the common thread running through these conversations provides the window to ascertain the public opinion. It facilitates " 'bottom up' approach with a focus on locally defined priorities and local priorities"(Cornwell & Jewkes, 1995:1667).

Prior to the popularity and acceptance of participatory approaches, most planning was based on the philosophy of 'trickle down effect'– implying that the experts plan and enrich the upper crust and its impact will automatically get transferred to the lower stratum of the society. Felt needs of the people were not taken into consideration and the planning was top heavy. Many intervention projects with large funding failed because they failed to meet the immediate requirements of the local people. For example – several housing projects for the *adivasi* in remote parts of tribal heartland remained unutilized or underutilized, as the houses were not built in accordance with the tribal worldview and customary lifestyle. Participatory approach first explores people's immediate needs as perceived by the 'local people' and then suggests policies for meeting those needs.

Participatory researchers argue, "that the key element of participatory research lies not in methods but in the attitudes of researchers, which, in turn, determine how, by and for whom research is conceptualized and conducted" (ibid) If the researcher is not able to establish rapport with the respondent, he will not be able to generate good data. It is also important in participatory researches that the researcher takes keen interest in the problems that are addressed in his research query.

It is important to remember that you cannot become a good participatory researcher by learning these methods only through a formal lesson in a classroom. It requires personal experience in the field and repeated interactions with a large number of individuals to master the craft of participatory research. A lesson on participatory research empowers you with tools like the necessity to respect confidentiality of your respondents, to refrain from encroaching on their privacy, to respect their distinct traditions and cultural values, to conduct your interviews without offending or getting into an argument with them, respecting their opinion and recording it objectively and being **ethical** at all times. Your honesty in data generation and data analysis will give the necessary quality weightage to your research. In this context, it is worth to remember that one may encroach on the respondent's time and privacy. While writing your research never disclose the name and location of your respondents. Seeking their consent for interviewing and writing about them is equally important.

Participatory Research: An Illustration

You enquire from an individual in the market about the price of onions in the month of July 2014, when the market sentiment suggests, that the price of onions is very high. Contrary to the assumed answer the respondent suggests that "no, it is not as high as last year because last year in the same month, I

bought onions for Rupees ninety a kilogram but this year it is much cheaper, as I have paid only rupees forty per kilogram". His response is relative to his experience in the past and similarly other responses may differ depending on the experiences of different individuals. But when you ask standardized question in a yes or no format and ask a direct question with only two options of 'yes' or 'no', your results will not reflect 'actual sentiments' in the market. In the participatory approach, you will have flexibility of questioning as also response. On the contrary in the non-participatory approach you will get only two kinds of responses either saying 'the price of onions is very high' or some saying 'no it is not high'. Thus there is inbuilt flexibility in the participatory approaches. They provide a better understanding of the real market situation, giving more options for planning and interventions.

Check Your Progress 3

1) Enlist the various methods of sampling used in participatory research.

.....

.....

.....

.....

2) What is inductive theorizing?

.....

.....

.....

3) State the steps involved in analysis of qualitative data.

.....

.....

.....

18.15 LET US SUM UP

In participatory research, data is generated from an insider's perspective known as EMIC view. Unlike quantitative research, in this approach, people provide answers through the process of reconstruction of the narrations. The tools used in data collection are – participation observation, non-participation observation, quasi participation observation, in-depth interview, oral and life histories, narratives, case study method and focus group discussion etc. Qualitative researchers use various non-random sampling techniques, these include – convenience sampling, quota sampling purposive sampling, and snowball sampling. The qualitative data is analysed by 'constant comparative method'. The personal bias is the important criticism levelled against this method. Reflexivity, flexibility and iterative are the key advantages of this method and hence, participatory research by using mixed methods approach is increasingly being used to address complex issues in social science research.

18.16 KEY WORDS

Focus Group Discussion : It is a technique for a eliciting descriptive information/data from specific population subgroups (a group of 8 to 12 persons).

Reflexivity : Reflexivity refers to a situation where an individual becomes a self by being able to take the attitude of others and thereby reflect on his or her own behaviour.

Snowball Sampling : A technique for gathering subjects through the identification of an initial subject who is used to provide the names of other actors.

Quota Sampling : A method of selecting respondents for service by assigning quotas to the interviewers that define groups of respondents by using a few key demographic characteristics.

Iteration : It is an act of repeating a process with the aim of approaching a desired goal, target or result.

18.17 SOME USEFUL BOOKS/READINGS

Borochowitz, Dalit Yassour. (2005). "*Teaching a Qualitative Research Seminar on Sensitive Issues*". *Qualitative Social Work* 4:347-62.

Bogdan, R.C. and Taylor, S.J. (1975). *Introduction to Qualitative Research Methods: A Phenomenological Approach to the Social Sciences*. New York: John Wiley.

Barbour, Rosaline. (2008). *Introducing Qualitative Research: A Student Guide to the Craft of Doing Qualitative Research*. London: Sage Publications.

Denzin, Norman. K. & Lincon, Yvonna. S. (eds) (2000). *Handbook of Qualitative Research* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Denzin, Norman K. and Yvonna S. Lincoln (2005). "*Introduction: The Discipline and Practice of Qualitative Research.*" Pp. 1-33 in *The SAGE Handbook of Qualitative Research* edited by Norman K. Denzin, and Yvonna S. Lincoln, California: Sage Publications.

Harding, Sandra. (1991). *Whose Science? Whose Knowledge?: Thinking from Women's Lives*. Ithaca, New York: Cornell University Press.

Watt, Daine. (2007). *On Becoming a Qualitative Researcher: The Value of Reflexivity. The Qualitative Report*. Volume 12 Number 1 March 2007. Pp 82-101.

Yin, R.K. (1984) *Case Study Research*: *Design and Method*. Newbury Park, CA: Sage.

Websites:

http://www.nova.edu/ssss/QR

18.18 HINTS/ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 18.2
2) See Section 18.2
3) See Section 18.3

Check Your Progress 2

1) See Section 18.4 and 18.9
2) See Section 18.7
3) See Section 18.9

Check Your Progress 3

1) See Sub-section 18.10.1
2) See Section 18.10
3) See Section 18.12

UNIT 19 CONTENT ANALYSIS

Structure

19.0 Objectives

19.1 Introduction

19.2 Historical Background of Content Analysis

19.3 Content Analysis: Concept and Meaning

19.4 Terms Used in Content Analysis

19.5 Approaches of Content Analysis

- 19.5.1 Conceptual Content Analysis
- 19.5.2 Relational Content Analysis

19.6 Procedure Involved in Content Analysis

- 19.6.1 Formulation of Research Question
- 19.6.2 Determining Materials to be Included
- 19.6.3 Developing Content Categories
- 19.6.4 Selecting Units of Analysis
- 19.6.5 Code the Materials
- 19.6.6 Analyze and Interpret the Results

19.7 Uses of Content Analysis

19.8 Advantages and Disadvantages of Content Analysis

- 19.8.1 Advantages
- 19.8.2 Disadvantages

19.9 Let Us Sum Up

19.10 Key Words

19.11 References and Suggested Books

19.12 Answers or Hints to Check Your Progress Exercises

19.0 OBJECTIVES

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

- explain the concept and meaning of the content analysis;
- identify the areas of research in social sciences wherein content analysis can be used;
- discuss the various types of content analysis;
- state the various approaches of content analysis;
- describe the various steps involved in using the content analysis;
- enable you to analyse and interpret the results; and
- discuss the advantages and disadvantages of the content analysis.

19.1 INTRODUCTION

Many research Scholars take up historical overviews of economics as an area of research. They probe the research questions like how the different streams of

economics like institutional economics, Neo-classical economics, behavioural economics etc. have evolved over a period of time and how researcher's focus had shifted from one area to another isolating a particular paradigm change. Sometimes a researcher aims to lay concepts and meanings from the social actors' account of realities related to his identified research problem. They use the sources like recorded interviews, letters, journals, newspaper stories and verbal materials. In such situations, content analysis as data (information) analysis technique is used by the social scientists. Much of the subject matter of social sciences including consumer studies is in the form of verbal and nonverbal behaviour. The exchange process in the market place and the communication of the values of the exchange depends upon the written or spoken words. Much of consumer research has concentrated on the characteristics, opinion, or behaviour of the interpreter of communication messages. In media economics, issues like monopolistic competition in TV channels, competition and diversity in newspaper industry, effects of group ownership on daily newspaper content etc. can be probed by content analysis alongwith other statistical techniques. Similarly content analysis can also be used to analyse the cultural issues in cultural economics to capture pattern form of explanation. Hence, in this unit, we will discuss the concept of content analysis, its various types and approaches, the procedures involved in its application, and its advantages and disadvantages. Let us begin with historical background of content analysis followed by describing its concept and meaning.

19.2 HISTORICAL BACKGROUND OF CONTENT ANALYSIS

We find different approaches to analysis and comparison of texts in hermeneutic contexts (e.g. bible interpretations) like newspaper analysis, graphological procedures, content analysis, and the dream analysis by Freud etc. The basis of quantitative content analysis had been laid by Lazarsfield and Lasswell in the USA during 1920's and 30's. Historically, content analysis was a time consuming process. Analysis was done either manually, or by slow mainframe computers to analyze punch cards containing data punched by human coders. Single study could employ thousands of these cards. Human error and time constraints made this method impractical for large texts. However, despite its impracticality, content analysis was utilized as research method by 1940's. Although initially limited to studies that examined texts for the frequency of the occurrence of identified terms (word counts), by mid-1950's researchers started to consider the need for more sophisticated methods of analysis, focusing on concepts rather than simply words, and on semantic relationships rather than just presence. The first textbook about this method was published by Berelson in 1952. In the sixties of 20th century, content analysis found its way into linguistics, psychology, sociology, history, arts etc. The procedures involved in content analysis were refined by incorporating models of communication; analysis of non-verbal aspects, contingency analysis, computer applications. Since the middle of 20th century objections had been raised against a superficial analysis without respecting latent contents and contexts, working with simplifying and distorting quantification (Kracauer, 1952). Subsequently qualitative approaches to content analysis had been developed. Today, the use of modern computers makes content analysis much simpler and the data less vulnerable to human error.

19.3 CONTENT ANALYSIS: CONCEPT AND MEANING

Content analysis is a set of procedures for collecting and organizing information in a standardized format that allows analysts to make inferences about the characteristics and meaning of written and other recorded materials.

Content analysis is a multipurpose research method developed specifically for investigating a broad spectrum of problems in which the content of communication serves as the basis of inference. Content analysis covers both the content of the material and its structure. Content refers to the specific topics or themes in the material. Structure refers to form. Whether an article is prominently featured on the first page of a newspaper or buried in the middle section is a structural question. Careful reading of the written materials is necessary for all kinds of researches in social sciences. But content analysis is something different from careful reading of written materials in many ways.

According to **Berelson**, content analysis is a research technique for the objective, systematic, and quantitative description of the manifest content of communication. The term content analysis is used here to mean the scientific analysis of communication messages. The method being "scientific" catholic in nature, requires that the analysis be rigorous and systematic. According to **Paisley**, content analysis is a phase of information processing in which communication content is transformed through objective and systematic application of categorization rules into data that can be summarized and compared.

This definition underlines the point that content analysis is a systematic technique for analyzing message content and message handling. It is a tool for observing and analyzing the overt communication behaviour of selected communicators.

Content analysis , while certainly a method of analysis, is more than that, it is a method of observation. Instead of observing people's behaviour directly, or asking them to respond to scales or interviewing them, the investigator takes the communications that people have produced and asks questions of the communication (Kerlinger, 1968, p.544).

Content analysis refers to any procedure for answering the relative extent to which specified references, attitudes or themes permeate a given message or document (Stone, 1964).

Neuendorf defines content analysis as " the systematic, objective, quantitative analysis of message characteristics".

Based on the above definitions, the following characteristics of content analysis emerge: objectivity, systematic and generality.

Objectivity: To have objectivity, the analysis must be carried out on the basis of explicitly formulated rules which will enable two or more investigators to obtain the same results from the same documents. This requirement of objectivity gives scientific standing to content analysis and differentiates it from literary crtiticism.

Systematic: In a systematic analysis, the inclusion and exclusion of content or categories is done according to consistently applied criteria of selection. This

requirement is meant to eliminate elements in the content which do not fit in the analyst's thesis.

Generality: By generality, we mean that the findings must have theoretical relevance, purely descriptive information about content, unrelated to other attributes of content or to the characteristics of communicator or recipient of the message, is of little scientific value.

Thus, content analysis can be defined as a research technique for collecting and organizing information in a standardized format for making inference systematically and objectively about the characteristics of message.

19.4 TERMS USED IN CONTENT ANALYSIS

The following concepts and terms are frequently used in content analysis. Let us understand these terms.

Manifest Content Analysis: It involves simply counting words, phrases, or "surface" features of the text itself. It provides reliable quantitative data that can easily be analyzed using inferential statistics.

Latent Content Analysis: It involves interpreting the underlying meaning of the text. Latent analysis is different to manifest analysis because researcher must have a clearly stated idea about what is being measured. The value of the latent content analysis actually depends upon the researcher's ability to expose previously marked themes, messages and cultural values within the text. This analysis is widely used in qualitative content analysis.

Content Units: There are two types of content units i.e. the unit of analysis and unit of observation. The unit of analysis concerns the general idea or phenomenon being studied. The unit of observation concerns the specific item measured at an individual level.

Coding: Coding is the process whereby raw data are systematically transformed and aggregated into units which permit precise description of relevant content characteristics. There are two methods of coding: (i) deductive measurement, (ii) inductive measurement.

Deductive Measurement: It requires the development of specific coding categories before a researcher starts a content analysis. Deductive measurement is useful with an established set of coding categories or if a clear hypothesis or research question exists at the outset of the analysis.

Inductive Measurement: This method supports the practice of emergent coding, which means that the basic research question or hypothesis for a formal content analysis emerges from the units of observation. It entails creating coding categories during the analysis process. Emergent coding is useful in exploratory content analysis.

Coding Scheme: Every analysis begins with an existing coding scheme. Coding scheme is another phrase of coding categories and coding book, within which all instances of the content are analyzed or noted. Every coding scheme consist of the following statements.

Master Code Book: It provides the coders with explicit instructions and defines each word/ phrase/ aspect to be analyzed. It explains how to use code sheet.

Code Sheet: Code sheet provides the coders with a form on which they note every instance of every word/ phrase/ aspect being analyzed. The code sheet lists all the coding categories.

Coding Dictionary: It is used with computer based content analysis. A set of words/ phrases, part of speech …. That is used as the basis for a search text.

Check Your Progress 1

1) Identify the areas of research in economics wherein content analysis can be used.

.....

.....

.....

.....

2) What is the distinction between manifest content analysis and latent content analysis?

.....

.....

.....

.....

19.5 APPROACHES OF CONTENT ANALYSIS

Instead of being a single technique, content analysis is a collection of different approaches to the analysis of texts or more generally of messages of any kind. Important approaches frequently used in Social Sciences are discussed below:

19.5.1 Conceptual Content Analysis

Traditionally, content analysis has most often been thought in terms of conceptual analysis. In conceptual analysis, a concept is chosen for examination and the analysis involves quantifying and tallying its presence. It is also known as thematic analysis. The focus here is on looking at the occurrence of selected terms within a text or texts, although the terms may be implicit as well as explicit. While explicit terms obviously are easy to identify, coding for implicit terms and deciding their level of implication is complicated by the need to base judgments on a somewhat subjective system.

19.5.2 Relational Content Analysis

Relational content analysis, like conceptual analysis, begins with the act of identifying concepts present in a given text or set of texts. However, relational analysis seeks to go beyond presence by exploring the relationships between the concepts identified. Relational analysis has also been termed semantic analysis.

In other words, the focus of relational analysis is to look for semantic, or meaningful relationships. Individual concepts, in and of themselves, are viewed as having no inherent meaning. Carley (1992) asserts that concepts are 'ideational kernels.' These kernels can be thought of as symbols which acquire meaning through their connections to other symbol. Relational content analysis approach can be of three types:

i) **Affect extraction:** This approach provides an emotional evaluation of concepts explicit in a text. It is problematic because emotion may vary across time and populations. Nevertheless, when extended, it can be a potent means of exploring the emotional/psychological state of the speaker and/or writer.

ii) **Proximity analysis:** This approach, on the other hand, is concerned with the co-occurrence of explicit concepts in the text. In this procedure, the text is defined as a string of words. A given length of words, called a *window*, is determined. The window is then scanned across a text to check for the co-occurrence of concepts. The result is the creation of a concept determined by the *concept matrix*. In other words, a matrix, or a group of interrelated, co-occurring concepts, might suggest a certain overall meaning.

iii) **Cognitive mapping:** This approach is one that allows for further analysis of the results from the two previous approaches. It attempts to take the above processes one step further by representing these relationships visually for comparison. Whereas affective and proximal analysis function primarily within the preserved order of the text, cognitive mapping attempts to create a model of the overall meaning of the text. This can be represented as a graphic map that represents the relationship between concepts. In this manner, cognitive mapping lends itself to the comparison of semantic connections across texts. This is known as map analysis which allows for comparisons to explore "how meanings and definitions shift across people and time".

19.6 PROCEDURE INVOLVED IN CONTENT ANALYSIS

The various steps involved in content analysis are: (i) Formulation of research questions, (ii) Determining materials to be included, (iii) Developing content categories, (iv) Selecting and Finalizing units of analysis, (v) Code the materials, (vi) Analyze and interpret the results. These are discussed below:

19.6.1 Formulation of Research Questions

Content analysis begins with a specific statement of the objectives or research questions. The objective of content analysis is to convert recorded "raw" phenomenon into data, which can be treated essentially in a scientific manner so that body of knowledge may be built up. Objectives are precisely worded questions that investigator/s is/are trying to answer. By making a clear statement of the research question, the researcher can ensure that the analysis focuses on those aspects of content which are relevant for the research. The question should be based on a clear understanding of research needs and the available data. Therefore, the selection of the topic should be such that can be answered by analyzing the appropriate communication content.

In general content analysis can be used to answer " What" but not " Why" question. It helps analysts to describe or summarize the content of written material , the attitudes or perceptions of its writers, or its effects on the audience. For example, if analysts want to assess the effects of different women empowerment programmes on the lives of younger and middle aged women in rural and urban area. Content analysis of open–ended interview / responses could be used to identify their outlook, attitudes and security about their life.

19.6.2 Determining Materials to be Included

The next step of content analysis is to decide relevant communication content to answer the research question and to determine the time period to be covered. Content analysis can be used to study any recorded materials as long as the information is available to be reanalyzed for reliability checks. It is used most frequently to analyze written materials to study any recorded communication including TV programmes, movies, photographs, regulations, other public documents, workplaces, case studies, reports, answer to survey questions, newspapers, news release, books, Journals article, letters etc. Speeches and discussions can also be analyzed.

Sampling: Sampling is necessary if the population is too extensive to be analyzed. Thus a sample should be selected from the population in order to make valid conclusion and generalization about a population. Simple random sampling, interval sampling, cluster sampling and multistage sampling techniques are used in content analysis.

19.6.3 Developing Content Categories

Content analysis is no better than its categories., since they reflect the formulated thinking , hypotheses, and the purpose of the study. Categories provide structure for grouping and recording units. Developing the category system to classify the text is the heart of content analysis. Berelson (1952) has emphasized the importance of formulating coding categories by quoting that " content analysis stands or falls by its categories. Particular studies have been productive to the extent that categories were clearly formulated and well adequate to the problem and to the content".

To be useful, every content category must be thoroughly defined, indicating what type of material be and not to be included. Chadwick et.al. (1984) have also emphasized the following three characteristics of content categories i.e. (i) Categories should be exhaustive so that all relevant items in the material being studied can be placed within a category. (ii) It should be mutually exclusive so that no item can be coded in more than one category. (iii) Thirdly categories should be independent so that recording of unit's category assignment is not affected by the category assignment of other recording units.

Category format: Categories can be conceptualized in many ways. Some common category formats are grouping, scales, and matrices. Structured category format increase coding efficiency especially when the number of categories are large. Scales provide the rank ordering information. Matrices are useful formats when analysts seek more information about issues than simply whether they are present or absent.

19.6.4 Selecting Units of Analysis

Once the categories have been identified, the analyst would be interested in determining the unit of content for classification under the content categories and the system of enumeration for the same. The unit of analysis is the smallest unit of content that is coded under the content category. The unit of analysis vary with the nature and objective of the analysis. Thus, the unit of analysis might be a single word, a theme, a letter, a symbol, a news story, a short story, a character or an entire article etc. There are two kinds of unit of analysis : *Recording unit and Context unit.*

The Recording Unit is the specific segment of content in which the occurance of a referene /fact is counted or the unit can be broken down so that reference/ facts can be placed in different categories.

The Context Unit is the larger body of the content that may be searched to characterize the recording unit. Context units set limits on the portion of written material that is to be examined for categories of words or statements. A recording unit is the specific segment of context unit in written material that is placed in a category. For example, if the coding unit is the word, then the context unit may be the sentence or the pargraph in which the word appears and characterizes the recording unit.

Word: The smallest unit generally used in the content analysis as a unit is a word. Lasswell(1952) calls word as a symbol and may include word compounds e,g, phrases as well as single word. In this type of research one might study the relative occurrence of key symbols or value laden terms until the content has been systematically examined relevant to the hypotheses of the study.

Theme: The theme is a single assertion about a subject. The theme is among the most useful units of content analysis because issues, values, beliefs, and attitudes are usually discussed in this form.

Character: Character may be defined as a use of fictional or historical character as the recording unit is also employed.

Item: The item is the whole natural unit employed by procedures of symbolic material. It may be the entire speech, radio programme, letter to the editors, or news story.

Space and Time Measures: Some studies have classified content by physical division such as column inch, the line or paragraph, the minutes or the foot film.

Finalizing units of analysis: In content analysis, the counting or quantification of the units is performed by using three methods of enumeration : 1) Frequency, 2) Intensity or Direction, and Space/ Time.

Frequency: Frequency simply means counting whether or not something occurs and how often (how many times).

Direction: Direction is nothing but the direction of messages in the content along some continuam e,g, positive, negative, supporting or opposed.

Intensity: Intensity is the strength or power of a message in a direction. For example, the characteristic of forgetfulness can be minor (e.g. not remembering to take the keys when leaving home, taking time to recall the name of someone whom you have not seen in years) or major (e.g. not remembering your name, not recognizing your children).

Space: A researcher can record the size of the text messages or the amount of space or volume allocated to it. Space in written text is measured by counting words, sentences, paragraphs, or space on a page (e.g. square inches) for video or audio text. Space can also be measured by the amount of time allotted.

To explain the differences among the above described quantification levels and how they relate to constructing categories, let us give a hypothetical example of 'Importance of FDI in public sector organizations'. The analyst has a major source of information as newspaper, articles, public documents, transcripts of interview with political leaders, and public officials. For each issue of each newspaper in the sample, the analysts add together number of column inches from all news articles/editorials to find he total number of space for each position in addition to the coding, the name , location, and date of each newspaper. The analyst who use this level of quantification have to assure that the differences they find in amounts of space are valid indicators of relative emphasis or importance. At the next level of quantification, the analyst can code the frequency of recording units by tallying the number of times each issues or statements occurs in the text. At the third level of quantification, analyst provide code for intensity. Frequencies are counted but each coded statement or issues is also adjusted by a weight that measures related intensity.

19.6.5 Code the Materials

Coding the unit of analysis into content category is called coding. Defining categories and preparing coding schedule for analysis and coding of the content are almost simultaneous steps. Material can be coded either manually or by computers, depending on the sources available and format of the material. After deciding how the material will be coded, the analyst writes the instructions for coding. According to Kripendorff (1980), the guidelines for coding instructions include definition of recording units and procedures for identifying theme, descriptions of variables and categories, outline of cognitive procedures used in placing data in categories, and instructions for using and administering data sheets.

Pre-testing: Pretesting is an important step before actual coding begins. It involves coding a small portion of the material to be analyzed. A pretest enables the analyst to determine whether (1) the categories are clearly specified and meet the necessary requirement, (2) the coding instructions are adequate and, (3) the coders are suitable for coding. These are determined on the basis of reliability among coders band consistency in individual coding decision. If analyst find that material can be coded with high reliability then actual coding begins.

19.6.6 Analyzing and Interpreting the Results

The main objective of content analysis is to analyze the collected information with regard to the proposed objective of the analysis. The analysis involves summarizing the coded data, discovering patterns and relationships within the data, testing hypotheses about the patterns and relationship to assess the validity of the analysis.

The absolute and relative frequencies are most commonly used for summarizing the data. Absolute frequency might be the number of times statements or issues are found in the sample. A relative frequency might be represented by a percentage of the sample size.

Another way of analyzing content analysis is to examine relations among variables by cross tabulating the cooccurrence of variables. Other techniques for discovering patterns of relationship in data include contingency analysis, cluster analysis and factor analysis.

One important development in analyzing content analysis of the data is the use of computer. Computer programme like "General Inquirer" can identify within a body of text , those words and phrases that belong to specified categories (Stone et.al., 1996). Other computer programme like SPSS, Atlas titi, Minitab etc. are also very useful in both quantitative and qualitative content analysis.

Whatever the technique used, a final and important task is to assess the validity of the result by relating them to other data that are known to reasonably valid. The inference a researcher can or cannot make on the basis of research is critical in content analysis. Content analysis describes what is in the text. It cannot reveal the intentions of those who created the text or the effects that messages in the text have on those who receive them.

Check Your Progress 2

1) State the different sources of content analysis.

.....

.....

.....

.....

2) Which methods are used to quantify the units of analysis?

.....

.....

.....

.....

.....

3) Briefly describes the different steps involved in the process of content analysis.

.....

.....

.....

.....

.....

19.7 USES OF CONTENT ANALYSIS

Content analysis is a scientific, objective, systematic, quantitative and generalizable description of the communication content. This method can be used to understand a wide range of themes such as social change, cultural symbols, changing trends in the theoretical content of different disciplines, changes in mass media content, nature of news coverage of social issues, election issues as reflected in mass media. Content analysis can be used to examine anything that you see around you.

It is used in several mass media and literature to cultural studies, psychology, economics, political science, gender, age issues, as well as many other fields where inquiry is made. Written documents, pictures, videos, can also be used for content analysis.

In consumer behaviour and marketing, content analysis has been used to study the following questions (Cited in Kassarjian, 1977).

1) What are product and company images of selected consumer goods as reflected in the mass media. (Stone, Dunphy & Bernstein, 1966).

2) Which of the several decision making models ,for example conmpensatory, lexicographic, risk etc. are used by magazine and television advertisers (Wright and Barbour, 1975).

3) What is the ease of readability of various marketing, advertising and consumer research Journals. (Lacho, Stearns, & Villere, 1975).

4) What are the changing values in society as reflected in the analysis of mass periodical fiction (Johns-Heine, & Gerth, 1949)

The above example indicates the wide applicability of content analysis. Many researchers have explored changes in women's role, sexual behaviour and health and violence by analyzing the content in television and movies messages (Olson, 1994).

Application of Content Analysis: An Illustration

A Research Study entitled "Gandhi's Constructive Programme: A Study of Its Contemporary Relevance": Undertaken by Ms. Raunak Ahmad, Research Scholar, (Gandhi and Peace Studies) IGNOU, New Delhi for award of the Ph.D Degree.

The main objectives of the study were to:

(i) find out the gaps in the value and principles of social work profession and its applicability in the Indian social setup, (ii) examine the perception of social work students in India towards Gandhian principles of Constructive Programme, (iii) study the relevance of principles of Gandhian Constructive Programme from the perspective of students of social work discipline in India and (iv) to pin point the importance of Gandhian Constructive Programme to strengthen the indigenous base of contemporary social work education.

Research Methodology: Keeping in view the objectives and issues to be probed in the study, the method of content analysis was considered appropriate to conduct the study. The study was undertaken for an in-depth and systematic analysis into the convolution of the phenomenon of Gandhi's Constructive Programme and its contemporary relevance.

In order to interrogate the values and principles of social work profession and its subsequent applicability in Indian social setup, a detailed study of the comprehensive MSW (Master of Social Work) syllabus of seven central universities was undertaken. These seven universities were: University of Delhi, Jamia Millia Islamia, Pondicherry University, Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Banaras Hindu University, Central University of Rajasthan and Maulana Azad National Urdu University. Through conceptual content analysis, the occurrence of the selected terms in the current social work curriculum was studied. The text in the curriculum was carefully examined and was classified under these categories: foundation of social work, theory and practice of group work, working with communities, social work research, social group work, social casework and counseling, history and philosophy of social work etc. These categories were further subcategorized for deeper analysis. Texts depicting similar meaning were put under single category and none of the category was repeated. The process of content analysis further went into detail when the common foreign scholars work and common Indian scholars works prescribed under the curriculum of universities were created as two separate categories. It helped to present a comparative picture. Categories were then coded and the coded data were summarized and analyzed.

Main Results

1) Methods of social work such as community work, casework, group work, social welfare management have been covered in the curriculum of all the seven universities.

2) Majority of universities in the list are also offering social work history and foundation as well as social work research.

3) The social action and social policy and planning being development oriented are more meaningful in Indian context but are offered by minimum number of universities.

4) The Gandhian ideology of social work is marginally touched in social work curriculum of most of the universities under study.

5) 67 per cent of the suggested readings recommended in the curriculum are of foreign scholars work.

6) The books on Gandhi, Gandhian ideology of social work written by Gandhi are barely included in the suggested reading list of the curriculum of these universities.

Apart from above, this study also made use of the method of content analysis in studying the Gandhian principles of Constructive Programme. For this purpose, the various documents introduced by Gandhi in the form of vows, pledges, principles etc. were used. In this manner, the method helped to draw the final set of principles of Constructive Programme which can further be studied in the context of its relevance in social work education.

19.8 ADVANTAGES AND DISADVANTAGES OF CONTENT ANALYSIS

19.8.1 Advantages

The Advantages of content analysis as a research technique can be summarized as follows:

1) The greatest advantage of content analysis is its economy in terms of time and money. There is no requirement for a large research staff. No special equipment is needed.

2) The methods allows the correction of errors. In content analysis, it is usually easier to repeat a portion of the study than it is in other research method.

3) Content analysis permits the study of processes occurring over a long time.

4) Content analysis has the advantage of all unobtrusive measures that it has any effect on the subject being studied.

5) It can present an objective account of events, themes, issues, and so forth, that may not be immediately apparent to a reader or viewer .

6) It deals with large volume of data. Processing may be laborious but of late computer has made it easy.

19.8.2 Disadvantages

1) It is limited to the examination of recorded communication. Such communication may be oral, written, or graphic, but they must be in some fashion which permits analysis.

2) Content analysis may not be as objective as it claims since the researcher must select and record data accurately. In some instances the researcher must make choices about how to interpret particular form of behaviour.

3) It has both advantages and disadvantages in terms of reliability and validity. Problems of validity is unlikely unless researcher happen to be studying communication process itself.

4) It describes, rather explains people's behaviour. It does not tell us what behaviour means to those involved and those watching.

5) By attempting to quantify behaviour, this method may not tell us very much about the quality of people's relationship,

Check Your Progress 3

1) What are the different areas where the content analysis can be used as a research technique?

.....

.....

.....

.....

2) Do you think that the technique of content analysis maintain objectivity? Give reasons.

.....

.....

.....

.....

.....

19.9 LET US SUM UP

Content analysis is a multipurpose research method developed specifically for investigating a broad spectrum of problems in which the content of communication serves as the basis of inference. Content analysis is a set of procedures for collecting and organizing information in a standardized format that allows analysts to make inferences about the characteristics and meaning of written and other recorded materials. Broadly content analysis is observed in two forms – manifested content analysis and latent content analysis. Mainly two approaches are followed in application of content analysis: conceptual analysis and Relational analysis. In conceptual analysis, a concept is chosen for examination, and the analysis involves quantifying and tallying its presence. Relational analysis also termed as semantic analysis too begins with the act of identifying concepts present in a given text or set of texts. However, relational analysis seeks to go beyond presence by exploring the relationships between the concepts identified. The various steps involved in content analysis include: objective or formulation of research question, determining materials to be included developing content categories, selecting and finalizing units of analysis, code the materials, and analyze and interpret the results. Content analysis can be used to understand a wide range of themes such as social change, cultural symbols, changing trends in the theoretical content of different disciplines, election issues as reflected in mass media. It can be used to examine anything that you see around you. Content analysis has several advantages and limitations.

19.10 KEY WORDS

Objectivity : To have objectivity, the analysis must be carried out on the basis of explicitly formulated rules which will enable two or more investigators to obtain the same results from the same documents.

Systematic : In a systematic analysis the inclusion and exclusion of content or categories is done according to consistently applied criteria of selection.

Generality : By generality, we mean that the findings must have theoretical relevance unrelated to other attributes of content or to the characteristics of communicator or recipient of the message.

Manifest Content Analysis: It involves simply counting words, phrases, or "surface" features of the text itself.

Latent Content Analysis : It involves interpreting the underlying meaning of the text.

Unit of Analysis : The unit of analysis concerns the general idea or phenomenon being studies.

Unit of Observation : Unit of observation concerns the specific item measured at an individual level.

Word : The smallest unit generally used in the content analysis as a unit.

Theme : The theme is a single assertion about a subject.

Character : Character may be defined as a use of fictional or historical character.

Coding : Coding is the process whereby raw data are systematically transformed and aggregated into units which permit precise description of relevant content characteristics.

Coding Scheme : Coding scheme is another phrase of coding categories and coding book, within which all instances of the content being analyzed or noted.

Master Code Book : It provides the coders with explicit instructions and defines each word/ phrase/ aspect to be analyzed.

Code Sheet : Code sheet provides the coders with a form on which they note every instance of every word/ phrase/ aspect being analyzed.

19.11 REFERENCES AND SUGGESTED BOOKS

Altheide, David L. (1996). *Qualitative Media Analysis. Qualitative Research Methods Vol. 38*. Thousand Oaks: Sage Publications.

Barcus, F.E., (1959). Communication analysis : analysis of the research, 1900-1958 Unpublished doctoral dissertation, University of Illinois.

Berelson, B. (1952). *Content Analysis in Communication Research,* New York: The Free Press

Budd, R.W., Thorp, R.K. & Donohue, L. (1967). *Content Analysis of Communications*, New York: The Macmillan Co.

Carley, K. (1992). Coding choices for textual analysis: A comparison of content analysis and map analysis. Unpublished Working Paper.

Chadwick, B.A., Bahar, H.M. & Albrecht, S.L. (1984). Content analysis. In B.A. Chadwick et.al., *Social Science Research Methods* (pp. 239-257), New

de Sola Pool, I. (1959). *Trends in Content Analysis*. Urbana, Ill: University of Illinois Press.

Gerbner, G., Holsti, O. R., Krippendorff, K., Paisley, W.J., and Stone, P.J., eds. (1969). The analysis of Communications Content : Developments in Scientific Theories and Computer Techniques, New York : Wiley. Jersey: Prentice –Hall.

Kerlinger, F.N. (1986). *Foundations of Behavioural Research* (3rd ed), New York: Holt, Rinehart and Winston.

Kracauer, S. (1952). The challenges of qualitative content analysis. Public Opinion Quarterly ,16,631-642.

Krippendorff, K. (1980). *Content Analysis: An introduction to its Methodology*,

Lacho, K.J., Stearns, G. K., and Villere, M.F. (1975).An analysis of the reliability of marketing journals. Combined Proceeding, American Marketing Association, 489-497. London: Sage Publications.

Mostyn, Barbara (1985). The content analysis of qualitative research data: A dynamic approach. In Michael Brenner, Jennifer Brown & David Canter (Eds.), *The research interview, uses and approaches* (pp.115-145). London: Academic Press.

Olson, B. (1994). Sex and the soaps: A comparative content analysis of health issues, *Journalism Quarterly*, 71(4): 840-850.

Paisley, W.J., (1969). Studying style as deviation from encoding norms. In The Analysis of Communications Content : Developments in Scientific Theories and Computer Techniques,, eds. G. Gerbner, et al., New York : Wiley, 133-146.

Stone, P. J., Dunphy, D. C.m and Bernstein, A. (1966). The analysis of product image. In The General Inqurier : A Computer Approach to Content Analysis, eds. P.J. Stone, et al , Cambridge , Mass : The MIT Press.

Wittkowski, Joachim (1994). *Das Interview in der Psychologie. Interviewtechnik und Codierung von Interviewmaterial*. Opladen: Westdeutscher Verlag.

Wright, P. And Barbour, F . (1975). The relevance of decision process model in structuring persuasive messages.

19.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 19.1 and Section 19.7

2) See Section 19.4

Check Your Progress 2

1) Written materials, recorded communications including TV programmes, movies, photographs, regulations & other public documents, work places, case studies, reports, newspapers, new press releases, books, journals, articles, letters etc.

2) See Sub-section 19.6.4

3) See Sub-section 19.6.5 and 19.6.6

Check Your Progress 3

1) See Section 19.7

2) See Sub-section 19.8.2

UNIT 20 ACTION RESEARCH

Structure

20.0 Objectives

20.1 Introduction

20.2 Historical Background of Action Research

20.3 Definition of Action Research

20.4 Principles of Action Research

20.5 Characteristics of Action Research

20.6 Models of Action Research

- 20.6.1 Kenmis and McTaggart's Spiral Model
- 20.6.2 Elliot's Action Research Model
- 20.6.3 O' Leary's Cycles' of Action Research Model
- 20.6.4 Stringer's Interacting Spiral Action Model
- 20.6.5 Kurt Lewin's Action Research Spiral Model
- 20.6.6 Calhoun's Action Research Cycle Model
- 20.6.7 Bachman's Action Research Spiral Model
- 20.6.8 Riel's Progressive Problem Solving Model
- 20.6.9 Hendricks's Action Research Model

20.7 Steps Involved in Action Research

20.8 Advantages and Disadvantages of Action Research

- 20.8.1 Advantages
- 20.8.2 Disadvantages

20.9 Let Us Sum Up

20.10 Key Words

20.11 References

20.12 Answers or Hints to Check Your Progress Exercises

20.0 OBJECTIVES

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

- state the meaning, type and importance of action research;
- describe the principles guiding the action research;
- explain the various models of action research;
- learn the steps involved in action research; and
- discuss the advantages and disadvantages of action research.

20.1 INTRODUCTION

We have learned in Unit 1, 5, and 6 that two approaches – quantitative research associated with dominant post positivist paradigm and qualitative research associated with interpretative paradigm and critical theory paradigm have been followed in undertaking research in Social Sciences. Under critical theory paradigm, research and practice are integrated activities and both (researcher

and practitioner) guide each other. With emergence of critical theory approach as an alternative to positivism and post positivism, participatory method and emancipatory methods are also used to conduct the research studies in economics. You have already studied participatory method in Unit 18. Emancipatory method relies upon the **process of knowing through doing** rather than generation of knowledge through conceptualization and theorizing. This strategy of knowledge generation is termed as **action research** wherein it is believed that theory is only really useful if it is put in service of a practice focusing on achieving social change.

Action research is a powerful tool for change and improvement at the local level. Kurt Lewin's , one of the pioneer person for action research, work was intended to change the life of disadvantaged groups in terms of housing, employment and their working conditions. Action research is essentially research through action. It is usually a collaborative activity which involves input from people who are likely to be affected by the research. It involves deep inquiry into one's professional action. The researchers examine their work and look for opportunities to improve. They work with others to propose a new course of action to help their community, improve its work practices. They seek evidence from multiple sources to help them, analyze reactions to the action taken. They recognize their own view as subjective and seek to develop their understanding of the event from multiple perspectives. The researcher uses data collection to characterize the forces in ways that can be shared with practitioner. This leads to a new plan for action.

Thus, action research aims to contribute both to the practical concerns of people in an immediate problematic situation and to further the goals of social sciences simultaneously. There is a dual commitment in action research to study a system and concurrently to collaborate with members of the system in changing it in what is together regarded as a desirable direction. Accomplishing this twin goal requires the active collaboration of researchers and client and thus it stresses the importance of co-learning as a primary aspect of the research process.

Hence, in this unit, we shall discuss the concept and type of action research, its various models, steps involved in action research, its uses, advantages and disadvantages. Let us start with the historical evolution of action research.

20.2 HISTORICAL BACKGROUND OF ACTION RESEARCH

It is not certain who invented action research. Origin of Action research is considered to begin from the work **Kurt Lewin in 1940's**. He was strong exponent of research action in its concern with power relations between researcher and researched and the rights of the individuals. He was concerned with the social problems and focused on participative group process for addressing conflict, crises and change within organization. Lewin first coined the term action research in 1946 in his paper "Action Research and Minority Problems". He characterized action research as: "A comparative research on the conditions and effect of various forms of social action and research leading to social action, using a process of a spiral steps, each of which is composed of a circle of planning, action and fact – finding about the result of the action".

Eric Trist was another major contributor to the field. He was a social psychiatrists. Both, Lewin and Trist applied their research to systematic change in and between organizations. They emphasized direct professional – client collaboration and affirmed the role of group relations as basis for problem solving. Both were strong proponents of the principle that decisions are best implemented by those who help them. Alternatively, **Deshler and Ewart (1995)** suggested that action research was first used by John Collier to improve race relations at the community level when he was the Commissioner of Indian Affairs prior to and during the Second World War, and Cooke (undated) appears to provide strong support for this. It is, therefore, unlikely that we will ever know when or where the method originated, simply because people have always investigated their practice in order to make better or improve. **Rogers' (2002) account of John Dewey's (1933)** notion of reflection, for instance, shows that it is very similar, and one could also point to the ancient Greek empiricists as using an action research cycle. Action research is difficult to define for two linked reasons: first, it is such a natural process that it comes in many different guises, and second, it has been developed differently for different applications. Almost immediately upon Lewin's coining of the term in the literature, action research was seen as a general term for four different processes: diagnostic, participant, empirical and experimental.

20.3 DEFINITION OF ACTION RESEARCH

According to **Frost (2002)** 'Action research is a process of systematic reflection, enquiry and action carried out by individuals about their own professional practice'. **Hopkins (1985)** suggests that action research is the combination of action and research rendering action as a form of disciplined inquiry, in which a personal attempt is made to understand, improve and reform practice. **Ebbutt (1985)**, too, regards action research as a systematic study that combines action and reflection with the intention of improving practice.

Corey (1953) views action research as a process in which practitioners study problems *scientifically* so that they can evaluate, improve and steer decision-making and practice. 'Action research is a term used to describe professionals studying their own practice in order to improve it'. **Kemmis and McTaggart (1992)** also argued that 'to do action research is to plan, act, observe and reflect more carefully, more systematically, and more rigorously than one usually does in everyday life'. 'Action research is ... usually described as cyclic, with action and critical reflection taking place in turn. 'Action research is a flexible spiral process which allows action (change, improvement) and research (understanding, knowledge) to be achieved at the same time' (Dick, 2002). **Cohen and Manion (1980)** described it as "essentially on the spot procedure designed to deal with a concrete problems located in an immediate situation. This means that a step by step process is constantly monitored over varying periods of time and by a variety of techniques ….. diaries, interviews, case studies, etc., so that ensuing feedback may be translated into modifications, adjustments, directional changes, redefinitions as necessary".

A more philosophical stance on action research is taken by Carr and Kemmis, who regard it as a form of 'self-reflective enquiry' by participants, which is undertaken in order to improve their understanding of their practices in context with a view to maximizing social justice... **Kemmis and McTaggart (1992)** suggest that: Action research is concerned equally with changing *individuals*, on

the one hand, and, on the other, the *culture* of the groups, institutions and societies to which they belong. Action research is designed to bridge the gap between research and practice (Somekh 1995) thereby striving to overcome the perceived persistent failure of research practice.

Thus, Action research combines diagnosis, action and reflection, focusing on practical issues that have been identified by participants and which are somehow both problematic yet capable of being changed. Over the last decade, action research has begun to capture the attention of teachers, administrators, and policymakers around the country. Kemmis and McTaggart (1988) defined it very effectively by encompassing the different definitions of action research.

> Action research is a form of *collective* self-reflective enquiry undertaken by participants in social situations in order to improve the rationality and justice of the own social or educational practices, as well as their understanding of these practices and the situations in which these practices are carried out. . . . The approach is only action research when it is c*ollaborative*, though it is important to realize that the action research of the group is achieved through the *critically examined action* of individual group members.
>
> (Kemmis and McTaggart 1988)

Most simply put, action research (AR) implies a process of research where the purpose of the research is not only to study the existing reality but to engage in an effort to transform it. In this model, research assumes a catalytic role and produces both a new dynamic and concrete, suitable changes in the reality that can then be actively inducted into the process of knowledge creation. A requirement of AR is that the separation between "researcher and researchee" is dissolved—so as to avoid the weakness of conventional methods which view "affected persons and groups" as being passive and incapable of analysing their own situation and identifying solutions to their own problems. (Mukherjee, 2006.)

Check Your Progress 1

1) How action research is distinct from pure and simple applied research?

.....

.....

.....

.....

.....

2) What are the aims of action research?

.....

.....

.....

.....

.....

3) State the concept of action research coined by Kurt Lewin.

.....

.....

.....

4) Do you think that action research play a catalysist role? Give reasons for your answer in two sentences.

.....

.....

.....

20.4 PRINCIPLES OF ACTION RESEARCH

A set of six principles according to winter guide an action research. These principles are discussed below:

1) **Reflective Critical Analysis**

It is the process of becoming aware of our own perceptual biases. For example, description or narration of a situation in any notes, transcripts, or official documents imply that it is factional and true. It claims to be reliable and authentic. But truth in any social setting or social reality is in relative to the teller. The principle of reflective critical analysis ensures people to reflect on issues and practices and make explicit interpretations, biases, assumptions and concern upon which judgments are made. In this way, practical description can give rise to theoretical considerations.

2) **Dialectical Critical Analysis**

It is a way of understanding the relationships between the elements that make up various phenomena in our context. Social reality is validated or shared through language. Phenomena are conceptualized in dialogue. Therefore a dialectical critical analysis is required to understand a set of relationship both between the phenomenon and its context, and between the elements constituting the phenomenon. The important point is to focus attention on those constituent elements that are unstable or opposing one another simply because these elements are most likely to create changes.

3) **Collaborative Resource**

The principles of collaborative resources presuppose that each person's ideas are equally significant as potential resources for creating interpretive categories of analysis, which are negotiated among the participants. It makes possible the insight gathered between many viewpoint and within a single viewpoint. Hence everyone's view is taken as a contribution to understand the situation. Participants in action research are co-researchers.

4) **Risk**

It is an understanding of our own taken-for-granted processes and willingness to submit them to critique. Normally, the change process is considered a threat to all previously established ways of doing things, thus creating

psychiatric fears among the participants. One of the most prominent fears emanates from the risk of ego generation from open discussion of one's interpretations, ideas and judgment. Initiator of action research will use this principle to reduce other's fears and invite participation by assuring that they, too, will be the subject to the same process and that whatever be the outcome, learning will take place.

5) **Plural Structure**

It involves developing various accounts and critiques, rather than a single authoritative interpretation. The action research comprises of a multiplicity of views, Commentaries and critical analysis which lead to multiple possible actions and interpretation. This inquiry requires much texts for reporting. Thus, there will be many explanations made clear and range of options for action presented. A report, therefore, acts as a support for ongoing discussion among collaborators rather than a final conclusion of fact.

6) **Theory, Practice and Transformation**

Theory and practice are considered two inter-dependent yet complementary phases of the change process. For action researchers, theory informs practice and practice refines theory in a continuous transformation. In any setting, people's actions are based on implicitly held assumptions, theories and hypotheses. With every observed result, theoretical knowledge is enhanced. Thus, a single change process has two interwined aspects. It is up to the researchers to make the theoretical justifications explicit for action, and to question the basis of that justification. The ensuring practical applications as a follow up are subjected to further analysis in five cycles that continuously alternates emphasis between theory and practice.

20.5 CHARACTERISTICS OF ACTION RESEARCH

Action research has the following characteristics:

1) **It is *collaborative, i.*e**. everyone's view is taken as a contribution in understanding the situation. Moreover, if a problem is faced by a practitioner in a particular situation (say a school), action research can be collaborative where practitioners facing similar problems in nearby schools can collaborate to find solutions of a problem.

2) **Action research helps in systems planning and restructuring**. For example, if a primary teacher finds that in his/her class the students are not able to concentrate, the teacher starts finding the reasons for the same. After analyzing the situation, the teacher finds that most of the children often observe other children playing in the playground because of opening of classroom window towards playground. They, therefore, are not able to concentrate in their studies in the class. Now, what do you think a teacher should do? Well, in such a case a teacher can change the seating plan of the classroom. This is the way a teacher gets involved in restructuring the class.

3) **Action research is a small-scale intervention**. Its objective is to bring out changes in the functioning of the practitioner himself/herself. It may or may not have applicability for others. Action research is a narrowly focused research undertaken by teachers and other practitioners in a given specific situation and context.

4) **Contextual nature** is an important characteristics of action research. For example, a teacher of a particular school may face a particular problem in the form of errors committed by fifth grade students in English comprehension in a school but the same problem may not be observed by him/her in other schools.

5) **It enhances the *competencies* of the practitioners**. Action research enables practiceners to have a clear vision of the problematic situation. This becomes helpful in identifying ways and means to tackle the problem.

6) **Action research seeks to *improve the quality of human* *relationships.***

7) **Action research seeks to understand particular complex** social situations whether it is a class, school or community**.**

8) **Action research allows us to identify remedial measures for improvement**. It is specific in nature, i.e. specific to a particular class, school or situation. Therefore, results cannot be generalized.

9) ***It is* a *systematic and scientific process*** but not very rigorous.

10) Action research is *participatory*: it is research through which people work towards the improvement of *their own practices*.

11) Action research makes use of quantitative and qualitative methodologies.

12) Action research is open-minded about what counts as evidence (or data) – it involves not only *keeping records* which describe what is happening as accurately as possible . . . but also *collecting and analysing our own judgements, reactions and impressions* about what is going on.

13) Action research is a *political process* because it involves us in making changes that will affect others.

14) It is a *systematic learning process* wherein people act deliberately, but remaining open to surprises and responsive to opportunities.

15) Action research allows us to build *records* of our improvements and changing *activities and practices.* It records the changes in the *language and discourse* we describe, explain and justify our practices. It also documents changes in the *social relationships and forms of organization* which characterize and constrain our practices.

20.6 MODELS OF ACTION RESEARCH

Numerous authors and researchers have proposed models for undertaking the action research. Because the process involved in action research is somewhat dynamic, various models look a bit different from one another but possess numerous common elements. Starting with a central problem or topic, action research models involve some observation or monitoring of current practice, followed by the collection and synthesis of information and data. Finally, some sort of action is taken, which serves as the basis for next stage of action research. Some models are simple in their design, while others appear relatively complex. Let us discuss some models of action research as an illustration.

20.6.1 Kemmis and McTaggart's Spiral Model

These two authors describe this model as participatory research. In this model, action research involves a spiral of self-reflective cycles of:

- Planning a change.
- Acting and observing the process and consequences of the change.
- Reflecting on these processes and consequences and then re-planning.
- Acting and observing.
- Reflecting, etc.

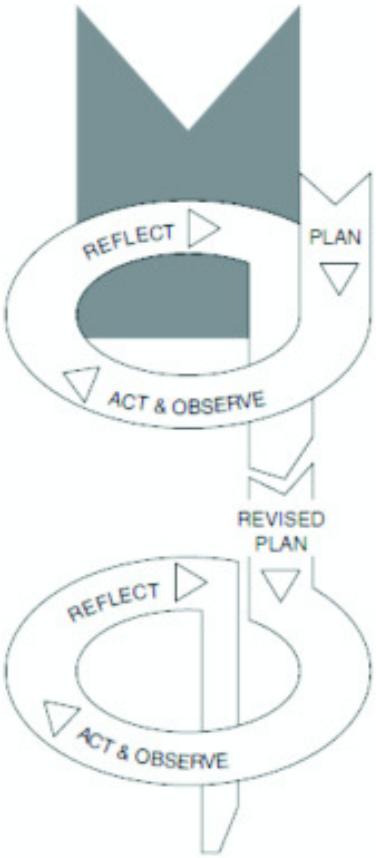


Fig. 20.1: Kemmis and McTaggart's Action Research Spiral

Kemmis and McTaggart's (2000) do not recommend to use it as a rigid structure. They maintain that in reality the process may not be as neat as the spiral of self-contained cycles of planning, acting, observing, and reflecting suggests. These stages, they maintain, will *overlap*, and initial plans will quickly become obsolete in the light of learning from experience. In reality the process is likely to be more fluid, open, and responsive.

We find the spiral model appealing because it gives an opportunity to visit a phenomenon at a higher level each time and so to progress towards a greater overall understanding. By carrying out action research using this model, you can understand a particular issue within a healthcare context and make informed decisions with an enhanced understanding. It is therefore about empowerment. However, Winter and Munn-Giddings (2001) point out that the spiral model may suggest that even the basic process may take a long time to complete.

20.6.2 Elliot's Action Research Model

The model employed by Elliot (1991) shares many of the features of that of Kemmis and McTaggart and is based on Lewin's work of the 1940s. It includes identifying a general idea, reconnaissance or fact-finding, planning, action, evaluation, amending plan and taking second action step, and so on, as can be seen in Figure 20.2.

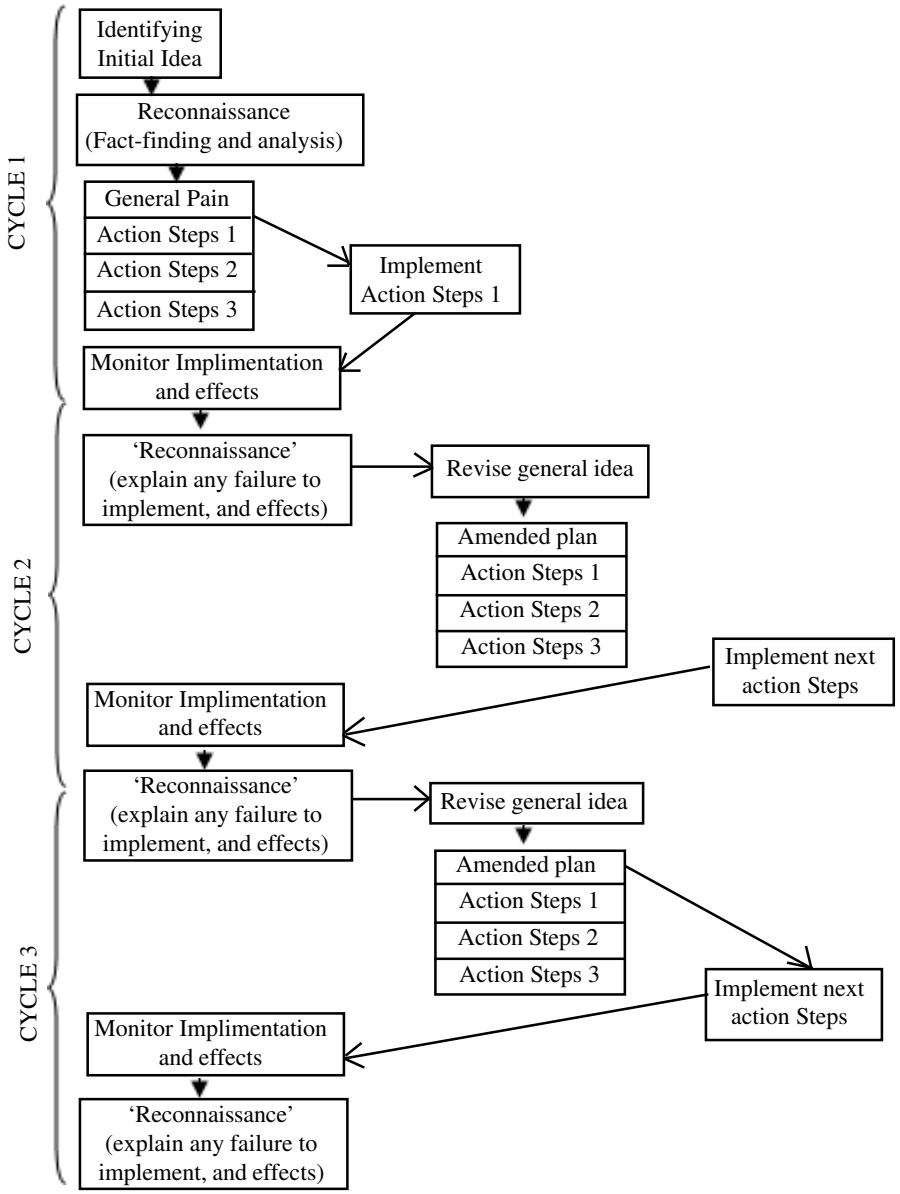


Fig. 20.2: Elliot's Action Research Model

Source: Ellliot, J. Action Research for Educational Change

20.6.3 O'Leary's Cycles of Action Research Model

Cycles of action research shown in Figure 20.3, portray action research as a cyclic process which takes shape as knowledge emerges. In this model, it is stressed that 'cycles converge towards better understanding of situation and improved action implementation. Cycle of actions are based in evaluative practice that alters between action and critical reflection'. O'Leary sees action research as an experiential learning approach to change, where the goal is to continually

refine the methods, data, and interpretation in the light of understanding developed in each earlier cycle.

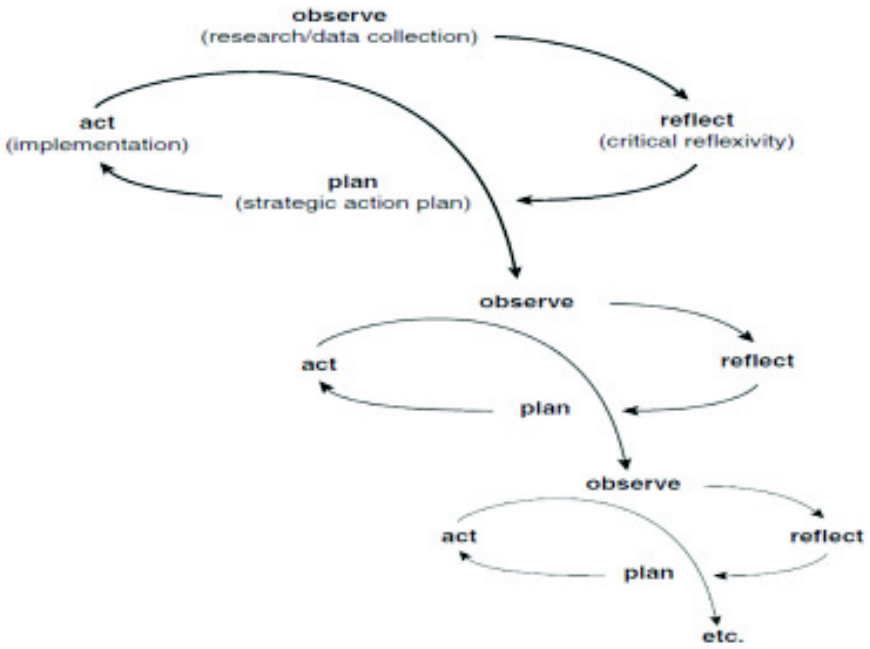


Fig. 20.3: O'Leary's Cycles of Research

20.6.4 Stringer's Interacting Spiral Action Research Model

"Simple, yet powerful framework" consisting of a "look, think, and act" routine. During each stage, participants observe, reflect, and then take some sort of action. This action leads them into the next stage (see Figure 20.4).

Fig. 20.4: Stringer's Interacting Spiral Action Research Model

20.6.5 Kurt Lewin's Action Research Spiral Model

Stringer propounded interacting spiral model and viewed that "action research"—also depicts an action research spiral, which includes fact finding, planning, taking action, evaluating, and amending the plan, before moving into a second action step (see Figure 20.5)

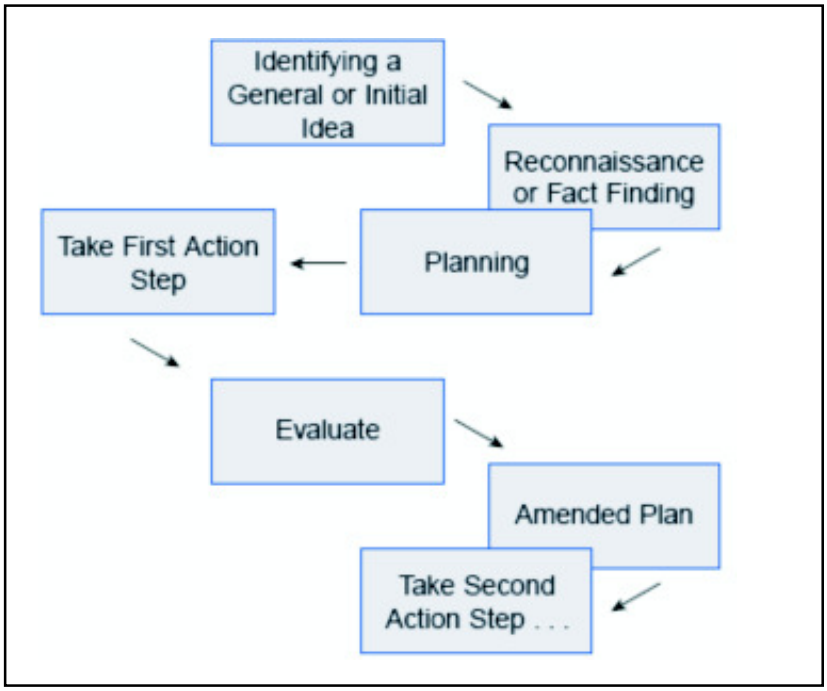


Fig. 20.5: Lewin's Action Research Spiral

Source: Adapted from *Encyclopedia of Informal Education.* (www.infed.org).

20.6.6 Calhoun's Action Research Cycle Model

While not appearing as a "spiral," still represents a process that is built around a cyclical notion. As she describes, the solid lines indicate the primary direction of the action research cycle through the phases in numerical order. The dotted lines indicate backward and forward movement within the cycle as refinement or clarification of information is warranted (see Figure 20.6).

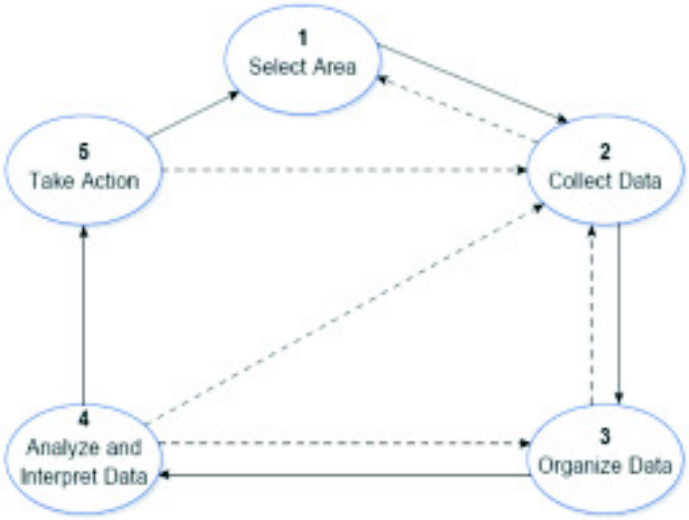


Fig. 20.6: Calhoun's Action Research Cycle

20.6.7 Bachman's Action Research Spiral Model

Action research spiral continues this notion of the cyclical nature of action research (see Figure 20.7). His downward spiral suggests that participants gather information, plan actions, observe and evaluate those actions, and then reflect and plan for a new cycle of the spiral, based on the insights that were gained in the previous cycle.

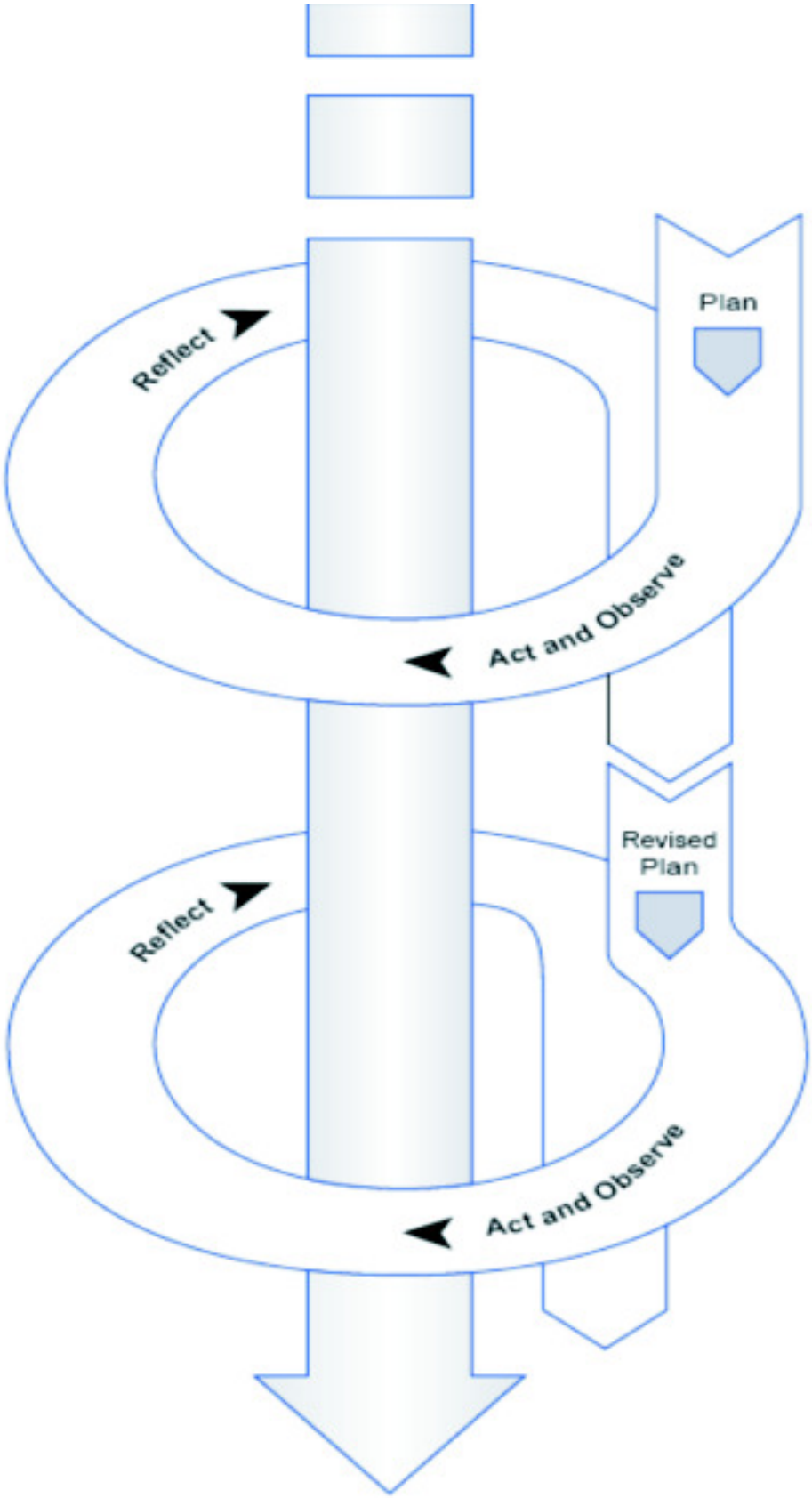


Fig. 20.7: Bachman's Action Research Spiral Model

20.6.8 Riel's Progressive Problem Solving Model

Through action research model takes the participants through four steps in each cycle: planning, taking action, collecting evidence, and reflecting (see Figure 20.8).

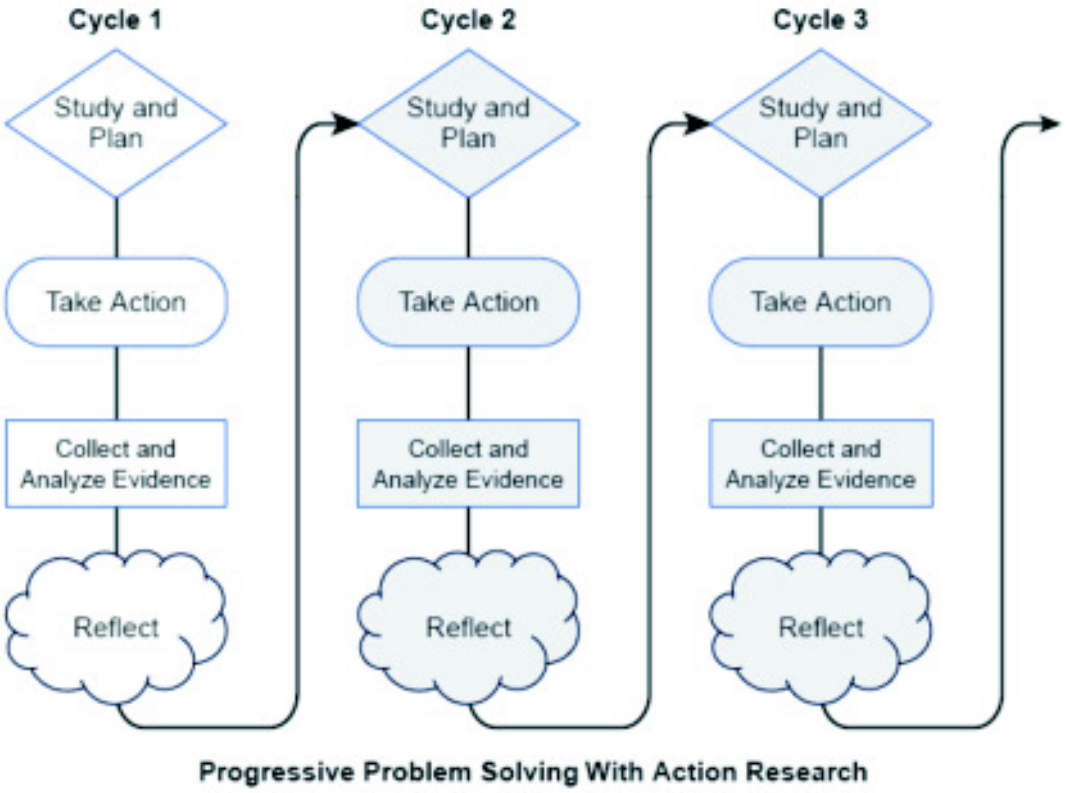


Fig. 20.8: Riel's Action Research Model

20.6.9 Hendricks's Action Research Model

It is shown in Figure 20.9. In her model, which she has placed in a school-based context, focuses on acting, evaluating, and reflecting.

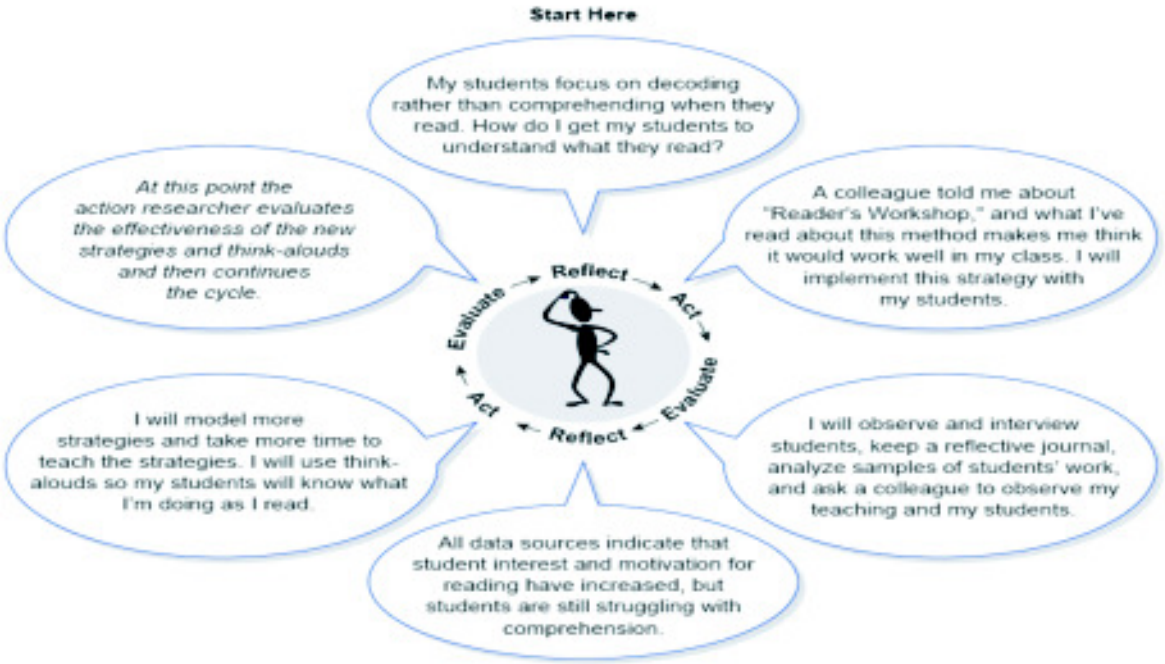


Fig. 20.9: Hendricks's Action Research Process

Source: Adapted from *Improving Schools Through Action Research: A Comprehensive Guide for Educators*, by Cher Hendricks, 2009, Boston: Allyn& Bacon.

From the above different models of action research, it can be inferred that the following four stages are essential features of the model. However, this does not mean that this is how all action research projects will work. The flexibility of action research based on constant evaluation and reflection means that the cycles may be truncated as new ways to proceed further.

i) **Planning**

Identifying the issue to be changed looking elsewhere for information. Similar projects may be useful, for developing the questions and research methods to be used to develop a plan related to the specific environment. In the school setting this could involve personnel, budgets and the use of outside agencies.

ii) **Acting**

Trialling the change following your plan, collecting and compiling evidence, questioning the process and making changes as required.

iii) **Observing**

Analysing the evidence and collating the findings, discussing the findings with co-researchers and /or colleagues for interpretation, writing the report, sharing your findings with stakeholders and peers.

iv) **Reflecting**

Evaluating the first cycle of the process, implementing the findings or new strategy, revisiting the process.

These features might be represented diagrammatically as given below

Fig. 20.10

While working through action research remember that:

- It is cyclical and progress is made in small chunks,
- It is heavily based on critical reflection you can use.

- A wide range of methods for collecting data but it may be advisable to limit these to a manageable number,
- Participants should have meaningful roles in the collection and presentation of data because of the flexibility of the process and the constant reflection,
- Not every cycle will be complete.
- There may be times when it is advisable to stop mid stream and start a new cycle.

Check Your Progress 2

1) What do you understand by dialectical critical analysis?

.....

.....

.....

.....

.....

2) How do multiplicity of views are accommodated in action research?

.....

.....

.....

.....

3) Match the followings:

1) Kemmis and Mc Taggart	a) Cycles of Action Research Model
2) Stringer	b) Action Research Cycle Model
3) Kurt Lewin	c) Spiral Model
4) Rial	d) Intracting Spiral Action Model

4) Enlist the features of flexibility of action research.

.....

.....

.....

.....

20.7 STEPS INVOLVED IN CONDUCTIONING ACTION RESEARCH

Action research is a systematic process for finding the solution of the problem. It can be conducted either individually or in collaboration with others. Different models of action research have envisioned different stages. In general, following steps are involved in conducting action research:

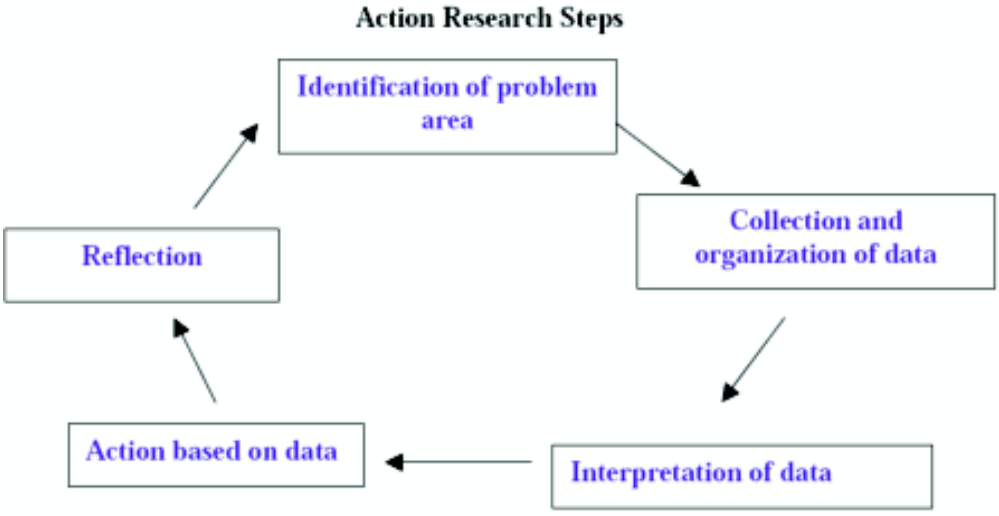


Fig. 20.11

1) **Identification of the Problem Area and Developing a Focus**

Suppose a teacher may have several questions or problems to encounter such as poor reading ability among his/her students, pronunciation problem among students, effective monitoring of the various programs and many more. Therefore, the focus of action research is on what students are experiencing or have experienced? For example, a teacher can study how to improve problem-solving skills in mathematics among the students or increase reading ability among students and so on. It is therefore crucial to choose the problem which is meaningful so that it can be solved in the stipulated time.

The need for action research arises because of perceived dissatisfaction with an existing situation. It is followed with the idea of bringing out improvement in the situation. The focus is on the following: (i) what is the cause of problem? (ii) Why is it happening? (iii) As a practitioner or a researcher, what can I do about it? (iv) Which steps can I take to solve the problem? The answers to all such questions are helpful in perceiving a problem as it exists which is a pre-requisite for undertaking any action research problem.

2) **Formulating the Problem**

Once, the problem is identified, the next step is to formulate it. The researcher tries to find causes underlying the problem along with various issues that are related to causes. These probable causes need to be stated in concise and unambiguous terms.

3) **Stating the Research Questions and Development of Propositions**

After formulating the problem, the researcher needs to state the research questions and develop a tentative theory in the form of propositions keeping in view the genesis of the problem. It is necessary to develop a conceptual and functional relationship, tentatively to understand and explain the given situation.

4) **Data Collection**

The collection of data is the most important step in deciding what action is needed for solving the problem. For example, in the school, there could be

multiple sources of data, which a practitioner can use to identify causes and developing, and implementing remedial measures. These include.

- Videotapes and audio tapes,
- Report cards,
- Attendance, samples of student work, projects, performances,
- Interviews with the parents, students etc.,
- Cumulative records and Anecdotal records,
- School Diaries,
- Photos,
- Questionnaires,
- Focus groups discussions,
- Checklists,
- Observation schedules.

Select the data that are most appropriate for the study. After collecting the data, these are arranged on the basis of gender, classroom, grade level, school, etc. The practitioner may use purposive samples of students or teachers from each grade level in case of larger groups.

5) **Analysis and Interpretation of Data**

After the data has been gathered, the next step is to analyze the data in order to identify trends and themes. The qualitative data obtained can be reviewed to take out the common elements or themes and may be summarized in the suitable table formats. The quantitative data can be analyzed with the use of simple statistics such as percentages, simple frequency tables, or by calculating simple, descriptive statistics.

At this step, the data is turned into information, which can help the practitioner or the researcher in making decisions. Therefore, this stage requires maximum time. After the analysis, it becomes clear what important points do these data reveal and which important patterns or trends are emerging.

6) **Discussion and Evaluating Actions**

After the careful analysis of the data, review of current literature is done for taking decisions and necessary actions. Following points need be kept in mind while conducting the literature review:

i) Identifying topics that relate to the area of the study and would most likely yield useful information.

ii) Gather or collect research reports, research books and videotapes relating to the problem.

iii) Organise these materials for drawing inferences in the light of result of the action research study.

Suggesting a plan of action will allow the practitioner to make a change. This is well informed decision-making. It is advisable to suggest one action at a time and then observe its outcome in improving the situation.

20.8 ADVANTAGES AND DISADVANTAGES OF ACTION RESEARCH

20.8.1 Advantages

1) Action research lends itself to use in work or community situations. Practitioners, people who work as agents of change, can use it *as part of their normal activities.* Mainstream research paradigms in some field situations can be more difficult to use. Action research offers such people a chance to make more use of their practice as a research opportunity.

2) When practitioners use action research, it has the potential to increase the amount they learn consciously from their experience. The action research cycle can also be regarded as a learning cycle. Systematic reflection is an effective way for practitioners to learn.

3) Action research is usually participative. This implies a partnership between you and your clients. You may find this more ethically satisfying. For some purposes it may also be more occupationally relevant.

4) Action research is helpful in determining policy related to the problem.

20.8.2 Disadvantages

1) Excuting an action research project is more difficult than conventional research. Here a researcher takes responsibilities for change as well as for research. In addition, as with other field of research, it involves you as a researcher in more work to set it up, and you don't get any credit for that.

2) It doesn't accord with the expectations of some examiners. Deliberately and for good reasons it ignores some requirements which have become part of the ideology of some conventional research. In that sense, it is counter-cultural.

3) As a researcher, one is not exposed much about action research. Action research methodology is something that a learner has to learn almost from scratch.

4) The library work for action research is more demanding. In conventional research you know ahead of time what literature is relevant. In most forms of action research, the relevant literature is defined by the data you collect and interpret. That means that you begin collecting data first, and then go to the literature to challenge your findings.

5) Action research is more difficult to report, at least for thesis purposes. If you stay close to the research mainstream, you don't have to take the same pains to justify what you do. For action research, you are obliged to justify your overall approach.

Check Your Progress 3

1) What does motivate you to undertake action research?

.....

.....

.....

2) How do research questions enable to formulate the problem?

.....

.....

.....

3) Identify the possible sources of data for an action researcher.

.....

.....

.....

4) What are the advantages of action research over conventional research?

.....

.....

.....

5) Do you think that action research is a tough challenge before a conventional researcher? Give two reasons in support of your answer.

.....

.....

.....

20.9 LET US SUM UP

Action Research is associated with emancipatory method of critical theory paradigm approach wherein knowledge is generated in the process of knowing through doing rather through conceptualization and theorizing. In this process, it becomes difficult to make a distinction between researcher and practitioner as their activities become integrated.

Action research is a powerful tool for change and improvement at the local level. Researcher in action research aims to contribute both to the practical concerns of people in an immediate problematic situation and to further the goals of social sciences simultaneously.

Origin of Action research is traced in the work Kurt Lewin in 1940's. He was strong exponent of research action in its concern with power relations between researcher and researched and the rights of the individuals. Action research is a form of *collective* self-reflective enquiry undertaken by participants in social situations to improve the conditions of researched.

A set of six principles guide action research. These include: reflective critical analysis, dialectical critical analysis, collaborative resource, risk, plural structure, theory, practice and transformation. Different models of action research discussed in this unit are: Kemmis and McTaggart action research model, *Elliot's Action Research Model*: *O'Leary's Cycles of Research (2004),* Stringer's model (2007), Lewin's Action Research Spiral, Calhoun's Action Research Cycle, Bachman's

Action Research Spiral, Riel's Action Research Model and Hendricks's Action Research Process.

The steps followed in action research include: Identification of the Problem Area and Developing a Focus, Formulating the problem, Stating the Research Questions and Development of Propositions, Data Collection, Analysis and Interpretation of Data, Discussions and Evaluating Actions. Action research has several advantages over conventional research. However, it has several limitations as well.

20.10 KEY WORDS

Action Research : *Learning by doing*. A group of people identifies a problem, do something to resolve it, see how successful their effects were and if not satisfied, try again.

Reflective Critical Analysis : It is the process of becoming aware of our own perceptual biases.

Dialectical Critical Analysis : It is a way of understanding the relationships between the elements that make up various phenomena in our context.

Collaborative Resource : Which is intended to mean that everyone's view is taken as a contribution to understand the situation.

Risk : It is an understanding of our own taken-for-granted processes and willingness to submit them to critique.

Plural Structure : It involves developing various accounts and critiques, rather than a single authoritative interpretation.

Collaborative : Everyone's view is taken as a contribution in understanding the situation.

Reconnaissance : Fact finding and explanation.

Planning : Identifying the issue to be changed, developing the questions and research methods to be used developing a plan related to the specific environment.

Acting : Trialling the change following your plan, collecting and compiling evidence.

Observing : Analysing the evidence and collating the findings, discussing the findings with co-researchers and/or colleagues for the interpretation.

Reflecting : Evaluating the first cycle of the process, implementing the findings or new strategy, revisiting the process.

20.11 REFERENCES

Calhoun, E. F., (1994). *How to use Action Research in Self Renewing School*. Alexandria. V A, ASLD.

Carr, W. and Kemmis, S. (1986). *Becoming Critical: Education, Knowledge and Action Research*. London: Falmer.

Ebbutt, D. (1985). *Educational Action Research: Some General Concerns and Specific Quibbles*. In R. Burgess (ed.) Issues in Educational Research: Qualitative Methods. Lewes: Falmer, 152–74.

Elliot, J. (1991). *Action Research for Educational Change*. Buckingham: Open University Press.

Ernest T. Stringer (2007). *Action Research (Third Edition)*. London: Sage Publications, 279pp (pb), ISBN: 978-1-4129-5223-1, $46.95.

G. E. Mills (2011). *Action Research: A Guide for the Teacher-Research (4th ed.)*. Pearson.

Hendricks. C. (2009). *Improving School Through Action Research; A Comprehensive Guide for Educator. (2nd ed.)*. Upper Saddle River, N J: Pearson.

Hopkins, D. (1985). *A Teacher's Guide to Classroom Research, Philadelphia*: Open University Press.

Kemmis, S. and McTaggart, R. (1992). *The Action Research Planner (third edition) Geelong*, Vic.: Deakin University Press.

Kemmis, S. and McTaggart, R. (2000). "*Participatory Action Research*", in N.K. Denzin and Y.S. Lincoln (eds) Handbook of Qualitative Research (2nd ed.). Sage, CA, pp. 567–605.

McNiff, J., and Whitehead, J. (2002). *Action Research: Principles and Practice (second edition)*. London: RoutledgeFalmer.

O'Leary, Z. (2004). *The Essential Guide to Doing Research*. London: Sage Publications.

Richard. Winter and Carol Munn-Giddings(2001). *A Handbook for Action Research in Health and Social Care*. by Routledge, New York NY 10017.

Selener, D. (1997). *Participatory Action Research and Social Change*. New York: Cornell Participatory Action Research Network.

Zuber-Skerritt, O. (ed.) (1996). *New Directions in Action Research, London*; Falmer Press.

20.12 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 20.1
2) See Section 20.1
3) See Section 20.2
4) See Section 20.3

Check Your Progress 2

1) See Sub-section 20.4.2
2) See Sub-section 20.4.5
3) See Section 20.6
4) See Sub-section 20.6.4

Check Your Progress 3

1) See Section 20.7
2) See Section 20.7
3) See Section 20.7
4) See Sub-section 20.8.1
5) See Sub-section 20.8.2

Indira Gandhi
National Open University
School of Social Sciences

MEC-109
Research Methods in Economics

Block

6

DATA BASE OF INDIAN ECONOMY

UNIT 21

Macro-Variable Data: National Income, Saving and Investment 9

UNIT 22

Agricultural and Industrial Data 37

UNIT 23

Trade and Finance 63

UNIT 24

Social Sector 85

Expert Committee

Prof. D.N. Reddy
Rtd. Professor of Economics
University of Hyderabad, Hyderabad

Prof. Harishankar Asthana
Professor of Psychology
Banaras Hindu University
Varanasi

Prof. Chandan Mukherjee
Professor and Dean
School of Development Studies
Ambedkar University, New Delhi

Prof. V.R. Panchmukhi
Former Professor of Economics
Bombay University and Former
Chairman ICSSR, New Delhi

Prof. Achal Kumar Gaur
Professor of Economics
Faculty of Social Sciences
Banaras Hindu University, Varanasi

Prof. P.K. Chaubey
Professor, Indian Institute of
Public Administration, New Delhi

Shri S.S. Suryanarayana
Rtd. Joint Advisor
Planning Commission, New Delhi

Prof. Romar Korea
Professor of Economics
University of Mumbai, Mumbai

Dr. Manish Gupta
Sr. Economist
National Institute of Public Finance and Policy
New Delhi

Prof. Anjila Gupta
Professor of Economics
IGNOU, New Delhi

Prof. Narayan Prasad **(Convenor)**
Professor of Economics
IGNOU, New Delhi

Prof. K. Barik
Professor of Economics
IGNOU, New Delhi

Dr. B.S. Prakash
Associate Professor in Economics
IGNOU, New Delhi

Shri Saugato Sen
Associate Professor in Economics
IGNOU, New Delhi

Course Coordinator and Editor: Prof. Narayan Prasad

Block Preparation Team

Unit	Resource Person	IGNOU Faculty (Format, Language and Content Editing)
21-24	Shri S.S. Suryanarayan Ex Joint Advisor Planning Commission Renamed as Niti Ayog New Delhi Updated by Mr. Purnendu Kishore Banerjee Dy. Registrar General and Census Commissioner New Delhi	Prof. Narayan Prasad Professor of Economics IGNOU, New Delhi

Print Production

Mr. Manjit Singh
Section Officer (Pub.)
SOSS, IGNOU, New Delhi

October, 2015

© Indira Gandhi National Open University, 2015

ISBN-978-81-266-

All rights reserved. No part of this work may be reproduced in any form, by mimeograph or any other means, without permission in writing from the Indira Gandhi National Open University.

Further information on Indira Gandhi National Open University courses may be obtained from the University's office at Maidan Garhi. New Delhi-110 068.

Printed and published on behalf of the Indira Gandhi National Open University, New Delhi by the Director, School of Social Sciences.

Laser Typeset by : Tessa Media & Computers, C-206, A.F.E.-II, Okhla, New Delhi
Printed at :

BLOCK 6 DATA BASE OF INDIAN ECONOMY

For undertaking any meaningful research in terms of situational assessment, testing of models, development of theory, evolving economic policy, assessing the impact of such policy etc., data is crucial. Data on different variables is thus an essential input for assessment and analysis of economic situations. The availability or otherwise of data, therefore, determines the scope of analysis. The reliability of the conclusions arrived at also depends on the availability and veracity of data. Hence, researcher's knowledge about the availability of data is important for conducting a meaningful research. Keeping in view that a student of economics at postgraduate level is expected to know the available databases, the present block deals with the different databases of Indian Economy.

The block comprises of 4 units. **Unit 21** deals with the data available on major macro variables relating to the economy – national income, saving and investment, etc. **Unit 22** throws light on the kind of agricultural and industrial databases. **Unit 23** deals with data on trade and finance.

Unit 24 discusses the availability of data on other social sectors like employment and unemployment, education, health, shelter and amenities, environment, quality of life etc. These four units lay particular emphasis on the different concepts used in data collection including the agencies involved in the compilation of data.

Abbreviations Used in the Block

ADR	-	American Deposit Receipt
AHSD	-	Animal Husbandry Statistics Division
AIES	-	All India Educational Survey
ANM	-	Auxiliary Nurse-Midwife
ASI	-	Annual Survey of Industries
ASDR	-	Age – Specific Death Rate
ASG	-	Agricultural Statistics at a Glance
BAHS	-	Basic Animal Husbandry Statistics
BC	-	Backward Classes
BD	-	Budgetary Deficit
BMI	-	Body Mass Index
BOD	-	Biological Oxygen Demand
BoP	-	Balance of Payments
BR	-	Birth Rate
CACP	-	Commission on Agricultural Costs and Prices
CAD	-	Current Account Deficit
CAS	-	Current Account Surplus
CB	-	Commodity Boards
CBHI	-	Central Bureau of Health Intelligence
CCPS	-	Cumulative Convertible Preference Shares
cfc	-	Consumption of Fixed Capital (depreciation)
CGHS	-	Central Government Health Scheme
CGWB	-	Central Ground Water Board
CHC	-	Community Health Centre
CIF	-	Cost, Insurance and Freight
CIS	-	Commonwealth of Independent States
CMI	-	Census of Manufacturing Industries
CMIE (EIS)	-	Centre for Monitoring Indian Economy (Economic Intelligence Services)
CPI-AL	-	Consumer Price Index – Agricultural labour
CPI-IW	-	Consumer Price Index – Industrial Workers
CPI-UNME	-	Consumer Price Index – Urban Non-Manual Employees
CSO	-	Central Statistical Organisation
CWC	-	Central Water Commission
CWS	-	Central Weekly Status
DCI	-	Dental Council of India
DDP	-	District Domestic Product
DE	-	Directory Establishments
DEs	-	Directorates of Employment
DESMOA	-	Directorate of Economics & Statistics, Ministry of Agriculture

DES(s)	-	Directorate(s) of Economics & Statistics
DGE&T	-	Directorate General of Employment & Training
DIPP	-	Department of Industrial Policy and Promotion
DR	-	Death Rate
EBs	-	Electricity Boards
EMI	-	Employment Market Information
EPC	-	Export Promotion Council
EPFO	-	Employees Provident Fund Organisation
EPW	-	Economic and Political Weekly
EPWRF	-	Economic and Political Weekly Research Foundation
ESI Act	-	Employees' State Insurance Act
ESIC	-	Employees State Insurance Corporation
EXIM Bank	-	Export Import Bank
FDI	-	Foreign Direct Investment
FIs	-	Financial Institutions
FIIs	-	Foreign Institutions Investors
FIPB	-	Foreign Investments Promotion Board
FISIM	-	Financial Intermediation Services Indirectly Measured
f.o.b.	-	free on board
FRG	-	Federal Republic of Germany
FSI	-	Forest Survey of India
FSS	-	Farmers' Service Societies
GDI	-	Gender Development Index
GDCF	-	Gross Domestic Capital Formation
GDFCF	-	Gross Domestic Fixed Capital Formation
GDI	-	Gender Development Index
GDP	-	Gross Domestic Product
GDR	-	Global Deposit Receipt
GDS	-	Gross Domestic Saving
GFCE	-	Government Final Consumption Expenditure
GFD	-	Gross Fiscal Deficit
GNP	-	Gross Primary Deficit
GPD	-	Gross Primary Deficit
GSDP	-	Gross State Domestic Product
GTT	-	Gross Terms of Trade
HA	-	Health Assistant
HDFC	-	Housing Development Finance Corporation
HDI	-	Human Development Index
IAMR	-	Institute of Applied Manpower Research
IBM	-	Indian Bureau of Mines
ICICI	-	Industrial Credit and Investment Bank of India
ICSSR	-	Indian Council of Social Science Research

IDBI	-	Industrial Development Bank of India
IIP	-	Index of Industrial Production
IIPS	-	International Institute for Population Sciences
IMF	-	International Monetary Fund
IMR	-	Infant Mortality Rate
I-O TT	-	Input Output Transaction Table
IPO	-	Initial Public Offering
IPP	-	Index of Prices Paid
IPR	-	Index of Prices Received
ISCED	-	International Standard Classification of Education
ISCO	-	International Standard Classification of Occupations
ISIC	-	International Standard Industrial Classification
ISIC – Rev. 3	-	International Standard Industrial Classification – Revision 3
ISM	-	Indian Systems of Medicine
ITC (HS)	-	Indian Trade Classification (based on Harmonised Commodity Description and Coding System)
ITC – Rev. 2	-	Indian Trade Classification – Revision 2
ITES	-	Information Technology Enabled Services
ITT	-	Income Terms of Trade
KVIC	-	*Khadi* and Village Industries Commission
LAMPS	-	Large – sized *Adivasi* Multi-Purpose Societies
LHV	-	Lady Health Visitor
LIC	-	Life Insurance Corporation of India
MCI	-	Medical Council of India
MF	-	Mutual Fund
MHA	-	Male Health Assistant
MHRD	-	Ministry of Human Resource Development
MI	-	Minor Irrigation
MMR	-	Maternal Mortality Rate
MOEF	-	Ministry of Environment and Forests
MoLE	-	Ministry of Labour & Employment
MoSPI	-	Ministry of Statistics & Programme Implementation
MPCE	-	Monthly Percapita Consumption Expenditure
MSP	-	Minimum Support Price
NABARD	-	National Bank for Agricultural and Rural Development
NABS	-	National Advisory Board on Statistics
NAS	-	National Accounts Statistics
NBFC	-	Non-Banking Finance Companies
NBNFC	-	Non-Banking Non-Finance Companies
NCERT	-	National Council for Educational Research and Training

NCO	-	National Classification of Occupations
NCI	-	Nursing Council of India
NCS	-	Net Capital Stock
NDC	-	National Development Council
NDE	-	Non-Directory Establishments
NDCF	-	Net Domestic Capital Formation
NDFCF	-	Net Domestic Fixed Capital Formation
NDP	-	Net Domestic Product
NDS	-	Net Domestic Saving
NEER	-	Nominal Effective Exchange Rate
NFCS	-	Net Fixed Capital Stock
NFD	-	Net Fiscal Deficit
NFHS	-	National Family Health Survey
NFE	-	Non-Formal Education
NHB	-	National Housing Bank
NIC	-	National Industrial Classification
NNP	-	Net National Product
NNR	-	Nano-Natal Rate
NPA	-	Non-Performing Assets
NPD	-	Net Primary Deficit
NPISH	-	Non-Profit Institutions Serving Households
NRI	-	Non-Resident Indian
NSDP	-	Net State Domestic Product
NSE	-	National Stock Exchange
NSRC	-	National Statistical Commission Report
NSSO	-	National Sample Survey Organisation
NTMIS	-	National Technical Manpower Information System
NTT	-	Net Terms of Trade
OAE	-	Own Account Enterprises
OECD	-	Organisation for Economic Co-operation and Development
OPEC	-	Oil and Petroleum Exporting Countries
PACS	-	Primary Agricultural Credit Society
PCI	-	Pharmacy Council of India
PFCE	-	Private Final Consumption Expenditure
PHC	-	Public Health Centre
PHSC	-	Public Health Sub-Centre
PISS	-	Priced Information Service Systems
PLR	-	Prime Lending Rate
PNR	-	Post – Natal Rate
PPF	-	Public Provident Fund
PRD	-	Primary Revenue Deficit

QI	-	Quantum Index
RAS	-	Regional Accounts Statistics
RBI	-	Reserve Bank of India
REC	-	Rural Electrification Corporation
REER	-	Real Effective Exchange Rate
RD	-	Revenue Deficit
RGI	-	Registrar General of India
RRBs	-	Regional Rural Banks
RS	-	Remote Sensing
SAS	-	State Accounts Statistics
SCBs	-	Scheduled Commercial Banks
SBR	-	Still Birth Rate
SC & ST	-	Scheduled Casts & Scheduled Tribes
SDDS	-	Special Data Dissemination Standards
SDP	-	State Domestic Product
SDR	-	Special Drawing Rights
SEBI	-	Securities Exchange Board of India
SIA	-	Secretariat for Industrial Approvals
SIC	-	Standard Industrial Classification
SIDBI	-	Small Industries Development Bank of India
SITC – Rev. 2	-	Standard International Trade Classification – Revision 2
SLR	-	Statutory Liquidity Ratio
SNA	-	System of National Accounts
SPM	-	Solid Particulate Matter
SRS	-	Sample Registration Scheme
SSMI	-	Sample Survey of Manufacturing Industries
UGC	-	University Grants Commission
UN	-	United Nations
UPS	-	Usual Principal Status
UPSS	-	Usual Principal and Subsidiary Status
UTI	-	Unit Trust of India
UVI	-	Unit Value Index
VRC	-	Vocational Rehabilitation Centre
WC Act	-	Workmen's Compensation Act
WPI	-	Wholesale Price Index
WPR	-	Worker-Population Ratio
WTO	-	World Trade Organisation

UNIT 21 MACRO-VARIABLE DATA: NATIONAL INCOME SAVING AND INVESTMENT

Structure

21.0 Objectives

21.1 Introduction

21.2 The Indian Statistical System

21.3 National Income and Related Macro Economic Aggregates

- 21.3.1 System of National Accounts (SNA)
- 21.3.2 Estimates of National Income and Related Macroeconomic Aggregates
- 21.3.3 The Input-Output Table
- 21.3.4 Regional Accounts – Estimates of State Income and Related Aggregates
- 21.3.5 Regional Accounts – Estimates of District Income
- 21.3.6 National Income and Levels of Living

21.4 Saving

21.5 Investment

21.6 Let Us Sum Up

21.7 Exercises

21.8 Some Useful Books

21.9 Answers or Hints to Check Your Progress Exercises

21.0 OBJECTIVES

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

- know the various approaches followed by the Indian statistical system in generating data on various economic and social phenomena;
- explain the SNA (System of National Accounts) methodology of data compilation;
- identify the various aggregates on which estimates are made by the National Accounts System;
- compare the interrelationship among different sectors of the economy through input-output transaction table (I-OTT) compiled by CSO; and
- describe the data on National Income, Saving, Investment and other related macro economic variable and the estimates of the state income which enable analysis of various aspects of economy.

21.1 INTRODUCTION

You would soon be a professional economist and would possibly be engaged in research or would function as an economic analyst or adviser in a Government or non-Government organization. As a professional economist, you will have to analyse situations and arrive at conclusions about them or develop solutions

for further action to tackle such situations. You will, for this purpose, identify the variables that, in your opinion, shape or affect the problem at hand, incorporate these in a suitably chosen model, test the efficacy of the model in reflecting the situation under examination and proceed further with the analysis. A basic input for such activity is data on the variables incorporated in the model. The availability of data ***relevant*** to the proposed analysis enhances or restricts the scope of your analysis and the reliability of the conclusions arrived at on the basis of the analysis. Indeed, it might even restrict the kind of variables and model you would like to use in your analysis, unless you want to conduct a survey yourself to collect the data that are not available. Your technical and professional equipment as an economist is, therefore, incomplete without knowledge of where data on the variables in question are available and how good they are for the purpose of the analysis at hand. This knowledge should naturally cover the entire data base of the Indian economy since an economist's work can cover any part or the whole of the Indian economy.

We shall look at the data base of the Indian economy in this Block. This Unit deals with data available on National Income, Saving and Investment – the major macro variables relating to the economy and associated macroeconomic aggregates. Agricultural and Industrial data, data relating to the Social Sector, Trade and Finance will be discussed in the subsequent units of this block.

Let us begin by taking a look at the Indian statistical system.

21.2 THE INDIAN STATISTICAL SYSTEM

The Indian Statistical System generates data on a variety of economic and social phenomena, essentially through six approaches. First, Central Acts like the Census Act, 1948; and the Collection of Statistics Act, 2008, etc. enable the Government agencies to conduct large-scale sample surveys for collection of data at regular intervals. Second, statutory returns prescribed under several other Acts like the Factories Act, 1948 the Companies Act, 1956, the Reserve Bank of India Act, 1953; the Registration of Births and Deaths Act, 1969 and the implementation of these Acts generate data on matters not covered by the surveys. Third, data collected by individual Ministries, Departments and organizations of the Central and State Governments as part of their specific functions reflect the emerging situation in different sectors and sub-sectors of the economy and administrative divisions of the country. Fourth, the administrative reports of these organizations supplement such data. Next, information derived from the data flows mentioned above, like the **National Accounts Statistics (NAS)**, index numbers of prices and production and indices like the Human Development Index and Gender Development Index provide readily usable inputs for research and policy and evaluation of the impact of policies and programmes of development in terms of the health of the economy and the well-being of the society. Lastly, a large number of surveys and research studies conducted by various institutions, public and private, on a variety of subjects constitute another flow of data and information.

The Ministry of Statistics and Programme Implementation of the Government of India is the apex body in the official statistical system of the country. The Ministry is headed by the Chief Statistician of the country. The Ministry consists of the Central Statistics Office (CSO), the National Sample Survey Office (NSSO) and the Computer Centre (CC). The Directorates of Economics and Statistics function at the level of the State/UT Governments. The CSO

coordinates the statistical activities in the country, lays down and maintains statistical norms and standards and provides liaison with Central, State and International statistical agencies. There is also a National Statistical Commission (NSC), comprising of Chairman and Members, who are eminent economists and statisticians from research institutions, representatives of Central Ministries and Departments and the State Directorates of Economics and Statistics, to provide (i) guidance for an overall perspective for statistical development in the country, (ii) guidance to Government on policy issues, and (iii) ensure effective coordination of all statistical activities of the Government of India. CSO is the Secretariat of the NSC.

The CSO, as the Central Statistical Authority, is responsible for coordination of statistical activities in the country and for evolving and maintaining statistical standards. Mention should be made in this connection of three publications of CSO, namely, the Sources and Methods of National Accounts Statistics, the National Industrial Classification and the Consumer Price Index. The first two are *adhoc* while the third a monthly one. The first one covers the methods and classification principles to be followed for preparation of the macro economic aggregates, the second one gives the classification principle to be followed for national and international comparability and the third one provides State/UT wise price indices.

Certain other agencies concerned with economic development also bring out publications that cover data relating to most of all the sectors of the economy at one place. These are **the Reserve Bank of India (RBI) Bulletin (monthly), the Financial Stability Report (including Trend and Progress of Banking in India), RBI's Handbook of Statistics on the Indian Economy** and **the website of the RBI**, the pre-Budget **Economic Survey** and the **Budget presented to Parliament** by the Finance Minister every year. **The Five-Year Plan documents** of the **Planning Commission of the Government of India used to** make up yet another set of data sources that provides important data across sectors. **The Human Development Report** was prepared by the Planning Commission **for the first time in the year 2001. This and similar reports prepared subsequently by many of the State Governments** for individual States and for district constituted another set of sources for data across sectors at the national, State, district and even sub-district levels. The **District Census Handbooks** published after successive population censuses by the Office of the **Registrar General of India (ORGI) and the Directorates of Census of different States and Union Territories** provide comprehensive information across sectors on individual districts, sub-districts, towns and villages. The ORGI, in the year 2014 has also launched a digital library of all census tables published since Census 1991 in their website www.censusindia.gov.in. These tables together provide a very useful time series data on demographic characteristics of the country and its States/ UTs. Most of the data and reports published by Government agencies are available in electronic format, that is, in DVDs/CDs and are accessible at the websites of the agencies concerned. In the year 2013, the Planning Commission has also set up a central web-based repository, namely, www.data.gov.in to bring published data of different Government agencies under one roof.

The Indian statistical system has also recently been reviewed so as to bring about improvements in it. A National Statistical Commission was appointed by the Government of India to examine critically the deficiencies of the existing statistical system from the point of view of timely availability, reliability and adequacy of data and to recommend measures to correct these deficiencies and

revamp the statistical system to generate reliable statistics for the purpose of policy and planning at different levels of Government. The Commission submitted its report to Government in September, 2001. **The Report of National Statistical Commission** (2001), published by the **Ministry of Statistics and Programme Implementation (MoSPI)**, is itself a useful reference book on the database of the Indian economy. This can be viewed or downloaded from the website **www.mospi.nic.in/nscr/hp.htm**. Most of data available in Government websites can be downloaded free of cost, either directly or as a registered user. In most cases, this registration is free of charge. For example, one can make a free registration in the website of the MoSPI and with the username, freely download all the reports and tables published by different offices under the MoSPI. For the Office of the RGI, one can simply visit its website www.censusindia.gov.in click on the link digital library and download all the census tables published from 1991 census onwards.

There are also non-Government sources that publish these for the use of researchers and other data users (Certain private websites like **www.indiastat. com** provide India centric, sector specific and state specific data for the research fraternity, usually for a fee). The **Centre for Monitoring Indian Economy (CMIE), Mumbai** and its **Economic Intelligence Services (EIS)** provide detailed and up-to-date information on the Indian economy through a **Monthly Review of Indian Economy** and annual documents each covering a specific sector or subject in great detail. These publications bring together at one place not only data available with Government statistical agencies but also data collected by CMIE itself. The **Economic and Political Weekly Research Foundation (EPWRF), Mumbai** is another organization that publishes detailed time series data on a variety of subjects along with a description of related conceptual and methodological issues and quality of data, besides presenting statistics on specific areas in **subject-specific Special Issues of the Economic and Political Weekly**. The search for data by an analyst like you can well start from one or more of the comprehensive sources enumerated above and move on, as necessary, to the primary agency concerned with the specific area of your interest.

Data on any variable have to be classified suitably for facilitating analysis. And the classification system needs to be standardized when data on any variable are collected by several agencies and utilized by many in order to ensure meaningful collection of data and comparability of data across space and time. At the international level, different agencies of the United Nations, like, the United Nations Statistics Division (UNSD), the OECD, the International Labour Office (ILO), etc. prepares these classifications for cross-country comparison. For individual countries, these classification systems are recommended. The countries are advised to maintain full compatibility up to a certain level and then make additions/ alterations to suit the country specific needs. In India, usually, such classifications are expanded to suit the need of the country, while maintaining international comparability. Two important classificatory systems developed for use may be referred to here, as its utility cuts across sectors. One is the National Industrial Classification (NIC), developed by the Central Statistics Office (CSO) in consultation with all concerned. It groups the entire spectrum of economic activities by means of a five-digit code structure. NIC has been revised from time to time in keeping with changes in the structure and variety of economic activities over the years. A Standard Industrial Classification, 1962 (SIC 62) developed by CSO was being used in the Sixties until NIC 70 was introduced. This was replaced by NIC 1987, which in turn was replaced by NIC 98 in December 1998. The most

recent revision of the NIC series is NIC 2008. All the NIC classifications are based on the relevant UN International Standard Industrial Classification (ISIC). NIC 2008 was, for instance, based on ISIC- Revision 4. NIC 2008 has been developed and released for use in the place of NIC 04 in September, 2008. NIC 2008 classifies the whole range of economic activity into 21 Sections, 88 Divisions, 238 Groups, 403 Classes and 1,304 Sub-classes. NIC 2008 is comparable completely with ISIC-Rev.4 up to the four-digit level. The fifth digit (sub-class) inside each of the class (4^{th}digit) takes care of the special features of the Indian economy. NIC is used for classifying data on variables like output, employment and income by economic activity. In each release of the NIC, concordance tables are made available to facilitate recasting of past data as per the new classification, to the extent possible.

The other classificatory system is the **National Classification of Occupations (NCO)**, developed by the **Directorate General of Employment & Training (DGE&T) of the Ministry of Labour & Employment**. **NCO 1958**, a five-digit code structure, was developed in the Fifties. This was revised to **NCO 1968**. This was in use till recently. This has now been replaced by **NCO 2004 (NCO04)**. The Indian NCOs have been based on the international occupational classification system developed by the International Labour Organisation (ILO) and revised from time to time. NCO 1968 was patterned on the **International Standard Classification of Occupations 1966 – ISCO 66 – of the ILO**. While revising ISCO 66 to prepare **ISCO 88**, ILO incorporated the concept of skill for effective performance in the relevant occupation, utilizing for the purpose the **International Standard Classification of Education (ISCED)**. **NCO 04 has followed the approach adopted in ISCO 88**. It takes note of the skill required for satisfactory performance in the relevant occupations and has a six-digit code structure. NCO 04 has 8 occupational divisions (the first digit of the code), 95 occupational groups (the second digit), 462 occupational families (the third digit) and 2484 occupations (the fifth and the sixth digits). NCO is usuful for classification of the job seekers, the jobs and employment, thereby facilitating analysis of trends in the labour market and in the employment potential of the economy. Past data can be recast as per NCO 04 with the help of **concordance tables provided in NCO 04**.

A word of caution is necessary while using data assembled from sources other than one's own. Data extracted from printed publications or websites have to be checked for printing errors/errors that occur while data are being posted on the websites. A check of the row and column totals of the tables and other identity relations relevant to the data under reference in the tables should take care of such errors. More importantly, you should also look into aspects of the data, like the concepts, definitions and methodology adopted by the source-agency for collecting and compiling the data, coverage of data (like, whether data was collected from entire country or some specific areas, whether the sample drawn was based on a statistically reliable random sample or it was a purposive sample of a few selected locations, etc.), the methods used for collecting data, reliability and so on periodicity and timeliness in the availability of data, integrity (confidentiality), the data are firmly moored to what the data are actually supposed to reflect. These aspects are collectively referred to as "**metadata"**. All these details are generally available in the publication containing the data or in a related publication of the agency publishing the data or in the relevant website of the agency.

It would sometimes be useful and instructive to judge the Indian situation in an international setting. Publications of the United Nations and its agencies as also

agencies like the International Monetary Fund (IMF) and the World Bank and regional agencies such as the Organisation for Economic Cooperation and Development (OECD)[1] provide data on different facets of the economies of the member countries including India along with information regarding comparability of country-wise data. The IMF has formulated a "**Special Data Dissemination Standards**" **(SDDS)** covering the real sector (national accounts, production index, price indices, etc.), Fiscal Sector, Financial Sector, External Sector and Socio-demographic data for different countries to facilitate transparency in the compilation and dissemination of data on important aspects of the economy of individual countries and cross-country comparison of such data. Countries like India who have accepted the SDDS provide to **IMF** a **National Summary Data Page** in respect of each of the areas and sub-areas listed in SDDS and metadata relating to such data, as per a Dissemination Format prescribed in SDDS. In addition, individual country agencies concerned disseminate an advance release calendar, which gives notice of the precise dates of release of data three months ahead of the date of release of the data, on the internet of the IMF's Data Dissemination Bulletin Board (DSBB). Such information, as provided by any country covered by SDDS, can be accessed on the internet with the help of the Google search engine using the search parameter Special Data Dissemination Standards IMF. You can also visit the site of the United Nation's Statistical Division databases at http://unstats.un.org/unsd/databases.htm to access data compiled by different UN agencies on a host of economic and social parameters.

Your capacity to utilize the vast amount of data available in the Indian Statistical System and examine critically the state of the society can be enhanced considerably by an intelligent combination of data across data domains. You have already come across some examples of such efforts – poverty ratios, concentration ratios for measuring the level of inequality, employment elasticity, labour productivity and capital productivity. (Unit 4 on Poverty and Inequality – Policy Implications and Unit 5 on Employment and Unemployment – Policy Implications in Block 1 of MEC-005 on Indian Economic Policy in your first year Course). Other examples are the Human Development Index and Gender Development Index (Units 3 and 4 in the same Block of MEC-005).

Check Your Progress 1

1) State the major functions of CSO.

……

……

……

……

……

……

……

……

[1] India is not a member of OECD, which is an organization in which several countries in Europe are members.

2) Name any three web sources where you can get data for your research, free of cost.

……

……

……

……

……

3) Which precautions should be taken care of while using data assembled from various sources?

……

……

……

……

……

4) What do you understand by the term Special Data Dissemination Standard?

……

……

……

……

……

Let us now turn our attention to data on macroeconomic variables like national income, saving and investment.

21.3 NATIONAL INCOME AND RELATED MACRO ECONOMIC AGGREGATES

21.3.1 System of National Accounts (SNA)

How do you assess the performance of an economy? You may be able to look at the trend in the production of rice or wheat to be able to say something about the performance of paddy or wheat crop. You can make a similar assessment about the production of steel. If you want to say something on agricultural production, where you find that some crops have done well and others have not, you think of making an overall assessment of agricultural performance by constructing an index of agricultural production to review the performance of agricultural production. But we should like to go beyond levels of output or production and look at performance in terms of incomes flowing from output in the form of rent, wages, interest and profit to those participating in the creation of output namely the factors of production – land, labour, capital and entrepreneurship (Since 1993, the role of the Government in influencing the economy, through various taxes and subsidies have also become an important item of analysis). Alternatively, we would like to base our judgment of

performance on value addition made by the production system namely, value of output net of the (intermediate) costs incurred in creating the output. It is (i) this overall value addition computed for all sectors/activities of the economy, that is referred to as the National Product, (ii) macro-aggregates related to it, and (iii) trends in (i) and (ii), that can help you in analyzing the performance of an economy.

As you know, National Income (NI) is the Net National Income (NNI). It is also used to refer to the group of macroeconomic aggregates like Gross National Income (GNI), Gross Domestic Product (GDP), Gross Value Added (GVA) and Net Value Added (NVA). All these of course refer to the total value (in the sense mentioned above) of the goods and services produced during a period of time, the only differences between these aggregates being depreciation and/or net factor income from abroad. There are other macroeconomic aggregates related to these that are of importance in relation to an economy. What data would you, as an analyst, like to have about the health of an economy? Besides a measure of the National Income every year or at smaller intervals of time, you would like to know how fast it is growing over time. What are the shares of the national income that flow to labour and other factors of production? How much of the national income goes to current consumption, how much to saving and how much to building up the capital needed to facilitate future economic growth? What is the role of the different sectors and economic activities – in the public and private sectors or in the organized and unorganized activities or the households in the processes that lead to economic growth? How does the level and pattern of economic growth affect or benefit different sections of society? How much money remains in the hands of the households for consumption and saving after they have paid their taxes (Personal Disposable Income) – an important indicator of the economic health of households? What is the contribution of different institutions to saving? How is capital formation financed? Such a list of requirements of data for analyzing trends in the magnitude and quality of, and also the prospects of, efforts for economic expansion being mounted by a nation can be very long. Such data, that is, estimates of national income and related macroeconomic aggregates from part of a system of National Accounts that gives a comprehensive view of the internal and external transactions of an economy over a period, say, a financial year and the interrelationships among the macroeconomic aggregates. National Accounts thus constitute an important tool of analysis for judging the performance of an economy vis-à-vis the aims of economic and development policy.

The United Nations (UN) has been recommending guidelines in the form of a System of National Accounts (SNA) in order to promote international standards for the compilation of national accounts as an analytical tool and international reporting of comparable national accounting data. The first version of SNA was recommended in 1953. This was revised in 1968 and then in 1993. The fourth version (2008 SNA) was prepared under the auspices of the Inter-Secretariat Working Group on National Accounts consisting of the Commission of the European Communities (Eurostat), International Monetary Fund (IMF), Organisation for Economic Cooperation and Development (OECD), the UN and the World Bank and adopted by the UN Statistical Commission in 2008. This system has been harmonized with other major statistical systems like the **Balance of Payments Statistics and Government Finance Statistics of the IMF**. The **2008 SNA** contains a coherent, consistent and integrated set of macroeconomic accounts based on a set of internationally agreed concepts, definitions, classifications and accounting rules. It provides a

comprehensive accounting framework within which data can be compiled and presented in a format that is designed to facilitate economic analysis, formulation of policy and decision-making. You can read the entire SNA by accessing it at http://unstats.un.org/unsd/nationalaccount/docs/SNA2008.pdf.

The informal sector or the unorganized sector occupies an important place in our economy, in terms of contribution to national income, employment and exports. It is necessary, therefore, to bestow adequate attention to unorganized sector in the compilation of National Accounts and in other data domains and in international efforts at promoting collection of data on informal sector activities. The 15th International Conference of Labour Statisticians (ICLS) (January, 1993) adopted a resolution on informal sector statistics with a view to helping member countries of the International Labour Organisation (ILO) in reporting comparable statistics of employment in the informal sector. The resolution was endorsed by the UN Statistical Commission and by 1993 SNA.

National Accounts are compiled in India using a mix of 1968 SNA and 1993 SNA. India is also in the process of moving towards the full implementation of SNA methodology to the extent feasible.

21.3.2 Estimates of National Income and Related Macroeconomic Aggregates

a) **Estimates Released by CSO**

The Central Statistics Office (CSO) in the Ministry of Statistics and Programme Implementation (MoSPI), Government of India compile and publish National Accounts, which include estimates of National Income and related macroeconomic aggregates like NNP, GNP, GDP and NDP, PFCE (private final consumption expenditure), saving, capital formation and so on for the country and for the public sector for every financial year. Quarterly estimates of GDP are also prepared and released. Estimates prepared for any year at the prices prevailing in that year are called estimates *at current prices*. Estimates of national income and other aggregates at current prices is generally not used in analyzing changes in the magnitude of these aggregates over time, as price levels over time gets affected by inflation or deflation of an economy. Suppose we need to compare the performance of the economy in terms of national income or other macroeconomic aggregates, over a period of time, say, five years. A comparison of estimates of, say, national income at *current prices* in the opening year and the final year of the five-year period will give us the increase or decrease, as the case may be, in national income at current prices. We can easily note that the quantum of change observed in national income *is the composite measure of the change in national income and the changes in the prices of goods and services between the two points of time*. How then to get at the actual change, or the change in real terms in national income (or any other macroeconomic aggregate) over the period? This is done by removing the effect of changes in prices while comparing the aggregate in question at different points of time. Estimates of the aggregate for different years are, therefore, prepared at the estimates of the aggregate at constant (base year) prices. The comparison of estimates of the aggregate at constant (base year) prices at different points of time is comparison of the magnitude of the aggregate in real terms and measures the real change over time in the magnitude of the aggregate. The base year now in use is 2011-12.

CSO changes the base year from time to time in order to take into account the structural changes that take place in the economy and depict a true picture of the economy. It was 1948-49, initially, when estimates of National Income were first prepared by CSO and published in 1956. It was shifted to 1960-61 in August, 1967, to 1970-71 in January, 1978, to 1980-81 in February,1988, to 1993-94 in February,1999 to 1999-2000 in January, 2006; to 2004-05 in January 2010and to 2011-12 in January 2015. Note that, in the beginning, the base years were the years in which the decennial population census was conducted, that is, the first year of every decade from 1961 to 1981. The choice of the subsequent base years, namely, 1993-94, 1999-00 and 2004-05 were a departure from this practice. What are the reasons for this? Estimates of workforce participation rates (WPR), that is, the proportion of workers to population for the benchmark years play an important role in the compilation of estimates of national income. CSO had been making use of data on WPR available from decennial population censuses till 1981 while compiling national income estimates and for this reason the base year was the relevant decennial census year. It was, however, observed that the Employment and Unemployment Surveys of the National Sample Survey Organisation (NSSO) captured WPR better than the population censuses. The base years chosen after 1981 were thus 1993-94, 1999-2000 and 2004-05, making use of data on employment based on the NSSO 50th, 55th, 61stRound and 68th round Surveys on Employment and Unemployment[2] conducted during July '93 – June '94, July '99 – June 2000, July '03 – June '04 and July 2011- June 2012 respectively. After every revision of base year, the CSO publishes a set of results which provide comparable estimates of the years prior to the base year, so that a long time series data on national income aggregates can be analysed by the researchers and policy makers.

Base year revisions differ from annual revisions in National Accounts primarily because of nature of changes. In annual revisions, changes are made only on the basis of updated data becoming available without making any changes in the conceptual framework or using any new data source, to ensure strict comparison over years. In case of base year revisions, apart from a shift in the reference year for measuring the real growth, conceptual changes, as recommended by the international guidelines, are incorporated. Further, statistical changes like revisions in the methodology of compilation, adoption of latest classification systems, and, inclusion of new and recent data sources are also made. Changes are also made in the presentation of estimates to improve ease of understanding for analysis and facilitate international comparability.

Estimates of national accounts aggregates for different years are published in considerable detail in CSO's Annual Publication "National Accounts Statistics (NAS)". The latest available NAS estimates are detailed in the publication NAS 2014 and the Press Notes of MoSPI dated the 31st January, 2015[3]. The prerss note of January 2015 provides, for the first time, estimates of macro-economic aggregates using the base year 2011-12. Note that in this release, for the first time, Gross Domestic Product (GDP) "at factor cost" has been discontinued. As is the practice internationally, industry-wise estimates have been presented as Gross Value Added (GVA) at basic prices, while "GDP at market prices has been referred to as GDP. Estimates of GVA at factor cost

[2] Some employment data from Census 1991/2001 have also been used to supplement the data from NSSO.

[3] This publication and these Press Notes can also be accessed in the **Ministry of Statistics and Programme Implementation website** *www.mospi.nic.in*

(earlier called "GDP at factor cost") can be compiled by using the estimates of GVA at basic prices and production taxes less subsidies. These have been given in Statement 3.1 of this note. For the years 2011-12, 2012-13 and 2013-14, GVA at factor cost have been compiled and are presented in Statements 10.1 & 10.2 of the press release.[4]Note that GVA at basic prices = GVA at factor cost + net taxes on production and GDP at market prices = GVA at basic prices + net taxes on products.

CSO releases, every January, "first revised estimates", earlier called the "Quick Estimates" of GDP, National Income, per capita National Income, Private Final Consumption Expenditure, Saving and Capital Formation by broad economic sectors[5] for the financial year that ended in March of the preceding year. Quick Estimates for any financial year are thus available with a time lag of ten months. It also releases, along with Quick Estimates for any financial year, revised estimates of national accounts aggregates for earlier financial years. Further, CSO brings out **Advance Estimates of GDP, GNI**, **NNI and per capita NNI**. **The GDP**, **from the base year 2011-12**, **is published at market prices**, **as per the latest international norms**. **Previously, it was published at factor cost**. Advance estimates of the GDPfor the current financial year is published in February, two months before the close of the financial year. These advance estimates are revised thereafter and the undated advance estimates are released by the end of June, three months after the close of the financial year. Meanwhile, by the end of the March, Quarterly Estimates of GDP for the quarter ending December of the preceding year are also released. Thus by the end of March every year, that is, by the end of every financial year, advance estimates of national income for the financial year that just ended, first revised estimates of national income for the preceding financial year and the quarterly estimates of national income up to the preceding quarter, that is, up to the quarter that ended in December of the financial year that just ended become available.

CSO sets before itself an advance release calendar for the release of national accounts statistics over a period of two years.[6] For instance, the Advanced Release Calendar for the years 2015 indicates that the Quarterly Estimate of GDP for any quarter during the period January, 2015 to end of September, 2015 would be available at the end of the next quarter. It also indicates the dates on which Advance Estimates, Quick Estimates and Revised Advance Estimates pertaining to the financial years ending on the 31st March, 2006, 30th June, 2006, 29th September, 2006 and 29th December, 2006, respectively would be released. Similarly, Revised Advance Estimates for the financial year 2004-05, Quick Estimates for the financial year 2004-05, Advance Estimates for the financial year 2005-06 and Revised Advance Estimates for the financial year 2005-06 would be released on the 30th June, 2005, the 31st January, 2006, the 7th February, 2006 and the 30th June, 2006, respectively. CSO has already (30th January, 2015) released the following estimates, as proposed:

i) First Revised Estimates for 2013-14 at current and constant (2011-12) prices,

[4] These 2 paragraphs have been taken from the Press release of the CSO dated 30th January 2015.

[5] 1. Agriculture, Forestry & Fishing. 2. Mining & Quarrying, 3. Manufacturing, 4. Electricity, Gas & Water Supply, 5. Construction, 6. Trade, Hotels & Restaurants, 7. Transport, Storage & Communication, 8. Financing, Insurance, Real Estate & Business Services, 9. Community, Social & Personal Services.

[6] This also takes care of SDDS referred to earlier.

ii) First revised Estimates of GDP, National Income, per capita National Income, Consumption Expenditure, Saving and Capital Formation for 2013-14 (along with revised estimates of the earlier two years)at current and constant (2011-12) prices,

iii) Advance Estimates of GDP, GNP and NNP and per capita NNP for 2014-15 at current and constant (2011-12) prices (on the 9th February, 2015).

The **NAS 2014** presents estimates of GNP, NNP and the corresponding domestic products GDP andNDP at factor cost and market prices at current and constant (2004-05) prices[7]. Besides giving estimatesof the components of GDP, estimates like Government Final Consumption Expenditure(GFCE), Private Final Consumption Expenditure (PFCE) in the domestic market, Gross Domestic Saving (GDS), Gross Domestic Capital Formation (GDCF), Exports, Imports, net factor income from abroad and the share ofthe public sector in GDP it presents[8], estimates of the following aggregates *at current and constant prices(2011-12 prices*) for 2013-14 and the period 2011-12 to 2013-14:

i) The contribution to GDP and NDP of different economic activities[9].

ii) Quarterly estimates of GDP by broad economic sectors;

iii) GDP of sub-groups of each of the economic activities (see footnote 9) 1.1 and also separately for its livestock sub-sector[10], 1.2,1.3,3.1,3.2,4, the sub-sectors of 4 namely, electricity, gas and water supply.

iv) NDP for the sub-group 10, namely, Public Administration & Defence, by (a) Central Government and Union Territories; (b) individual State Governments; (c) local authorities; and (d) quasi government bodies.

v) Factor incomes (compensation for employees and operating surplus/mixed income) of NDP for each economic activity (at current prices).

vi) Factor incomes of NDP of the unorganized segment of each economic activity (at current prices).

vii) Property incomes (rent and interest) and Financial Intermediation Services Indirectly Measured (FISIM) in the organized and unorganized segments of each economic activity (at current prices).

viii) Net National Disposable Income, Private Income, Personal Income and Personal Disposable Income.

---

[7] As mentioned earlier, the shift in the base for constant prices to 1999-00 was made in January, 2006, that is, after the publication of NAS 2005 in May, 2005. NAS 2006, due in June, 2006, would be presenting estimates at constant prices at 1999-00. Besides the press note of MoSPI referred to in the text, the **Economic Survey 2006 of the Ministry of Finance** presented by the Finance Minister to Parliament on the 27th February, 2006 also contains estimates released by CSO at current and constant (1999-00) prices for the years 2005-06 (Advanced Estimates), 2004-05 (Quick Estimates), 2003-04 and 2002-03.

[8] Estimates of some of these macro aggregates are also published in the **Monthly Bulletin of RBI**.

[9] 1. Agriculture and Allied Activities; 1.1. Agriculture; 1.2. Forestry& Logging; and 1.3. Fishing; 2. Mining & Quarrying; 3. Manufacturing; 4. Electricity, Gas, Water Supply and other utility services; 5. Construction; 6. Trade, Repair, Hotels & Restaurants; 7. Transport, Storage, Communication and services related to broadcasting; 8. Financial services, 9. Real Estate Ownership of dwelling &Professional Services; 10. Public administration and defence and 11 Other Services;

[10] The livestock sub-sector is a sub-sector of the sub-group 'agriculture' (1.1 in the preceding footnote) and includes animal husbandry, (fuel & manure), silk worm cocoons and honey.

ix) GDP and NDP of the Public Sector by type of public sector institutions[11] and economic activity.

x) Factor incomes of Public Sector NDP for each economic activity by type of institutions (current prices).

xi) Property incomes and FISIM in each economics activity of Public Sector by type of institutions.

xii) Value of output, inputs and consumption of fixed capital (CFC) for a number of economic activities.

xiii) Financial Assets and Liabilities of the Household sector.

xiv) Economic and Purpose Classification of Current and Capital Expenditure of Administrative Departments (current prices).

xv) Production Accounts of (a) Railways, (b) Communication, (c) Departmental Enterprises other than (a) and (b), and (d) Producers of Government Services.

xvi) Detailed external transactions accounts.

xvii) Depreciation as provided in the books of accounts of public sector institutional groups[12] and private corporate sector; and

xviii) Time series of GNI, GDP NNP, NDP, Per capita NNI and other macroeconomic aggregates like CFC, net factor income from abroad, PFCE in the domestic market, GFCE exports, imports and mid-year population from 1950-51 to 2013-14[13].

Estimates of the aggregates referred to at (i), (vii), (xiii), (xvi) and PFCE and its distribution by object and by type of goods, GFCE for the period 2004-05 to 2012-13 at current and constant (2004-05)Prices and the time series estimates referred to at (xvii) above at current and constant (2004-05) prices have been released by CSO in NAS 2014 by the end of May 2014 and posted on the **MoSPI website.**

What about separate estimates of National Income aggregates for rural and urban areas? CSO publishes estimates of NDP for rural and urban areas for each economic activity (see footnote 8) at current prices for each of the base year, namely, 1970-71, 1980-81, 1993-94, 1999-2000 and 2004-05.

b) **Other Publications Giving CSO's Estimates**

The Handbook of Statistics on the Indian Economy[14] – 2013-14, published by the **Reserve Bank of India (RBI)** also publishes time series from 1952-53 of macro economic aggregates (that would otherwise have to be compiled from different issues of the NAS document) like GDP at factor cost, CFC, NDP at factor cost, indirect taxes less subsidies, NDP at market prices, net factor income from abroad, GNP at factor cost, NNP at factor cost, GNP at market price, NNP at market price, personal disposable income, GDP and NDP of

[11] Administrative Departments, Departmental Enterprises, Non-departmental Enterprises and Quasi-Government Bodies.

[12] Administrative Departments, Public Sector departmental Enterprises subdivided into (i) Railways, (ii) Communication; and (iii) others and Public Sector Non-Departmental Enterprises subdivided into (i) financial companies; and (ii) non-financial companies.

[13] The **Pre-budget Economic survey** also presents the time series on GNP, GDP, NNP and per capita NNP and mid year population.

[14] Published annually. This can downloaded from the RBI WEBSITE http://rbidocs.rbi.org.in/rdocs/Publications/PDFs/000HSE13120914FL.pdf

public sector (from 1960-61), gross and net domestic capital formation, per capita GNP and NNP, at one place. The Economic and Political Weekly research Foundation (EPWRF) also presents comprehensive national income statistics in a time series format from time to time along with details of concepts, methodology and data sources for ready use by researchers. It provides data from **1950-51 to 2011-12 with 2004-05 as the base year (http://www.epwrfits.in/TimeSeriesDataResearch.aspx)"** is the latest. This can be obtained from EPWRF either as a subscriber or through a pay per module mode. National Accounts Statistics of most countries and areas of the world can be accessed at the UNSD website http://unstats.un.org/unsd/nationalaccount/madt.asp. At the time of writing the module, the latest year for which this data is available is 2013.

c) Limitations of the Estimates

The concepts and methodology used and the data sources utilized for making these estimates are set out in two publications of the CSO, namely, **"National accounts statistics : Sources and Methods" (available for the base 2004-05 series, published by the CSO in 2012)** and **"New Series on National Accounts (Base 2004-05)" (CSO, 2010). The MoSPI Press Note dated the 31st January, 2006** releasing Quick Estimates of National Income, etc., for 2004-05 indicates briefly the changes in the new series of National Accounts Statistics with base year 2004-05. The publication NAS- Manual of estimating state and District Income, 2008 provide the methodology for making estimates of GSDP and GDDP. Besides, the publication "National Accounts Statistics" brought out every year has a chapter titled "Notes on Methodology and Revision in the Estimates". This chapter in the NAS also contains tables explaining, sector-wise, reasons for the revision in GDP growth rate as per the different types of estimates (like revised advance, quick, etc., estimates) released in the preceding year. For instance, NAS 2014 gives the reasons for revision in GDP growth rate during.

i) 2012-13 between the provisional Estimates released in May, 2013 and first revised estimates released in January, 2014;

ii) 2011-12 between First Revised Estimates released in January, 2013 and Estimates released in January, 2014; and

iii) 2010-11 between Estimates released in January, 2013 and Estimates released in January, 2014.

Limitations of estimates of national income aggregates arise from insufficiency of data or the choice of data in capturing adequately income flows. Estimates of GDP/NDP at factor cost by sectors can be classified into two broad categories from the point of view of differences in the data base – *direct estimates* and *indirect estimates. Direct estimates* are based on statistics available annually on a regular basis so that these reflect year-to-year variations in the volume of the economic activities concerned. The translation of such annual statistics into National Accounts aggregates requires the use of certain norms and ratios or other assumptions. Resort to readily available indications of the economic activity in question has often to be made *when and if* there are delays on the part of the data-generating agencies in supplying the required regular annual data. As a result, revisions made in the estimates of national accounts aggregates when regular annual data become available later lead to major changes in the provisional estimates released earlier. *Direct estimates* mostly cover the institutional groups, (i) the public sector (and the Government

component within it), and (ii) the private corporate sector, or what is usually referred to as the 'organised' segment of the economy. On the other hand, *indirect estimates* require to be made when regular annual statistics are not available in respect of any economic activity. How are these made? Estimates based on periodic benchmark surveys are first derived for the survey year. These are then extrapolated forward or backward, as required, on the basis of physical indicators of the economic activities concerned. The degree of approximation involved in this procedure depends critically on the sensitivity of the indicator in reflecting year-to- year variations in the volume of the economic activity concerned. *Indirect estimates* are used in respect of the institutional groups, (a) households, and (b) non-profit institutions serving households (NPISH) and the "unorganized" segment of the economy. The share of *direct estimates* in aggregate GDP has increased from 57.6 per cent in the series with 1970-71 as the base year to 63.7 per cent in the series with 1980-81 as the base year and to 89.6 per cent in the series with 1993-94 as the base year. The share of *direct estimates* in different sectors in the GDP series with 2004-05 as the base year varied from 100 per cent in Mining, Registered Manufacturing, Electricity, Gas & Water Services, Railways and Public Administration &Defence sector to about 21 per cent in trade sector[15]. As such, the share of organized sector in 2012-13 had become only 44.7 per cent of the economy (with base year 2004-05, page 1i, NAS 2014).

The intention of this sub-section is not to comment adversely on the admittedly complex task of compiling national accounts statistics but to draw attention to limitations that need to be kept in mind by users as economists/statisticians/ analysts utilizing these data while carrying out their research, analytical and advisory work. Improvements in data and methodology for estimating national accounts aggregates constitute a continuous process, calling for efforts across sectors, in which Ministries in the central Government, regulatory agencies, the CSO, the State Directorates of Economics & Statistics and other research institutions participate. The interesting reader may like to familiarize himself/herself with the work being done by **the Indian Association for Research in National Income and Wealth**, **New Delhi**. **Their Bulletin** would show the efforts underway to bring about further improvements in national accounts methodologies and statistics. The recommendations of the **National Statistical Commission (2001)** will also pave the way to further enhancement of the quality of these estimates.

21.3.3 The Input-Output Table

Any economic activity is dependent on inputs from other economic activities for generating its output and the output from this economic activity serves as inputs for producing the output from other activities. Data relating to the interrelationship among different sectors of the economy and among different economic activities are thus important for analyzing the behavior of the economy and, therefore, for formulation of development plans and setting targets of macro variables like output, investment and employment. Such an input-output table will also be useful for analyzing the impact of changes in a sector of the economy or economic activity on other sectors of the economy and indeed the entire economy. An **Input-Output Transactions Table (IOTT)** is prepared by CSO from time to time. The latest to be published is the one relating to **2007-08**[16]. It gives, besides the complete table, the methodology

[15] This paragraph is based on Section 13.2 Chapter 13 of the Report of the National Statistical Commission.

[16] This can be accessed at http://mospi.nic.in/Mospi_New/upload/iott-07-08_6nov12.htm

adopted, the database made use of, analysis of the results and the supplementary tables derived from the IOTT giving the input structure and the commodity composition of the output. The Planning Commission updates and recalibrates the IOTT and prepares an Input-Output table for the base period of a Five Year Plan being formulated and another Input-Output table for the terminal year of the Five Year Plan. The detailed results of this exercise, carried out in the Planning Commission for the formulation of any Five Year Plan, are published by the Planning Commission as the **Technical Note to the Five Year Plan**. The Technical Note contains the relevant Input-Output Table, the methodology adopted and related material. The latest available in this series is the **Technical Note tothe Tenth Five Year Plan.** The IOTT of CSO and the Input-Output Table of the Planning Commission would be useful to researchers interested in exploiting the power of the input-output technique in economic and econometric analysis in their research work.

21.3.4 Regional Accounts – Estimates of State Income and Related Aggregates

a) **Estimates of States Domestic Product (SDP) Prepared and Released by State Governments and Union Territory Administrations**

State Accounts Statistics (SAS) consist of various accounts showing the flows of all transactions between the economic agents constituting the State economy and their stocks. The most important aggregate of SAS is the State Domestic Product (SDP) (State Income). Estimates of GSDP and NSDP at constant and current prices are being prepared and published by the **Directorates of Economics and Statistics (DES) of all State Governments and Union Territory Administrations except** the Union Territory Administrations of **Dadra & Nagar Haveli, Daman & Diu and Lakshadweep.**

b) **Other Publications Giving Estimates of SDP Made by State/UT DES**

The State Governments and Union Territory Administrations send their estimates of SDP to the CSO. These estimates at current and constant prices are available in the **CSO website**. The **pre-Budget Economic Survey** presented to Parliament by the Finance Minister every year contains these estimates (as a time series from 1993-94) of Net State Domestic Product (NSDP) at current prices. The **RBI Handbook** referred to above provides more detailed data drawn from the CSO website. It presents:

i) estimates of NSDP at current and constant prices – one set with the constant prices base 1980-81 for the period 1980-81 to 1998-99 and the other with the constant prices base 1993-94 for 1993-94 to 2003-04; and

ii) estimates of NSDP by economic activity and per capita NSDP at current and constant (1993-94) prices for the period 1993-94 to 2003-04; (new series).

Estimates of State Income (gross State Domestic Product) made by States and Union Territories by economic activity at *current as well as constant prices* are readily available in a (short) time series format at one place in the publication of the **CMIE (EIS)**, **Mumbai** referred to above. These estimates at current and constant prices by sectors for the period 1960-61 to 2000-01 have also been brought together at one place for the use of research scholars by the EPWRF in their publication **Domestic Product of States of India: 1960-61 to 2000-01**

(EPWRF, June, 2003). These are also available on the **website of the EPWRF** and in user-friendly and interactive **CD ROMS**.[17]

c) **Limitations of Estimates of SDP**

The preparation of estimates of SDP call for much more detailed data than for the preparation of national level estimates, especially on flows of goods and services and incomes across geographical boundaries of States/ Union Territories. Conceptually, estimates of SDP can be prepared by adopting two approaches. These are the *income originating approach* and the *income accruing approach.* In the former case, the measurement relates to the *income **originating** to the factors of production physically located within the area of a state.* In other words it is *the net value of goods and services produced within a state.* In the latter case, the measurement relates to the *income **accruing** to the factors of production physically located within the area of a State. Here,* the measurement relates to the *income **accuring** to the normal residents of a State.* The income accruing approach provides better measure of the welfare of the residents of the State. Preparation of estimates of SDP using the income accruing approach is not possible because data on inter-State flow of goods and services are not available. Compilation of other aggregates of State accounts is also problematic because the data required for the purpose, especially on inter-State flows of incomes, are not available.[18] Thus only the *income originating approach is used in preparing estimates of SDP.* This has to be kept in mind while using estimates of SDP.

Efforts have been made by the CSO over the years to bring about a good degree of uniformity across States and Union Territories in SDP concepts and methodology. The major milestones in these efforts are:[19]

i) the recommendations of the **First report of the Committee on Regional Accounts (CSO, 1974)** set up by the Government of India and the **Final Report of the Committee (CSO, 1976)** in the form of a set of Standard Tables to be prepared and a system of regional accounts, besides suggestions regarding steps to fill gaps in data *vis a vis* those needed for compilation of SDP, for the guidance of DES of States/Union Territories;

ii) the article "**Mahabaleshwar Accounts of States"** in the **October, 1976 issue of the Journal of Income & Wealth**;

iii) the article "**The Status of State Income Estimates"** appearing in the **October, 1980 issue of the Monthly Abstract of Statistics (CSO)**; and

iv) the article "Comparable Estimates of SDP – 1970-71 to 1975-76" appearing NAS, January, 1979 of CSO, which described the methodology of preparing estimates of SDP.

v) the publication "**National Accounts Statistics: Manual of Estimating State and District Income, 2008**" published by the CSO in 2008.

Are the estimates of SDP of different States/Union Territories comparable? Probably not. The successive Finance Commissions have found it necessary to get comparable estimates of NSDP and per capita NSDP made for their use for periods relevant to their work. These comparable estimates of SDP are published in the **Reports of the successive Finance Commissions** from the

[17] E-mail id for queries to EPWRF: epwrf@vsnl.com .

[18] The Report of the National Statistical Commission, Chapter 13 section 13.7.

[19] *ibid.* See also the Preface to EPWRF publication on SDP in the preceding sub-section (b) on other publications giving SDP estimates.

sixties. The **EPWRF publication on SDP** referred to earlier also provides comparable estimates of SDP and evaluates these with reference to those made for the periods 1960-61 to 1964-65, 1976-77 to 1978-79, 1982-83 to 1984-85, 1987-88 to 1989-90 and 1994-95 to 1996-97 for the use of the Finance Commissions. The question of comparability of estimates of SDP of different States/Union Territories is important from the point of view of their use in econometric work involving Inter-State or regional comparisons or the contribution of the regions or the States to the national product.

21.3.5 Regional Accounts – Estimates of Districts Income

The need for preparing estimates of district income has become urgent in the context of decentralization of governance and the importance of, and the emphasis on, decentralized planning or planning from below. Of late, it has gained a special importance, as this is one of the three indicators to compile the district-wise Human Development Index, prepared by some of the States. The State Governments have realized this need and estimates of District Domestic Product (DDP) are being prepared by a number of State DES and published by them in **State Statistical Handbooks/Abstracts/Economic Surveys** and are also posted on **their websites**. The position in different States and Union Territories (zone-wise) is as follows:

Type of estimate of DDP	State/ UT	Base year
1. for all years till 2013-14	Rajasthan	2004-05
2. for all years till 2012-13	Andhra Pradesh, Arunachal Pradesh, Assam, Madhya Pradesh, Maharashtra, Meghalaya, West Bengal	2004-05
3. for all years till 2011-12	Bihar, Haryana, Himachal Pradesh, Karnataka, Kerala, Chhattisgarh, Punjab, Tamil Nadu, Telangana, Uttar Pradesh	2004-05
4. for all years till 2010-11	Nagaland, Odisha, Uttarakhand	2004-05
5. Up to 2005-06/ 2006-07/ 2007-08	Jharkhand, Mizoram, Andaman &Nicobar Islands/ Manipur / Sikkim	1999 – 2000
6. Not available	Goa, Gujarat, Jammu & Kashmir, Tripura, Delhi, Puducherry	
7. Not required	Chandigarh, as the UT has only 1 district	

For some of the States in 6, it might be available with the State Directorate of Economics and Statistics, but not available centrally at one location.

The methodology for preparing estimates of district income at present is based on[20]

i) The recommendations of the **Report[21] of the Technical Group set up by the Department of Statistics** for recommending a suitable methodology for preparing estimates of district income;

ii) The methodology developed in August, 1996 jointly by the DES of Uttar Pradesh and Karnataka – a task entrusted to these organization by CSO in 1995.

[20] See Methodology of estimating State and District Domestic Product, CSO, 2008.
[21] Submitted in January, 1987.

The *income originating approach* is adopted for compiling district income estimates in the light of the kind of data that is available. As mentioned earlier, the *income accrual approach can* measure the welfare of the normal residents of the district in a more reliable manner. However, as flow of funds from district to district is not available and difficult to estimate, the income originating approach is followed. It is necessary to make adjustments, in the estimates based on the *income originating approach,* for flow of incomes across territories of districts that are rich in resources like minerals or forest resources and where there is a daily flow of commuters from other districts.

There are a lot of problems in the context of data availability separately for each district. For most of the commodity producing sectors, like agriculture, mining and registered manufacturing, data is fairly available. But, in remaining sectors, it is very scanty. Even within the agricultural sector, even if one can get the district-wise production, prices at each district sometimes may not be available. In case of entire unorganized sector, where national and State level estimates are based on periodic surveys, reliable district level estimates from these data cannot be prepared due to limitations in sample size.

21.3.6 National Income and Levels of Living

The foregoing paragraphs have talked about the availability of estimates of macroeconomic aggregates at the national and state levels, in different economic activities, in segments of economic and geographic spaces like public and private sectors, and households and rural and urban areas. What do trends in the magnitude of these variables say about the welfare of different sections of society? Precious little, perhaps, especially when these are considered without information on the manner in which the macroeconomic aggregates are distributed among these sections of society. Per-capita national income or even per-capita personal disposable income can only indicate overall (national) averages. Distribution of population by levels of income can be a big step forward in understanding how well the performance in the growth of GDP has translated into or has not translated into, improvements in levels of living for sections of society below levels considered the minimum desirable level. It would also help us analyse whether inequalities in levels of living have worsened or abated. What about the level of unemployment or levels and quality of employment? What about the health status of people? Or, to consider all these indicators of the state of a society and different sections of society, what are the levels of human development and gender discrimination? Such lines of analysis and the data required for the purpose are important from the point of view of planning for a strategy of growth with equity.

The Quinquennial Consumer Expenditure Surveys of the NSSO provides such data. These provide the distribution of households by monthly per capita consumption expenditure (MPCE) classes. You have already looked at such data and their implications for planning and policy in Unit 4 on “Poverty and Inequality – Policy Implications” in Block 1 on “Framework of the Indian Economy” in Course MEC-105 on “Indian Economic Policy” in your first year course. As you have seen there, the latest available data on this relate to 2011-12, flowing from **the 68th round NSSO survey of 2011-12**. The next such comprehensive survey (74thRound of NSS)is planned in 2016-17. You have also looked at the data on recent trends on growth rate of employment and the growth rate of the economy (employment elasticity) in Unit 5 on “Employment and Unemployment: Policy Implications” and comprehensive indicators like Human Development Index and the Gender Development Index

in Units 3 and 4 in Block 1 of Course MEC-005. You would have noted that the **Human Development Reports prepared by the Planning Commission (available on its website) and several State Governments** contain detailed data on these questions.

Check Your Progress 2

1) List the various components of the system of National Accounts.

……

……

……

……

……

2) Which have been the base years for compilation of national accounts statistics (NAS)? What is the difference between base year revision and revision of estimates in annual series of the NAS?

……

……

……

……

……

……

3) In addition to CSO, which other agencies bring out the data on macro economic aggregates?

……

……

……

……

……

……

……

4) By whom the estimates of state income and related aggregates are made?

……

……

……

……

……

……

21.4 SAVING

As you are aware, broadly speaking, GNP is made up of consumption, saving exports net of imports, besides net factor income from abroad. Saving is important in as much as it goes to finance investment, which in turn brings about growth of GNP. What is the volume of Saving relative of GNP? How much of it is consumed by the needs of depreciation? Who all contribute, and how much, to the total volume of Saving? Let us see what kind of data is available on such questions.

Estimates of Gross Domestic Saving (GDS) and Net Domestic Saving (NDS) in *current prices* and the Rate of Saving are made by **CSO** and published in the **National Accounts Statistics (NAS) every year** and the **Press Note of January of every year releasing First Revised Estimates**. These are first made for any year along with First Revised Estimates of GDP, etc., and revised and finalized along with its subsequent revisions. The structure of Savings, that is, the distribution of GDS and NDS by type of institution – household sector, private corporate sector and public sector are also available in this publication. Also presented in this document is the distribution of:

i) GDS, NDS and consumption of fixed capital (depreciation) by public sector, private corporate sector and household sector;

ii) Public sector GDS, consumption of public sector fixed capital and public sector NDS by type of institutions[22];

iii) Public sector GDS – by public authorities [sub-divided further into government administration and departmental (commercial) enterprises], non-departmental enterprises (broken up further into government companies and statutory corporations, the latter including port trusts also);

iv) Private sector GDS, consumption of fixed capital and NDS by private corporate sector and household sector;

v) Private corporate sector GDS by joint stock companies (distributed further into financial and non-financial companies) and cooperative banks and societies;

vi) GDS of the household sector into (net) financial saving and saving in physical assets and that of (net) financial saving, that is, currency, net deposits, shares and debentures, net claim on Government, life insurance funds and provident and pension funds;

vii) Financial assets [in greater detail than in (vi) above] and liabilities of the household sector leading to net financial saving of the household sector; and

viii) Item (xiii) in the list in sub-section (a) of Section 20.3.2 – these indicate estimates of net saving in these entities.

NAS 2014 presents the above estimates for 2012-13 and for the period 2004-05 to 2012-13. The **Press Note dated the 31st January, 2014 of MoSPI** releasing **Quick Estimates** provides estimates of GDS, NDS, and all the estimates referred to at (i) to (vii) above for 2012-13. Time series of estimates of GDS and NDS *in current prices* are, however, available from 1950-51 onwards. The

[22] See footnote 12.

time series on the structure of GDS by institutions (public, private corporate and household sectors) are presented in the **Pre-Budget Economic Survey (2014)** of the Ministry of Finance. **The Handbook of Statistics on the Indian Economy of RBI** presents time series data from 1952-53 on GDS and NDS and from 1970-71 onwards on (i) their components by institutions, and (ii) financial assets and liabilities of the household sector, item-wise, at current prices. **The EPWRF publication on NAS (fifth edition)** gives time series data from 1950-51 on all these variables and on Domestic Saving in public sector by type of institutions. The CMIE publication referred to earlier also contains time series data on GDS and NDS, though for a shorter period of time.

Estimates of Gross and Net Domestic Saving at the State and Union Territory levels are not being made at present by DESs of State/Union Territories.

The CSO publishes "**Sources and Methods"**, an additional publication showing the detailed methodology followed and data sources used for preparation of the National Income aggregates in the base year and the methodology to be followed since the change in base year. It may be added that estimates of Savings suffer from a number of limitations due largely to deficiency of data. This is particularly so in respect of estimates of various financial instruments from the private corporate sector and the household sector. Saving being the excess of income over expenditure, the major gaps in data relate to the household sector, non-profit institutions serving households (NPISH) and local bodies. Quality of data relating to the private corporate sector poses an additional problem. The estimate of Savings of the household sector in physical assets is thus derived as a residual, in the absence of direct data.[23]

21.5 INVESTMENT

Investment is Capital Formation (CF). Investment of money in the shares of a company or purchase of land is not investment but buying a house or machinery is investment. In other words, investment is creation of physical assets like machinery, equipment, building, inventories and so on, called "produced assets", adding to the capital stock (of such assets) in the economy, enhances the productive capacity of the economy. Investment or CF is another important component of GNP and the rate of investment – expressed as a proportion of GNP – largely determines the rate of growth of the economy. How is capital formation financed by the economy? What is the contribution of different sectors to capital formation or, how much is used up by different sectors? What is the capital stock available in the economy? These are all the questions that rise in one's mind when considering strategies for economic growth. What kind of data is available?

The annual NAS publication of the CSO presents estimates of Gross Domestic Capital Formation (GDCF), Gross Domestic Fixed Capital Formation (GDFCF), Change in Stocks, Consumption of Fixed Capital (CFC), Net Domestic Capital Formation (NDCF) and Net Domestic Fixed Capital Formation (NDFCF) in current prices and at constant prices. These estimates are made along with First Revised Estimates of National Income in January every year and the process of revision of these estimates proceeds along with that of the estimates of national income aggregates. The estimates of capital formation are presented in the NAS through two alternate approaches, namely,

[23] **Report of the National Statistical Commission**, Chapter 13, Section 13.6

through the flow of funds approach and through the commodity flow approach. The flow of funds approach is gross domestic savings plus net capital inflows from abroad, while the commodity flow approach, derived by the types of assets. **The successive issues of NAS** and the **Press Note of MoSPI of January every year** also present estimates of the distribution of the following aggregates at current and constant prices:

i) GDCF, GDFCF, Change in Stocks, CFC, NDCF and NDFCF by type of institutions (public sector, private corporate sector and household sector);

ii) GDFCF in the public sector, private corporate sector and household sector by type of assets (construction and machinery & equipment);

iii) GDCF, GDFCF, Change in Stocks, CFC, NDFCF and NDCF by economic activity[24];

iv) Net Capital Stock (NCS), Net Fixed Capital Stock (NFCS) and Inventory as on the 31st march of each calendar year by type of institutions (Government administrative departments, Departmental enterprises, Non-departmental enterprises, joint stock companies of the private corporate sector, cooperative and household sector);

v) NCS, NFCS and Inventory as on the 31st March of each calendar year by economic activity;

vi) Consolidated Account of the Nation – How Capital Formation is financed (current prices);

vii) Consolidated Account of the Nation – External Transactions showing current and capital transactions (in current prices);

viii) GDCF, GDFCF and Change in Stock in the public sector by economic activity and type of institutions;

ix) NDCF, NDFCF and Change in Stock in the public sector by economic activity and type of institutions;

x) GDFCF in each type of public sector institutions by type of assets (at current prices);

xi) (v) above for the public sector;

xii) Financing of Capital Formation in (a) Railways, (b) Public Sector Communication and (c) Administrative Departments including Departmental Enterprises other than (a) and (b), (d) Non-Departmental Financial and Non-Financial Enterprises (in current prices); and

xiii) Time series of GDCF and GFCF from 1951 to latest financial year.

NAS 2014 presents estimates of aggregates listed above at current and constant (2004-05) prices for the period 2004-05 to 2012-13. The **MoSPI Press Note of the 31st January, 2015** provide estimates of GCF, GFCF and consumption of fixed capital in current and constant (2011-12) prices as also of the distribution of these by industry of use and by institutional sectors of the economy, as defined in the SNA, namely, Public non-financial corporations, private non-financial corporations, public financial corporations, private financial corporations, general Government and Households for the period 2011-12 to 2013-14.

[24] See footnote 10.

Other publications providing time series of CSO data on capital formation are:

Sl. No.	Publication		Variable	Time Series From
1)	Pre-Budget Economic Survey (2006)		#GDCF, GFCF and Change in Stocks and their structure by by institutions	1950-51
2)	RBI Handbook of Statistics on the Indian Economy – 2005	*i)	GDCF and NDCF	1950-51
		*ii)	GDCF, GFCF and Change in Stocks and their structure by institutions.	1970-71
		#iii)	Financial assets and liabilities of the household sector (item-wise)	1970-71
3)	EPWRF publication on NAS (Fifth Edition)	*i)	data of the type in items (i) to (ix) and (xi) of NAS listed in the preceding paragraph of this section are generally given.	1950-51
		*ii)	Capital-output ratios by institutions;	1950-51
		*iii)	Average NFCS to Output	1950-51
4)	CMIE		*GDCF and GFCF	1960-61

at current prices *at current and constant (1993-94) prices

The publications/documents of **CSO** and other documents like the **Report of the National Statistical Commission** referred to in the section on National Income contain the methodology and the data utilized for preparing estimates of Capital Formation and related aggregates also.

Estimates of GFCF at the State level are being prepared by 14 States, namely, Punjab, Haryana, Himachal Pradesh, Madhya Pradesh, Orissa, Meghalaya, Assam, Andhra Pradesh, Tamil Nadu, Kerala, Karnataka, Maharashtra, Gujarat and Rajasthan. The status of these estimates are indicated in the **EPWRF publication on NAS,** referred to in the sub-section 20.3.4.

Gaps in data for the estimation of capital formation exist in a number of relevant areas[25]. Some of these are:

i) Non-availability of regular data on equipment and machinery (part of GFCF) in domestic production in unregistered manufacturing;

ii) Lack of annual enterprise surveys and surveys on NPISHs for estimating GFCF by institutional sectors of the economy;

iii) Complete accounts of local bodies are not available for some States/ UTs;

iv) Bank advances and other indirect approaches form the basis of estimation of Changes in Stocks, with benchmark estimates coming from quinquennial enterprise surveys; and

[25] **Report of the National Statistical Commission**, Chapter 13, Section 13.6.

v) Inadequacies in (a) the representative character of the sample of companies and (b) the blow-up factor used for estimation of saving and investment of the private corporate sector.

Check Your Progress 3

1) What approaches are followed by CSO to compile the saving data in National Accounts Statistics? Which estimates of investment are prepared by the CSO?

……

……

……

……

2) Go through the methodology contained in NAS 2015 and the data utilized for preparing estimates of capital formation and related aggregates.

……

……

……

……

3) From where can you get the data on contribution of different sectors to capital formation?

……

……

……

……

21.6 LET US SUM UP

The Indian Statistical System with the **Ministry of Statistics and Programme Implementation (MoSPI)** at its apex collects, compiles and disseminates an enormous amount of data on diverse aspects of the Indian economy, to the extent feasible, international standards prescribed by the **United Nations** and other international bodies. **MoSPI** sets standards to be followed by statistical agencies in the country and coordinates statistical activities carried out by these agencies. It also developes standardized classification of economic activities, namely, **NIC**, revising it from time to time keeping pace with the structural changes taking place in the economy. Data published by various Government agencies can be accessed through the portal www.data.gov.in . Other agencies also bring together and publish comprehensive data on some or all sectors of the economy, these publications being, the annual **Economic Survey** of the Ministry of Finance, the Budget documents, the Five-Year Plan documents, **RBI's Statistical Handbook on the Indian Economy**, **its monthly bulletin, the website on the database of the Indian Economy**, and the **Human Development Reports** prepared by the Planning Commission and the State Governments. Most of these data are available in electronic format – and/or on

the websites of the agencies concerned. One can usually download the published reports of the Government agencies from their respective websites free-of-charge. Various agencies also give anonymized data for analysis purposes. The **CMIE (EIU), Mumbai** and the **EPWRF, Mumbai** also bring together and publish comprehensive data on different sectors/aspects of the economy regularly. The **Statistical Handbooks of the State/Union Territory DES** and the **District Census Handbooks** published by the Census Directorates/Registrar general of India provide comprehensive information across sectors at the State and sub-State levels. Finally, the website of the United Nations Statistics Division provide access to methodological documents, data and report of various UN organisations which can be used for cross-country comparisons. The search for any kind of data can thus very well start from one of these sources and move on, if necessary, to the primary source.

National Accounts Statistics (NAS), which consists of estimates of National Income (NNI), GNI, GDP, NDP and related macroeconomic aggregates like Consumption Expenditure, Personal Disposable Income and Factor Incomes constitute an important tool for analyzing the performance of an economy. These compiled by **CSO**, **MoSPI**and published every year. In compiling these, CSO follows the guidelines given in 1993 **SNA of the United Nations** to the extent feasible. These estimates are published in great detail in CSO's annual publication **National Accounts Statistics (NAS)**. The latest issue of this publication – **NAS 2014 –** is also available on the **MoSPI website**. In addition, Advance Estimates of GNI, GDP, NNI and per capita income for a financial year are released two months before the end of the financial year through a **Press Note** and in the **MoSPI website**. First Revised Estimates of these aggregates relating to a financial year and revised estimates of earlier years are released by January of the next year, that is, within ten months after the close of the financial year through a **Press Note** and on the **MoSPI website**. Quartly Estimates of GDP are also prepared and released, generally within 45 days from the close of a quarter.

NAS estimates are presented in current and constant (base year) prices. **CSO** changes the base year from time to time to take note of structural changes in the economy and to depict a true picture of the economy. The base year was changed to 2004-05 in January, 2010 and to 2011-12 in January, 2015. Thus, as of the 1st March, 2015, Advance Estimates of GDP, etc., for 2014-15, First Revised Estimates of GDP, etc., for 2013-14, estimates of GDP for the quarter ending December, 2014 and for the period April-December, 2014, all in current and constant (2011-12) prices, have been released.

Estimates of Saving and Capital Formation form part of the National Accounts Statistics. These estimates are presented in detail in the publication **National Accounts Statistics, published in May of each year**. The press note of January, 2015, has provided the estimates referred to above with a changed base year 2011-12. The detailed estimates presented in **NAS** enable an analysis of various aspects of Saving and Capital Formation like the structure of Saving and Capital Formation, the manner in which Capital Formation is financed, the role of the public sector in Capital Formation and trends in the rate of Saving and Capital Formation and net capital flow from abroad.

Estimates of State Income are made and released by **DES of all States and Union Territories except Dadra & Nagar Haveli, Daman & Diu and Lakshadweep**. Estimates of GFCF are being made by 14 state DESs. District

Income estimates are made and released by DES of States. Estimates of State Income are also available in **MoSPI website**, besides the **Economics Survey**, **RBI Statistical Handbook on the Indian Economy** and the **CMIE** and **EPWRF (State Income)** publications referred to earlier. The last publication also give the State level GFCF estimates. The **RBI Handbook**, the **Economic Survey**, the **EPWRF publication on NAS** and the **CMIE** publication referred to earlier also present the national level estimates of GDP, etc., estimates of Saving and Capital Formation made by CSO. The **CMIE** and the **EPWRF** publications also contain almost as much detail as the publication **NAS** and in a time series format.

We have also noted the limitations and inadequacies to which estimates of national and state income and those of capital formation and saving are subject. Improvement in methodology and efforts to fill gaps in data constitute a continuous process.

21.7 EXERCISES

1) What type of data will you need to assess the performance of Indian Economy? Explain the various sources of such data?
2) State the difference between "Gross" and "Net" in the context of domestic and national income. How can one arrive at "Net National Income" from "Gross Value Added at basic prices"?
3) What are the relationships depicted in Input-Output Table? How are these tables useful in research work?
4) With the help of NAS 2014, explain the structure of saving in India.
5) What do you mean by the term capital formation? What is the capital stock available in India?

21.8 SOME USEFUL BOOKS

Central Statistics Office (2012): *National Income Statistics – Sources and Methods,* Ministry of Statistics and Programme Implementation, Government of India, New Delhi.

_____**(2008):** ***Methods*** *of estimating State and District Domestic Product.*

_____**(2014):** *National Accounts Statistics, 2014,*

_____**(2007):** *Input Output Transactions Table2007-08*

_____**(2015):** *Press Note New Series estimates of national income, consumption expenditure, saving and capital formation (base year 2011-12)*

International Labour Organisation (ILO) (1993) :*Resolution Concerning Statistics of Employment in the Informal Sector,* 15th International Conference of Labour Statisticians (ICLS), January, 1993. Geneva.

Indian Association for Research in National Income and Wealth & Institute of Economic Growth (1998): Golden Jubilee Seminar on Data Base of the Indian Economy, Delhi. A look at the Bulletins of the Association would also be useful.

http://mospi.gov..in : website of the Ministry of Statistics and Programme Implementation, Govt. of India.

http://rbi.org.in : website of the Reserve Bank of India.

http://unstats.un.org : website of the United Nation's Statistics Division.

Katyal, R.P., Sardana, M.G. and Satyanarayana, J. (2001): Estimates of DDP, Discussion Paper 2, National Workshop on State HDRs and the Estimation of District Income Poverty Under the State HDR Project Executed by the Planning Commission (GOI) with UNDP Support held in Bangalore in July, 2001, UNDP, New Delhi.

Ministry of Statistics & Programme Implementation (2001): Report of the National Statistical Commission, Ministry of Statistics and Programme Implementation, New Delhi.

System of National Accounts 1993 (SNA 1993) Commission of the European Communities, International Monetary Fund, Organisation for Economic Cooperation and Development United Nations, World Bank, Brussels/ Luxembourg, New York, Paris, Washington, D.C. **www.mospi.nic.in :** Website of the Ministry of Statistics and Programme Implementation.

21.9 ANSWERS ON HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) Coordination of statistical activities in the country maintenance of statistical norms and standards, providing liaison with control, state and international statistical agencies.

2) See Section 21.2

3) See Section 21.2

4) See Section 21.2

Check Your Progress 2

1) The data like estimates of national income and related macro economic aggregates (for example Personal Disposable Incomes, Saving, Consumption, Capital Formation etc.) which provide a comprehensive view of the internal and external transactions of the economy over a period form the system of National Accounts.

2) 1948-49, 1960-61, 1970-71, 1980-81, 1993-94, 1999-2000.

3) RBI – The Handbook of Statistics on the Indian Economy, CMIE – National Income Statistics, EPWRF publication on NAS and SDP etc.

4) Directorates of Statistics and Economics.

Check Your Progress 3

1) See Section 21.4

2) The annual document of NAS published by CSO.

3) See Section 21.5

UNIT 22 AGRICULTURAL AND INDUSTRIAL DATA

Structure

22.0 Objectives

22.1 Introduction

22.2 Agricultural Data

- 22.2.1 Agricultural Census
- 22.2.2 Studies on Cost of Cultivation
- 22.2.3 Annual Estimates of Crop Production
- 22.2.4 Livestock Census
- 22.2.5 Data on Production of Major Livestock Products
- 22.2.6 Agricultural Statistics at a Glance (ASG)
- 22.2.7 Another Source of Data on Irrigation
- 22.2.8 Other Data on the Agricultural Sector

22.3 Industrial Data

- 22.3.1 Data Sources Covering the Entire Industrial Sector
- 22.3.2 Factory Sector – Annual Survey of Industries (ASI)
- 22.3.3 Monthly Production of Selected Industries and Index of Industrial Production (IIP)
- 22.3.4 Data on Unorganised Industrial Sector
- 22.3.5 Industrial Credit and Finance
- 22.3.6 Contribution of Industrial Sector to GDP

22.4 Let Us Sum Up

22.5 Exercises

22.6 Some Useful Books

22.7 Answer or Hints to Check Your Progress Exercises

22.0 OBJECTIVES

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

- know the kind of agricultural data available;
- explain the different sources of agricultural data and different data generating agencies;
- describe the characteristics of multivariable industrial data; and
- know how to access the different types of agricultural and industrial data.

22.1 INTROCUTION

In the previous unit, we have looked at data on macro variables like National Income, Saving and Investment, which tell us about the dimension and broad structure of the economy and the direction in which the economy is moving.

We shall now look at two of the sectors of the economy in this unit. One is the agricultural sector that provides livelihood to roughly two-thirds of the total workforce in the country. The other is the industrial sector, which consists of manufacturing, power, gas and water supply.

22.2 AGRICULTURAL DATA

You are already aware of the importance of agriculture to the Indian economy and indeed to the Indian way of life. You would, therefore, like to examine several aspects of agriculture like the level of production of different crops and commodities, the availability and utilization of important inputs for agricultural production, incentives, availability of post-harvest services and the role of agriculture in development. Similarly, you would like to know about the livestock and their products, fisheries and forestry, people engaged in these activities and so on. All these analyses require enormous amount of data over time and space. Let us have a look at what kind of data are available and where.

The **Directorate of Economics and Statistics (DESMOA) in the Department of Agriculture and Cooperation (website http://eands.dacnet.nic.in)** and the **Animal Husbandry Statistics Division (AHSD) of the Department of Animal Husbandry, Dairy and Fisheries (website http://dahd.nic.in/dahd/statistics/animal-husbandry-statistics.aspx)** in the Ministry of Agriculture are the major sources of data on agriculture and allied activities. Among the major efforts at collection of data mounted at regular intervals, the following are the major ones:

i) the quiquennial agricultural census and the input survey;

ii) the cost of cultivation studies;

iii) annual estimates of crop production;

iv) the quinquennial livestock census; and

v) integrated sample survey to estimate the production of major livestock products.

DESMOA releases statistics on agriculture and allied activities flowing from its work and also collected from other division (like the **Agricultural Census Division)** and Departments of the Ministry of Agricultural and other Ministries and agencies in their comprehensive annual publication **Agricultural Statistics at a Glance (ASG)**[1]. Other publications include, **"Cost of Cultivation in India"**, the monthly bulletin **"Agricultural Situation In India"**. **AHSD** publishes the data flowing from its activities [which include the activities mentioned at (iv) and (v) above] and those collected from other divisions of the Department of Animal Husbandry, Dairy and Fisheries and other agencies in its biennial publication **Basic Animal Husbandry Statistics (BAHS)**.

22.2.1 Agricultural Census

Agriculture Census forms part of a broader system of collection of Agricultural Statistics. It is a large-scale statistical operation for the collection and derivation of quantitative information about the structure of agriculture in the country. An

[1] This can be accessed in the **Ministry of Agriculture (Department of Agriculture) website** . One can also access **www.data.gov.in** and click on

agricultural operational holding is the ultimate unit for taking decision for development of Agriculture at micro level. It is for this reason that an operational holding is taken as the statistical unit of data collection for describing the structure of agriculture. Through Agriculture Census it is endeavored to collect basic data on important aspects of agricultural economy for all the operational holdings in the country. Aggregation of data is done at various levels of administrative units. The agricultural census started in 1970-71. The ninth of these censuses relates to the reference year 2010-11.

Periodic Agriculture Censuses are important as these are the main source of information on basic characteristics of operational holdings such as land use and cropping patterns, irrigation status, tenancy particulars and the terms of leasing. This information is tabulated by different size classes and social groups including Scheduled Castes/Scheduled Tribes which are needed for development planning, socio-economic policy formulation and establishment of national priorities. The census also provides the basis for the development of a comprehensive integrated national system of agricultural statistics and has links with various components of the national statistical system. The whole project of Agriculture Census in the country is implemented in three distinct phases, which are statistically linked together but focus on different aspects of agricultural statistics. In Phase-I, a list of holdings with their area and social characteristics and gender of the holders is prepared. In Phase-II, detailed data on agricultural characteristics of holdings are collected from selected villages. Thus the whole operation of Agriculture Census in India is not really a complete Census. In fact, it is a combination of Census and Sample Survey.[2]

The third phase, which is better known as the **Input Survey** and conducted in the year following the census year[3], relates to collection of data on the pattern of input-use across crops, regions and size-groups of holdings. The main objective of the input survey is to generate data on consumption of various agricultural inputs, according to major size-groups of operational holdings, viz., marginal (below 1 ha.), small (1- 1.99 ha.), semi-medium (2- 3.99 ha.), medium (4- 9.99 ha.) and large (10 ha. and above), for getting an insight into the consumption pattern of inputs by various categories of farmers. This information is vital for planning production, imports and distribution of fertilizers. The inputs covered in the survey include chemical fertilizers, HYV seeds, pesticides, farmyard manures/compost, bio-fertilizers, agricultural implements and machinery, livestock and agricultural credit besides data on input use including use of certified/notified seeds, high yielding variety seeds, pest control measurements adopted by cultivators, educational qualification, age and size of households of operational holders are captured through Input Survey. The results of the first phase, including the all-India report of the first phase of the ninth agricultural census can be accessed at http://agcensus.nic.in/ agcen201011.html. The provisional results of the second phase have also been released. Data collection and processing of the third phase is progressing in the different States and UTs. Selective information is also available on the **internet[4] through** a query-based facility and can provide data on operational

[2] This part is reproduced from http://agcensus.dacnet.nic.in/, which can be accessed to get the latest data, up to tehsil level from Agricultural Census 2013.

[3] The input survey was initiated in the 1975-76 Agricultural Census and conducted in 1976-77.

[4] On the website http://agcensus.dacnet.nic.in/nationalholdingtype.aspx, one can access tables related to agricultural census years 1995-96 to 2010-11. Further, On the website http://inputsurvey.dacnet.nic.in/nationalsummary.aspx one can access the tables related to 1996-97 to 2006-07 so far.

holdings at the national, State, district and *tehsil* levels by type, characteristics and size class/group. You can, for example, get information on operational holdings by size class of the holding, social group (Scheduled Caste, Scheduled Tribe, Other and all groups) and tenancy characteristics of the holding (lease terms, land use, irrigation status, irrigation by sources, wells and tube wells, gross cropped area, cropping pattern, and area dispersal. Clearly, there is a time lag in the availability of agricultural census and Input Survey results.

22.2.2 Studies on Cost of Cultivation

DESMOA implements a comprehensive scheme for studying the cost of cultivation of principal crops in India and this results in the collection and compilation of field data on the cost of cultivation and production in respect of 29 principal crops leading to estimates of crop-wise and State-wise costs of cultivation and also computation of the *index of the terms of trade between agriculture and non-agricultural sectors*. The scheme covers 19 States and foodgrain crops, oil seeds and commercial crops and selected vegetables. Statistics of cost of cultivation of principal crops are published by **DESMOA** in its publication **"Cost of Cultivation in India"** and also in **ASG**. The **Commission for Agricultural Costs and Prices (CACP)** makes use of the estimates of cost of cultivation and production and the structure of cost of cultivation flowing from these studies, along with an analysis of a wide spectrum of data on variables like market prices, productivity of the crops concerned, domestic and global inter-crop price parity, the terms of trade between the agricultural and the non-agricultural sectors and the supply-demand situation in arriving at their recommendations to Government on Minimum Support Prices (MSP). The **CACP reports** thus contain not only the data thrown up by the cost of cultivation studies but also those assembled by it from various sources for its work. For example any of the **CACP report** presents:

i) estimates of cost of cultivation per hectare, cost of production per quintal, and the value of the main product and the byproduct;

ii) the breakup of the cost by inputs like human labour, bullock labour, machine labour, seed, fertilizers and manures, insecticides, irrigation charges and so on;

iii) a variable input price index; and

iv) month-end wholesale prices of the commodities concerned by variety and location.

for all the commodities dealt with in the report and for all states. In addition, it gives information on the index number of whole sale prices for farm inputs, index numbers of whole sale prices of cereals, pulses, etc., and data on a host of aspects relevant to the subject matter of the report like area, production, yield of major crops, buffer stock of cereals and month-wise and area-wise daily wage rates for agricultural labour[5]. MSPs for different commodities are also published in **ASG.**

[5] The cost of cultivation studies reports and related data can be accessed at http://eands.dacnet.nic.in/Cost_of_Cultivation.htm. Now, the data is available for reference years 2004-05 to 2011-12. The details regarding the work of CACP and the data presented in **CACP reports** are based on the information available in the **CACP website http://cacp.dacnet.nic.in** . The MSP up to agricultural year 2014-15 can be viewed in the link http://cacp.dacnet.nic.in/ViewContents.aspx?Input=1&PageId=36&KeyId=0A .

The other useful outcome of the cost of cultivation studies, as already mentioned, is the **index of the terms of trade between the agricultural and non-agricultural sectors.** This is defined as

$$\text{Index of the Terms of Trade} = \frac{\text{IPR [Index of Prices Received]}}{\text{IPP [Index of Prices Paid]}}$$

(Base Year : the triennium ending 1991-92)

The numerator relates to the prices that the agricultural sector gets for its produce while the denominator relates to the prices that the sector pays for products it purchases for final consumption, intermediate consumption and capital formation. The time series of ITT from 1981-82 onwards, complete with individual values of IPR and IPP and those of the component indices of IPP, namely, final consumption (weight in IPP 73.54), intermediate consumption (weight in IPP 21.63) and capital formation (weight in IPP 4.83) are published in **ASG**.

The Terms of Trade, specifically, the terms of trade between agriculture and industry is the ratio of agricultural prices to industrial prices, both measured as price indices. The terms of trade between agricultural and industry lies at the heart of the government's policy in the agricultural sector. A rise in the ratio (agricultural prices divided by industrial prices) means that the agricultural sector is better off in terms of its purchasing power of industrial goods. Suppose there are just two goods in the economy, bananas and pins and you produce and sell bananas. Both initially cost Rs. 10. If now both prices go up by 100 per cent, nothing happens to the terms of trade and purchasing power remains unchanged. If instead, banana prices go up by 100 per cent and the price of pins doesn't, then as a banana manufacturer you are better off since one banana can now buy two pins instead of one. The ratio or the terms of trade between bananas and pins has doubled. Substitute agriculture for bananas and industry for pins and you can get what a rise in agriculture's terms of trade does for farmers.[6]

The National Statistical Commission has pointed out that there is a sizeable time lag in the availability of results for use in the work of CACP and that steps should be taken to cut down delays in the release of the results of these studies[7].

22.2.3 Annual Estimates of Crop Production

DESMOA makes and releases annual estimates of area, production and yield in respect of principal crops of foodgrains, oil seeds, sugar cane, fibres and important commercial and horticulture crops. These crops account for about 87 per cent of the total agricultural output. Estimates of area are based on a reporting system that is a mix of complete coverage and coverage by a sample, those of yield are based on a **system of crop cutting experiments** and **General Crop Estimation Surveys**. The preparation of these estimates takes time. Estimates of crop production are, however, needed earlier and in fact even

[6] This part is taken from an article published in the economic times ET classrooms series, which can be accessed at http://articles.economictimes.indiatimes.com/2002-02-04/news/27340862_1_bananas-issue-prices-trade.

[7] Report of the national Statistical Commission, Chapter 4.

before the crops are harvested, for policy purposes. Advance estimates of crop production are, therefore, made. The first assessment of the *kharif* crop is made in the middle of September, that is, when the South-West Monsoon is about to be over. The second advance estimate – a second assessment of the earlier estimate of the *kharif* crop and the first assessment of the *rabi* crop – is made in January. The third advance estimate is made at the end of March or the beginning of April and the fourth in June. The methodology for estimating area and yield of crops and for making advance estimates are given in **ASG**. Estimates of annual[8] production of the principal crops mentioned above, gross area under different crops and yield per hectare of these crops and Index Numbers on these variables with base year the triennium ending 2007-08 = 100[9] are also available in **ASG** in the form of a time series. So are estimates of production of crops by States.

Check Your Progress 1

1) Give two major sources of data on Agriculture and allied activities like dairy and fisheries.

……

……

……

……

……

2) How can you retrieve the agricultural census data at district and tehsil level?

……

……

……

……

……

3) How are annual estimates of crop production made by the Directorate of Statistics and Economics?

……

……

……

……

……

[8] Season-wise in some cases.

[9] The base of these index numbers is a triennium in order to even out year to year to fluctuations of the underlying variable(s). The earlier base TE 1981-82 for these index number series was changed to TE 1993-94 in 2000-01 in order to update the index numbers and also for these to be in harmony with the index numbers of wholesale prices and industrial production with base 1993-94 = 100.

22.2.4 Livestock Census

The last quinquennial livestock census, the nineteenth in the series, was conducted in 2012. The census has collected information, district-wise on livestock, poultry, fishery and also agricultural implements. Livestock covers cattle, buffaloes, sheep, goats, pigs, horses and ponies, mules, donkeys, camels, yak, *mithun* and also, dogs and rabbits. These are classified by age (appropriate for the species). sex, breed, function ('breeding', 'work', 'both' and 'others' for males and 'in milk' or 'dry' in the case of females). Poultry covers cock, hen, duck, and drake, which are classified as *desi* and 'improved' varieties. Fishery covers fishing activity (inland capture, inland culture, marine capture and marine culture), members involved in fishing (male, female and child), persons engaged in fishing (part time male, full time male, part time female and full time female) craft/gear, namely, trawlers, gin-netters, liners, seiners, motorized canoe/*maran*, non-motorized canoe/*maran,* and miscellaneous gears like trawler-net, gill netter-net, cast-net, drag-net, hook lines, set barriers and others. Each of these classes except the last one is further classified into size (length) and horsepower. Agricultural implements cover annually operated implements, animal operated implements, plant protection equipment, water lifting devices, tractor and power operated implements, equipment for livestock and poultry and horticulture tools. Each group provides further data on four or five specific equipments like for example, thresher under the first group, ST Plough under the second, TPO Spray under the third, diesel under the fourth, Crawler Tractor under the fifth, incubator under the sixth and power operated tools under the seventh group.

The provisional results of the 2012 census are available on the website of the Department of Animal Husbandry, which can be accessed at http://dahd.nic.in/dahd/WriteReadData/Livestock.pdf. **ASG** and **BAHS** also present some data from the livestock census.

22.2.5 Data on Production of Major Livestock Products

AHSD is responsible for collection of statistics on animal husbandry, dairy and fisheries. These are published, as already mentioned, in **BAHS**. The latest relates to 2004. This presents data up to 2003-04, on

i) long time series of estimates of production of milk, eggs, meat and wool;

ii) State-wise short time series of such estimates;

iii) State-wise and national estimates of per capita availability of milk and eggs;

iv) contribution of cows, buffaloes and goats to milk production and of fowls and ducks to egg production in different States in 2003-04;

v) average annual growth rates of production of major livestock products, milk, eggs and wool;

vi) short time series on imports and exports of livestock and livestock products;

vii) short time series on area under fodder crops, pastures and grazing in different States;

viii) estimates of dry and green fodder production in different States during the three-year period ending 2002-03;

ix) a short time series on the no. of artificial inseminations performed in different States;

x) achievements in key components of dairy development;

xi) livestock and poultry (numbers) in India and in different States – 1992, 1997 and 2003; and

xii) time series on world livestock production and milk production.

22.2.6 Agricultural Statistics at a Glance (ASG)

The pocket book on agricultural statistics and the agricultural statistics at a glance are two annual publications available in the website http://eands.dacnet.nic.in/latest_2006.htm. One of the most important statistics, which is the statistics on Land utilization, is available in the **ASG**. As for inputs and access of individual operational holdings to such inputs, **ASG** provides data on:

i) Methodology of crop estimation,

ii) Socio-economic indicators: population and growth rates, State-wise classification of cultivators, Gross State Domestic Product for agriculture and allied activities,

iii) Outlays, expenditure and capital formation in agriculture,

iv) Area, production and yield of principal crops: target and achievement of production of major crops, three largest producing States of important crops, season-wise area, production and yield of food grains (rice, wheat, jowar, bajra, maize, total pulses, gram, tur, masur, nine oilseeds, groundnut, rapseed and mustard, soyabean, sunflower), cotton, jute, mesta, sugarcane, tobacco, guarseeed, etc.; yield rates of principal crops, area and yield under the high yielding varieties, area and production of horticulture and plantation crops (potato, onion, coconut, cashewnut, arecanut, garlic, ginger, sweet potato, turmeric, chillies, cardamom, pepper);

v) All-India index numbers of area, production and yield of principal crops;

vi) Area, production and yield of major crops in different countries;

vii) Minimum support prices/marketed surplus ratios of various agricultural commodities;

viii) Procurement by public agencies, per capita net availability of foodgrains, consumption and stocks of foodgrains and raw jute,

ix) Import and export .of agricultural commodities, tariffs and bound rates of major agricultural commodities, trends in wholesale price index of foodgrains and commercial crops;

x) Land use statistics: agricultural land by use in India, selected categories of land use;

xi) Inputs: production and use of agricultural inputs (fertilizers, crop-wise distribution of certified/quality seeds, consumption of electricity for agricultural purposes, flow of institutional credit for agricultural sector, state-wise number of kisan credit cards, state-wise/season-wise national agricultural insurance scheme, etc.

xii) Agricultural census data: number and area of operational holdings by size group – marginal (size less than one HA[10] area), small (1 to 2 HA), semi-medium (2 to 4 HA), medium (4 to 10 HA) and large (more than 10 HA), area irrigated by different source of irrigation and by size class and the position in different States;

xiii) Estimated number of rural households, farmer households, indebted farmer households by size class of land possessed;

xiv) Ceiling on land holding and wages for agricultural work;

xv) Livestock population in India, all-India and State-wise production of milk, eggs, meat and wool;

xvi) Fish production in India and States;

xvii) Rainfall scenario and management of natural disaster.

The subjects **Subsidies** is a much-debated subject nowadays and agricultural subsidies in developing countries and developed countries of Europe and in USA are also much in the news. The Ministry of Agriculture provides a time series of the amount of subsidy given to agricultural with its break-up into subsidy for (i) fertilizers (ii) electricity, and (iii) irrigation[11], (iv) other subsidies given to marginal farmers and Farmers Cooperative Societies in the form of seeds, development of oil seeds, pulses, etc. One can easily derive the size of the total subsidy relative to GDP in current prices. The limitations to which these figures are subject need to be kept in mind.

22.2.7 Another Source of Data on Irrigation

The data on irrigation presented in **ASG** (referred to the proceeding sub-section) is based on crop statistics collected by **DESMOA** from village through a mix of a reporting system and sample surveys. The **Central Water Commission (CWC) under the Ministry of Water Resources** collects hydrological data on the important river systems in the country through 877 hydrological observation sites. The Ministry also conducts periodic **Censuses of Minor Irrigation Works** along with a sample check to correct the Census data. The **five Census** conducted so far related to **1986-87, 1993-94, 2000-01**, **2006-07 and 2013-14**. This is conducted in all States/UTs except in Daman & Diu and Lakshadweep. The reports can be seen in the **website of the Ministry http://micensus.gov.in.The Report of the Minor Irrigation Census 2006-07** provides information on the 20.7 million minor irrigation works in the country like the share of ground water and surface water works, crop-wise utilization of the irrigation potential created, the manner in which the water is distributed in the field – sprinkler, drip, open channel or underground water irrigation and state-wise distribution of these variables.

The National Commission on Statistics had in their report made the following observations[12] on irrigation data available from **DESMOA** and the **Ministry of Water Resources**:

[10] HA: abbreviation for hectare.

[11] The rates for supply of water to farmers are kept low as a matter of policy. This results in a loss to the Government Irrigation System. The excess of operating costs over gross revenue is treated as imputed irrigation subsidy.

[12] Chapter 4 of the Report.

i) The **CWC** had a large volume of data on various aspects of irrigation without any statistical analysis of such data being carried out. The CSO and the CWC need to get together and ensure that the data collected in analysed and put to use by the statistical machinery.

ii) There is a large variation between the statistics on the "area irrigated" published by **DESMOA** and the "irrigation potential actually utilized" published by the Ministry of Water Resources.

iii) Data users should be made aware of the reasons for such variation – differences (if any) in concepts, definition etc.

iv) The reluctance on the part of the State Governments to furnish data in view of their vested interest in the sharing of water is an added problem.

v) Data from both the sources – **DESMOA** and the **Ministry of Water Resources** – are available after a large time lag and this needs to be reduced.

22.2.8 Other Data on the Agricultural Sector

Data on forest cover is, already mentioned, as part of land-use statistics presented on the basis of nine fold land-use classification in **ASG**. **Forest Survey of India (FSI)** also collects data on forest cover (dense forest, open forest and mangroves) through a biennial survey by making of Remote Sensing (RS) technology since 1987. The latest survey relates to 2013. Digital interpretation has reduced the time lag in the availability of such data obtained earlier through periodic reports from field formations. There are discrepancies between the data on forest area given by **ASG** and those given by **FSI** due to differences in the concepts and definitions used by the two organizations in collecting data. Data on production of industrial wood, minor forest produce and fuel wood are available with the **Principal Chief Conservator of Forests in the Ministry of Environment and Forests.**

The National Bank for Agriculture and Rural Development (NABARD) is another agency that is closely involved in agriculture, especially in the matter of credit to the sector through cooperatives. The **annual reports of NABARD and its publication** like **"Statistical Statements Relating to the Cooperative Movement in India"** (Part I deals with credit societies and the other with non-credit societies) and **"Key Statistics on Cooperative Banks"**, besides **NABARD's website** (https://www.nabard.org/english/home.aspx) would be useful source of information on agricultural credit.

Statistics on area, production and yield of crops and subsidies given to agriculture are also available as time series in the **Economic Survey, expenditure budget documents of the Ministry of Finance and the RBI Handbook of Statistics on Indian Economy (http://rbidocs.rbi.org.in/rdocs/Publications/PDFs/000HSE13120914FL.pdf)**. The third source is important in the context of the policy direction to banks and all financial institutions to ensure that 40 per cent of the total institutional credit should flow to the priority sector of which agricultural sector forms a part. The data warehouse of the RBI can be accessed at http://dbie.rbi.org.in. **The RBI Handbook of Statistics on Indian Economy – 2013-14** publishes time series data on:

i) National income, saving and employment from 1952-53 onwards;

ii) Output and prices of agricultural production, from 1956-57 onwards;

iii) Assets and liabilities of the RBI, components of money stock, etc. from 1962-63 onwards;

iv) Select aggregates of Scheduled Commercial Banks (SCBs), like demand deposit, time deposit, borrowings from RBI, liabilities to banks, investment in Government securities, food credit, non-food credit, cash in bank, balance with RBI, assets with banks from 1956-57 onwards;

v) Sectoral deployment of non-food gross bank credit; industry-wise deployment of gross bank credit;

vi) Direct and indirect institutional credit to agriculture and allied activities, 1975-76 onwards;

vii) SCB's direct finance to farmers according to size of land holdings, 1985-86 onwards;

viii) the distribution of non-food credit by priority sector credit and credit for other sectors;

ix) the share of agriculture in priority sector credit;

x) Financial market: structure of interest rate, etc.;

xi) Public finances: key deficit indicators of central Government, major components of receipts of central Government, key deficit indicators of state governments, pattern of receipts of state governments, etc.;

xii) Trade and balance of payment, etc.

Direct credit to agriculture is also available from other institutions and the Handbook furnishes data on:

i) the short and long term loans issued by each of the institutions – Cooperatives, Scheduled Commercial Banks (SCBs), Regional Rural Banks (RRBs) and the State Governments directly to beneficiaries or borrows for agriculture and activities allied to it during a year; and

ii) the loans outstanding at the end of each year.

Institutions like SCBs, Cooperative, RRBs and the Rural Electrification Corporation (REC) provide indirect credit to agriculture and allied activities. This is routed through some other agency, conduit or tier like the Electricity Boards (EBs), the Primary Agricultural Credit Societies (PACS), the Farmers' Service Societies (FSS), the Large-sized *Adivasi* Multi-Purpose Societies (LAMPS) and State-supported corporations and agencies. Such indirect credit is designed to promote agricultural productivity or to increase the agricultural income of the ultimate beneficiary, through a number of steps like distribution of fertilizers and other inputs to these activities, loans to EBs and loans to farmers through PACS, FSS and LAMPS. Other types of indirect credit include advances to State-supported corporations and agencies for onward lending of funds (up to Rs. 10,000/-) to weaker sections of society engaged in agriculture, namely, small and marginal farmers and those engaged in activities allied to agriculture.

We have already seen in the preceding Unit (on Macro-variable like national income) that **National Accounts Statistics (NAS)** publishes data on the contribution of agriculture and its sub-sectors to GDP and other measures of

national/domestic product. **NAS** also provides information on value of output of various agricultural crops including those of drugs and narcotics, fibers, fruits and vegetables, livestock products like milk group products, meat group products, eggs, wool and hair, dung, silkworm cocoons and honey, forestry products like industrial wood, firewood and minor forest produce, inland fish and marine fish. It also furnishes data on capital formation in agriculture and animal husbandry, forestry and logging and fishing.

Check Your Progress 2

1) How can you access the data on per-capita availability of milk, egg, wool, cow and buffaloes?

……..

……..

……..

……..

……..

2) What do you mean by the term 'cropping intensity'?

……..

……..

……..

……..

……..

3) Which document contains the data on subsidy given to agriculture?

……..

……..

……..

……..

……..

……..

4) Name the document which contains the data on food and non food credit provided by Scheduled Commercial Banks.

……..

……..

……..

……..

……..

……..

22.3 INDUSTRIAL DATA

The industrial sector can be divided into a number of subgroups. This grouping is different from grouping by economic activity. Such divisions arise from framework factors like applicability or coverage or certain laws, employment size of establishments of groupings for purposes of promotional support by Government. Such groupings are the organized and unorganized sectors, the factory sector (covered by the Factories Act, 1948 and the Bidi and Cigar Workers Condition of Employment Act, 1966), small-scale industries, cottage industries, handicrafts, *khadi* and village industries (KVI), directory establishments (DE), non-directory establishments (NDE), own account enterprises (OAE)[13]. Attempts have been made to get at a detailed look at the characteristics of some of these sub-sectors of the industrial sector, as the data sources covering the whole sector often do not provide information in such detail. Let us now examine the kind of data available for such individual subgroups of the industrial sector and those that cover the entire industrial sector.[14]

22.3.1 Data Sources Covering the Entire Industrial Sector

There are five sources that cover the whole spectrum of economic activity or non-agricultural activities and therefore, the entire industrial sector. However, four of these sources give data only on levels of industrial employment:

i) The first source is **the decennial Population Census**. This provides data on the levels of employment in various economic activities across the economic spectrum, and therefore, the industrial sector, down to the latest NIC[15] three-digit code levels. Thus, Census 1991 provides data at 3-digit level of NIC 1987, Census 2001 provides data at 3-digit level of NIC 1998 and Census 2011 will be providing data at 3-digit level of NIC 2008. It also provides (i) employment levels in each of the three-digit NIC code classified by Occupational Divisions [the first digit of the National Classification of Occupations 1968 (NCO 1968)], (ii) industrial employment in broad industrial sectors (a nine-fold classification of economic activities is considered[16]) classified by broad age groups, and (iii) industrial employment in broad sectors (nine-fold classification, as above) by levels of education. Such details are available up to district levels. These data, however, become available after a time lag of 4-5 years after completion of the Census field-work.

ii) The second source consists of the **quinquennial sample surveys** relating to labour force,employment and unemployment conducted by the National Sample Survey Office (NSSO). These surveys also provide similar type of data on industrial employment, down to State levels, separately for rural and urban and males and females. Data by the districts can also be tabulated, with a suitable caution for the district level sample sizes, as the

[13] DEs are those establishments that employ at least six persons. NDEs are establishments that employ at least one but not more than five employees. OAEs are the self-employed.

[14] One distinction between the economic activity classification and the industrial sector classification is the placement of Mining and Quarrying industries. Although this is a part of "primary sector" under the economic activities, it is considered as a part of industrial sector.

[15] See the section on India Statistical System in Unit 20 on Macro-Variable Data: National Income etc.

[16] Same as footnote 15 given above.

primary data collected (unit record data) can be obtained on CDs from NSSO[17]. Key results from the NSSO surveys are available after about six months, while the final report becomes available one year after the surveys are completed.

iii) **The third source is the Economic Census**, which has been conducted in 1977, 1980, 1990, 1998, 2005 and 2012. The Economic Census covers all economic enterprises in the country except those engaged in crop production and plantation and provides data on employment in these enterprises[18]. This provides a frame for the conduct of more detailed follow up (enterprise) surveys covering different segments of the unorganized non-agricultural sectors, which in turn throw up data on production and employment in these segments, useful for an analytical study of these segments and in the compilation of national accounts.

iv) **The fourth source is the Employment Market Information (EMI) programme of the Directorate General of Employment & Training (DGE&T), Union Ministry of Labour & employment** and the **Directorate of Employment** under the State Governments. This is based on the statutory quarterly employment returns furnished by non- agricultural establishments in the private sector employing 10 or more persons and all public sector establishments. This source of industrial employment provides data at quarterly intervals down to district level. The quarterly data are available with a time lag of about a year. Detailed data at the level of three digit NIC 1987 codes are available in the Annual Employment Reviews based on this programme with a time lag of about two years. While data at national and State levels would be available from the National reports, district level data would be available from the State reports. This source or course covers only the organized sector.

v) The fifth source consists of the periodic unorganized sector surveys of the NSSO, which covers the unorganized non-agricultural enterprises. These surveys, usually conducted as follow-up surveys of the Economic Census, provide data on characteristics of the enterprises, its 5-digit NIC code, operating expenses, receipts, gross value added, fixed assets, liabilities and the number of workers working on a regular basis in the enterprise. Similar characteristics for the informal sector enterprises are also tabulated and published from these survey data by the NSSO. These data are tabulated and published at State levels, separately for rural and urban. Data by the districts can also be tabulated, with a suitable caution for the district level sample sizes, as the primary data collected (unit record data) can be obtained on CDs from NSSO[19]. Key results from the NSSO surveys are available after about six months, while the final report becomes available one year after the surveys are completed.

---

[17] You can see the rate list of NSSO unit level survey data from http://mospi.nic.in/ Mospi_New/upload/nsso/ratelist_UnitData.pdf.

[18] **The report on the Economic Census 1998** is available on the website of the Ministry of Statistical and Programme Implementation (MoSPI) www.mospi.nic.in . The website of MoSPI announces that the site's new address is www.mospi.gov.in . **The report on the Economic Census 2005** has been released by the Ministry on the 12th June, 2006, according to news item "Rural enterprises growing faster – Employment rate high in J & K: Economic Census 2005" appearing on page 11 of The Hindu, Chennai Edition, dated Tuesday the 13th June, 2006. The Ministry's website does not say anything about the 2005 report till 14/6/06.

[19] You can see the rate list of NSSO unit level survey data from http://mospi.nic.in/ Mospi_New/upload/nsso/ratelist_UnitData.pdf.

There are, however, other sources that provide data on production or on a variety of interrelated aspects of industry like employment, output, inputs and so on. Let us consider these sources.

22.3.2 Factory Sector – Annual Survey of Industries (ASI)

Collections of industrial statistics on aspects like output, employment, input and value added was initiated as early as 1944, when an annual **Census of Manufacturing Industries (CMI)** was launched, covering factories registered under the Indian Factories Act, 1934 and employing 20 or more persons and using power. Subsequently, a **Sample Survey of Manufacturing Industries (SSMI)** was started from 1949. While the former was restricted to only 29 of the 63 groups of industries, SSMI covered the remaining 34 groups of industries. As soon as the Factories Act, 1948 came into force, the coverage of both CMI and SSMI was extended to cover factories registered under the Factories Act, 1948 employing 10 or more persons and using power and those employing 20 or more persons without using power. The **CMI and SSMI reports** constituted the main source of data on different aspects of the factory sector up to 1958[20]. A brief coverage of the ASI can be seen at the site http://www.csoisw.gov.in/CMS/En/1024-asi-manual.aspx and at http://www.csoisw.gov.in/cms/cms/Files/572.pdf.

The annual Survey of Industries (ASI) replaced the CMI and the SSMI in 1960 after the Collection of Statistics Act, 1953 and the Collections of Statistics Rules, 1953 came into force. Detailed industrial statistics relating to industrial units in the country[21] like capital, output, input, value added employment and factor shares are collected in ASI. The survey was launched in 1960 and has been conducted every year since then except in 1972. From 2011, the ASI is conducted under the Collection of Statistics Act 2008 and the Collection of Statistics Rules 2011, except in the State of Jammu and Kashmir, where it is conducted under the State Collection of Statistics Act, 1961 and the rules framed thereunder in 1964. The frames used for the survey consists of:

a) all factories registered under Sections 2m(i) and 2m(ii) of the Factories Act, 1948 employing 10 or more workers using power as well as those employing 20 workers but without using power;

b) *bidi* and cigar manufacturing establishments registered under the *Bidi* and Cigar Workers (Conditions of Employment) Act, 1966 with coverage of units as in (i) above;

c) certain servicing activities like water supply, cold storage, repair of motor vehicles and other consumer durables like watches etc. Though servicing industries like motion picture production, personal services like laundry services, job of dyeing, etc., are covered under the Survey but data are not tabulated separately, as these industries do not fall under the scope of industrial sector defined by the United Nations. Defence establishments, oil storage and distribution depots, restaurants, hotels, café and computer services and the technical training institutes, etc., are excluded from the purview of the Survey.

[20] The ASI section of the CSO website.

[21] The coverage was extended to Jammu & Kashmir with the passing of the Collection of Statistics Act, 1961 (Jammu & Kashmir) and the notification of the Collection of Statistics Rules, 1964 under that Act.

The frames for the ASI are revised every year. At the time of revision, the deregistered factories are removed and newly registered factories are added in the frame. In ASI 2013-14, more than 2.65 lakh units were in the ASI frame, up from about 2.52 lakh units in ASI 2012-13. The "factory sector" thus consists of factories, *bidi* and cigar manufacturing units, electricity sector and some service units. Electricity units registered with the CEA, the departmental units such as railway workshops, Road Transport Corporation workshops, Government Mints, sanitary, water supply, gas storage, etc., are not covered from 1998-99 as there are alternatives sources of data for CSO for purposes of compiling GDP estimates in respect of these segments of industry.

The reference period for the survey is the accounting year April to March preceding the date of the survey. The sampling design and the schedules for the survey were revised in 1997-98, keeping in view the need to reduce the time lag in the availability of the results of the survey. The survey will not, however, attempt estimates at the district level. The survey used NIC 1987 for classifying economic activities before ASI 1997-98 and shifted to NIC 1998 with effect from ASI 1998-99[22]. From ASI 2009-10, the NIC 2008 codes are being used.

CSO has released the final results of ASI 2011-12 and provisional results of ASI 2012-13. Results in respect of selected characteristics are available in electronic media at various levels of aggregation[23]:

- All India by 4-digit level of NIC 2008,
- State by 3-digit level of NIC,
- Unit level with suppressed identification, etc.

All the reports of previous years can be accessed on the **website of the Ministry of Statistics and Programme Implementation (MoSPI)**. These results are:

1) **Time series data (1980-81 to 2007-08)** of principal characteristics[24] of the factory sector;
2) principal characteristics by major industry group;
3) principal characteristics by major States;
4) estimates of some important characteristics by States for 2002-03;
5) estimates of some important characteristics by 3-digit code of NIC 1998;
6) rural-urban break-up of principal characteristics; and

[22] NIC 1987 and NIC 1998 contain concordance tables to enable data users to recast data of earlier years from one NIC structure to another.

[23] You can download the ASI 2011-12 reports from the ASI portal of the CSO at http://www.csoisw.gov.in/CMS/cms/Home.aspx.

[24] The principal characteristics are: 1) no. of factories, 2) fixed capital, 3) working capital, 4) invested capital, 5) outstanding loans, 6) no. of workers, 7) mandays of workers, 8) no of employees, 9) mandays – employees, 10) total persons engaged, 11) wages to workers, 12) total emoluments, 13) old age benefits, 14) social security benefits, 15) other benefits, 16) fuels consumed, 17) total inputs, 18) products, 19) value of output, 20) depreciation, 21) net value added, 22), rent paid, 23) interest paid, 24) net income, 25) gross fixed capital formation (GFCF) 26) (a) material, fuels, etc., (b) semi-finished goods, (c) finished goods, 27) total, 28) gross capital formation (GCF), 29) profits.

7) principal characteristics by type of organization[25].
8) estimates for structural ratios and technical coefficients for factory sector from 2006-07.

Data users can find out from the **website (click on ASI DATA COST)** to find out what data – **tables or unit record data –**are available and the cost of obtaining the same from CSO.

CSO had earlier published **ASI 2010-11 final (detailed) results in two volumes**. These volumes present data at the 3-digit and 4-digit level of NIC 1998 codes for the national level and at the 3-digit level of NIC 1987 codes for individual States. Data on the following aspects of the industries forming part of the factory sector are presented:

i) Capital employed, input, output and gross value added (GVA),
ii) Employment, mandays and wages,
iii) Fuel consumed,
iv) Material consumed,
v) Products and by-products.

These volumes as well as those relating to earlier years are available also on electronic media. In addition, **CSO has also released time series data on ASI in 5 parts – Volumes I relates to ASI 1959-71, Volume II, Series A to ASI 1973-74 to 1981-82**, **Volume II, Series B to ASI 1982-83 to 1988-89**, **Volume III, Series A to ASI 1989-90 to 1993-94 and Volume III, Series B to 1994-95 to 1997-98**. These volumes present data on important characteristics for all-India at two-digit and three-digit NIC code levels and for the States at two-digit NIC code levels. These publications are **also available in electronic media on payment**.

The database on the **Annual Survey of Industries, 1973-74 to 2009-10** also provides time series ASI data on the principal characteristics of the factory sector, along with concepts and definitions used. The data **can be procured from the EPWRF on payment**.

The data available from ASI can be used to derive estimates of important technical ratios like capital-output ratio, capital-labour ratio, labour cost per unit of output, factor shares in net value added and productivity measures for different industries as also trends in these parameters. The most important use of the detailed results arises from the fact that these enable derivation of estimates of

i) the input structure per unit of output at the individual industry; and
ii) the proportions of the output of each industry that are used as inputs in other industries, enabling us to use the technique of input-output analysis to evaluate the impact of a change effected in (say) the output of an industry on the rest of the economy. The construction of the Input-Output

[25] Types of organization are 1) proprietorship, 2) joint family (HUF), 3) partnership, 4) public limited company, 5) private limited company, 6) govt departmental enterprises, 7) public corporations, 8) corporate sector (4+5+6+7), 9) *Khadi*& Village Industry, 10) handloom industry, 11) cooperative sector, 12) others (including 'not recorded').

Transaction Tables (I-OTT) for the Indian economy is largely based on ASI data. As we know, I-OTT provides the basic ratios required for Input-Output Analysis, which has a wide range of uses, from analysis of economic structure and backward and forward linkages of different industries to the setting up of targets of production, investment, employment and so on at the macro level.

22.3.3 Monthly Production of Selected Industries and Index of Industrial Production (IIP)

CSO prepares and releases **monthly indices of industrial production (IIP)** and also the **monthly use-based index of industrial production (base year 2004-05)**. IIP with base year 2004-05 was first released in April 2005. The base year for IIP was earlier 1993-94[26]. IIP with the new base year has (i) a wider coverage of items than the IIP with base year 1993-94, (ii) the weighting diagram of the new index has taken into account the contribution of the unorganized manufacturing sector, (iii) small scale industry items have been given individual weights in the weighting diagram of the new index, (iv) the revised IIP series follows NIC 2004 (the earlier one followed NIC 1987) and (v) the new IIP is released within six weeks of the reference month. The IIP (2004-05) is a quantitative index based on production data received from 16 source agencies covering 682 items clubbed into 399 item groups in the basket of items of the index.

The Directorates of Statistics and Economics of the State Governments and Union Territory Administrations (DESs) had been preparing **IIPs for their respective areas**. These IIPs were not comparable with each other or with that prepared by CSO for the country as a whole because of differences in the base year, basket of items, data and methodology used for the construction of the indices. The question of preparing State-wise IIPs that are comparable with the national IIP was, therefore, examined and State Governments and Union Territory Administrations have initiated work I this regard on the basis of the recommendations made in this regard. The preparation and release of IIP is at different stages in different States and Union Territories. For example the **Governments of Tamil Nadu**, **Andhra Pradesh**, **West Bengal**, **Assam**, **Goa**, **Punjab**, **Haryana**, **Rajasthan and Maharashtra have released the new IIP prepared** as a result of these efforts and **others are in the process of preparing such IIPs**.

CSO releases Quick Estimates of IIP within six weeks of the closing of the reference month. CSO has released the **IIP for December, 2014** through the **Press Release dated 12th February 2015**. The Release provides estimates of

- monthly IIP – overall index and for sectors (mining, manufacturing and electricity) – for the period April, 2014 to December, 2014;
- monthly IIP at 2-digit level NIC 2004 code for December 2013 and December 2014 and the cumulative index for the period April-December 2014; and

[26] Change in the base of IIP is necessary to measure the real growth in the industrial sector. The United nations Statistical Office (UNSO) recommends that the base year should be changed quiquennially. IIP was first prepared in India with the base year as 1937. The base year was then changed to 1946, 1951, 1956, 1960, 1970, 1980-81, 1993-94 and 2004-05.

- monthly IIP use-based index (basic goods, capital goods, intermediate goods, consumer goods, consumer durables, consumer non-durables) for the period April, 2014 to December, 2014.

CSO follows the SDDS norms of the International Monetary Fund (IMF)[27] in respect of estimates of IIP and, accordingly, the Press Release of the 12th February 2015 announces that IIP for January 2015 would be released on the 12thMarch 2015. **This Press Release is accessible in the MoSPI website**. The website also has time series data at 2-digit level NIC code on (a) monthly IIP from April, 1994 (b) annual averages from 1994 and also on the monthly use-based IIP. Similar time series data on (a) monthly IIP with base year 1980-81 at 2-digit level of NIC 1970, and (b) monthly use-based IIP (base year 1980-81) are also available from April,1990.

The Reserve Bank of India's (RBI) "Handbook of Statistics on the Indian Economy" also presents IIP at two-digit-level, the use-based IIP and index numbers of Infrastructure Industries[28] as also data on production in selected industries. Monthly statistics of mineral production are published by **the Indian Bureau of Mines (IBM), Nagpur in their publication "Monthly Statistics of Mineral Production"**. It also releases mineral group-wise and State-wise value of mineral products. The latest data available while writing this chapter was from March 2013[29]. The **CSO publication "Energy Statistics"** bring together important data on different sources of energy in the country at one place. It presents time series data that give a broad picture of the trends in production, consumption and price indices of major sources of conventional energy in the country for last 30 years. **The ministries of Petroleum and Gas, Power and Non-Conventional Energy** provide information in their respective spheres of activity. The **Economic Intelligence Services (EIS) of CMIE, Mumbai** has volumes on **Energy and Infrastructure** that present detailed data on the trends in these sectors. A visit to the **EPWRF website** (click with words EPWRF through the Google search engine) would also be rewarding in terms of time series data on industrial production. At the global level, the International Energy Agency (IEA) is considered as one of the most authoritative organisations in dealing with energy issues. The IEA publishes the world Energy Outlook. The latest one, World Energy Outlook 2014, was published in November 2014. The IEA portal http://www.iea.org/statistics/ also provides for search of statistics by country. IEA online data services provide monthly data on oil and gas, quarterly energy prices, etc. However, data from this site are to be purchased, be it hard copy, CD-RoM or pdf files. The UNSD also compiles and publishes energy statistics in a comparable manner for 224 countries of the world. The reports, although a bit dated, can be downloaded free of charge from their website[30].

---

27 See Indian Statistical System in Unit 18 on "Macro Variable Data: National Income, Saving, Investment".

28 Infrastructure industries are electricity, coal, steel, cement, crude petroleum and refined petroleum products. These industries are part of the basket of items of IIP and the Index for Infrastructure Industries can be computed from the indices for the individual infrastructure industries.

29 These can be accessed from the IBM website http://ibm.nic.in/msmpmar13.htm.

30 You can visit the link http://unstats.un.org/unsd/energy/yearbook/2011/004-10.pdf and read the introduction to the Yearbook to have an idea of the data being made available by them.

Check Your Progress 3

1) Indicate the major sources of data on levels of industrial employment. Which one is the most and which one the least frequent? Which of the sources shall you use for your analysis, if the analysis are to be done i) for backward districts and ii) for major States of India?

……

……

……

……

……

2) How can you access the report of the economic census 1998?

……

……

……

……

……

3) What do you understand by quarry based system developed by CSO?

……

……

……

……

……

4) List two important uses of ASI data.

……

……

……

……

……

5) From which document can you have data pertaining to various sources of energy?

……

……

……

……

……

22.3.4 Data on the Unorganised Industrial Sector

The Development Commissioner for Small Scale Industries (DCSSI) in the Ministry of Small Scale Industries, in the Central Government and the Directorates of Industries of State Government and Union Territory Administration provide data on small-scale industrial units registered with the latter set of agencies. The DCSSI has conducted a census of small scale industrial units thrice – in 1973-74 (reference year 1972), 1990-91 (reference year (1987-88) and in November, 2002 (reference year 2001-02). The results of the third census can be seen in the publication **"Final Results: Third all India Census of SSI – 2001-02" released by the Ministry of Small Scale Industries**. Broad details of the performance of small-scale industries are available in the Annual Reports of the Ministry of Small Scale Industries. **Time series data on employment, production, labour productivity in small-scale industries (SSI)** and value of exports of the products of small-scale industry are also available in the **RBI Handbook**. Data on some parts of ***Khadi* and Village Industries Commission (KVIC)**, handlooms and handicrafts do get included in ASI but data relating exclusive to these sub-sectors are available in the Annual Reports of these organizations or in the **Annual Reports of the Ministries under which these Boards/Commissions function**.

Surveys of unorganized sector enterprises conducted once in five or six years contribute to the strengthening of the database of the unorganized sector. The following **rounds of the NSSO** provide data on input, output, value added, fixed assets, liabilities, number of working owner, hired worker and other worker (most of whom would be unpaid family workers) and related data on the unorganized non-agricultural sector. You can become a user in the website of the MoSPI and download relevant reports for your research, free of cost. You can also purchase relevant micro data from the MoSPI.

NSS Round No.	Reference Year	Subject coverage
67th	July 2010 – June 2011	Unorganised Non-agricultural Enterprises (excluding cconstruction)
63rd	July 2006 – June 2007	Unorganised service sector enterprises (excluding construction)
62nd	July 2005 – June 2006	Unorganised manufacturing Enterprises
57th	July 2001 – June 2002	Unorganised service sector enterprises (excluding construction and finance)
56th	July 2000 – June 2001	Unorganised manufacturing Enterprises
55th	July 1999 – June 2000	Informal sector non-agricultural enterprises (excluding finance)
53rd	January – December 1997	Unorganised trading Enterprises
51st	July 1994 – June 1995	Unorganised manufacturing Enterprises
46th	July 1990 – June 1991	UnorganisedNDTE and DTE (trading) Enterprises
45th	July 1989 – June 1990	Unorganised manufacturing Enterprises

At the global level, the UN International Labour Organisation (ILO) supports surveys on employment and child labour related issues. The Laborsta database and the Key Indicator of Labour market (KILM) of the ILO contains comprehensive country level comparable estimates on labour force[31], employment and unemployment, forced labour, informal economy and labour migration related data. The UN Industrial Development Organisation (UNIDO) has recently developed the world productivity database, providing total factor productivity and related indicators.

We have looked at data sources that provide data on production and also on certain related variables like capital employed, employment etc. How is the capital financed? Let us look at some of the sources that throw light on these matters.

22.3.5 Industrial Credit and Finance

The RBI Handbook on Statistics of the Indian Economy provides time series data on the sectoral deployment of non-food gross bank credit provided by Scheduled Commercial Banks to different sectors of the economy which enables you to study trends in the flow of bank credit to small-scale industry, medium and large industries, wholesale trade other than food production and export credit. This would enable you to evaluate the implementation of the policy regarding allocation of bank and institutional credit to the priority sector, which includes small-scale industries. It also provides time series data on deployment of bank credit to some 25 industrial categories and power. In addition, it presents time series data on the health of SSI and non-SSI units – (a) the number of SSI units that are sick, (b) the number that are weak, (c) the number of non-SSI units that are sick, and (d) the amounts outstanding (loans) from each of these categories of units.

The **ASI** provides, as you have seen, some data on financial aspects of industries – fixed capital, working capital, invested capital, loans outstanding and also the interest burden of industrial units (up to the 4-digit NIC code level). From where and how have the industries raised capital needed by them? We have looked at one source of capital or working capital, namely, bank credit. Time series data on new capital issues and the kinds of shares/instruments issued (ordinary, preference or rights share or debentures, etc.,) and the composition of those contributing to capital (like promoters, financial institution, insurance companies, government, underwriters and the public) are also presented. Data on assistance sanctioned and disbursed by financial institutions like Industrial Development Bank of India (IDBI), Industrial Credit and Investment Corporation of India (ICICI), Small Industrial Development Bank of India (SIDBI) and Life Insurance Corporation (LIC) of India can also be obtained from the **Handbook**. Similarly, some data on the financing of project costs of companies are available in the **Handbook**. All these data of course relate to companies, which may include non-industrial ventures too. The primary source of such data is the Ministry of Company Affairs. The publication of the Securities Exchange Board of India (SEBI) **"Handbook of Statistics on the Indian Securities Market"(the latest one published in 2012)** provides annual and monthly time series data on industry-wise classification of capital raised through the securities market – the industrial sector activities by which the classification is being made are, cement

[31] See the website http://labordoc.ilo.org/ .

and construction, chemical, electronics, engineering, entertainment, finance, food processing, health care, information technology, leather, metal mining, packaging, paper and pulp, petrochemical, plastic, power, printing and rubber. A reference to the two volumes of **EIS, CMIE,** one **"Industry: Financial Aggregates"** and the other **"Industry: Market Size and Shares"** would be rewarding.

Foreign direct investment (FDI) is another important source of capital finance for industrial expansion. Availability of data on FDI is dealt with in Unit 22 on Trade and Finance.

22.3.6 Contribution of Industrial Sector to GDP

The National Accounts Statistics (NAS), as we have seen in UNIT 18 on National Income, Saving, Investment, furnishes information on the contribution of the industrial sector to GDP. **NAS** presents a short time series of estimates of (i) value of output and GDP of each two-digit NIC code level industry in the registered and the unregistered sub-sector of the manufacturing sector, (ii) value of output of major and minor minerals and GDP and NDP of the mining & quarrying sector, and (iii) GDP and NDP of the sub-sectors electricity, gas and water supply.

Check Your Progress 4

1) Name the agencies which generate the data on small scale industries.

   ……..

   ……..

   ……..

   ……..

   ……..

2) How can you get the data on deployment of bank credit to SSI units?

   ……..

   ……..

   ……..

   ……..

   ……..

3) From which sources you can get the data on credit flowing to small scale sector?

   ……..

   ……..

   ……..

   ……..

   ……..

22.4 LET US SUM UP

The major efforts to collect data on different aspects of the agricultural sector are the quinquennial agricultural census, the quinquennial livestock census, the cost of cultivation studies, annual estimates of crop production and the integrated sample survey to estimate the production of major livestock products. Most of the data flowing from these efforts relate to production, inputs for production, land utilization patterns, intensity of land use, irrigation, costs and pricing of agricultural produce, procurement and distribution, subsidies and credit are found in the two publications **"Agricultural Statistics at a Glance" (ASG)** and **"Basic Animal Husbandry Statistics" (BAHS)**. Statistics on the characteristics of land holdings of different size with reference to the parameters referred to above are also available in **ASG**. More detailed information in this regard, collected in the succcessive **agricultural census** are available in the **Ministry's website**. Similarly, detailed information relating to sub-sectors – livestock, poultry and fishery as also agricultural implements, collected in the **livestock censuses, are available in the website of the Ministry**, though some information relating to the earlier census are available in **ASG** and **BAHS of relevant years**. Data on credit flowing to the agricultural sector are available in the **RBI Handbook of Statistics on the Indian Economy** and **NABARD's Publications**. Another source containing comprehensive information on agriculture at the global level is the FAO statistical yearbooks published by the UN Food and Agriculture Organisation.

The most detailed data available in the industrial sector relates to the factory sector provided by **ASI**. The multi-variable data covers about 30 characteristics – investment, inputs, output, gross and net value added, employment, factor shares, net value added at 3 or 4-digit NIC code levels by States and Union territories. Quick estimates of monthly production in selected industries and the Index of Industrial Production are available within six weeks of the reference month from the **CSO** and revised in the next two months. The State Governments and Union Territory Administrations are also taking steps to prepare and release monthly IIPs that are comparable to the national level IIP. Some have already started releasing such estimates. The **fourth census of micro, small and medium enterprises (MSME) relating to 2006-07** is the most recent and comprehensive source for data on production, employment and exports of the small-scale sector. Reports from the NSSO on surveys of the unorganized sector carried out once in five or six years provide useful data on the characteristics of this sector. Enquiries covering the entire industrial sector or the organized industrial sector such as the **population census, NSSO surveys on employment**, the **Economic Census** and the **Employment Market Information** Programme of the Ministry of Labour & Employment provide data only on employment. The **RBI handbook** gives data on industrial credit and financing of industrial capital. The **CMIE (EIS) volumes on industry**, one on **financial aggregates** and the other on **market size and shares** provide a fund of data on finance and market competition aspects of industry. At the global level, the International Labour Organisation (ILO) provides statistics on employment and the UNIDO provides statistics on total factor productivity.

22.5 EXERCISES

1) Which major efforts have been made to collect the data on different aspects of agricultural sector?

2) Discuss the characteristics of data flowing from the agencies involved in compilation of agriculture data.

3) "The detailed data available in the industrial sector relates to the factory sector" explain.

4) Do you think that data available on unorganized sector is inadequate? What suggestion would you like to make in this regard?

22.6 SOME USEFUL BOOKS

Amitabh Kundu & Alakh Sharma (Ed.) (2001) : Informal Sector in India – Perspectives & Policies, Institute for Human Development & Institute of Applied Manpower Research, IP Estate, Ring Road, New Delhi.

____National Sample Survey Office (NSSO) (Various issues) : Survey Reports of unorganized sector enterprise surveys **NSSO**, Ministry of Statistics & Programme Implementation, Govt of India, Sardar Patel Bhavan, New Delhi.

Department of Agriculture & Coop : Agricultural Statistics At a Glance, Directorate of Economics & Statistics (DES), Ministry of Agriculture, Govt. of India, Krishi Bhavan, New Delhi. Cost of Cultivation in India, DES, Ministry of Agriculture & Cooperation, New Delhi.

Reports of the Commission on Agricultural Costs and Prices (CACP), DES, Ministry of Agriculture & Cooperation, New Delhi.

Indian Horticulture Data Base, Ministry of Agriculture & Cooperation, New Delhi

___Ministry of Finance : Economic Survey, Ministry of Finance, North Block, New Delhi.

___Reserve Bank of India (various issues) : RBI Handbook of Statistics on the Indian Economy. The Reserve Bank of India, Mumbai

www.mospi.nic.in : Website of the Ministry of Statistics & Programme Implementation. (even the letters cso in google search will lead you to the (CSO website).

eands.dacnet.nic.in : Directorate of Economics and Statistics, Department of Agriculture and Cooperation, Ministry of Agriculture, Government of India

www.wrmin.nic.in : Website of the Ministry of Water Resources.

www.rbi.org.in : **W**ebsite of the Reserve Bank of India (RBI).

www.fao.org : UN Food and Agriculture Organisation (FAO) website.

www.ilo.org	**:**	UN International Labour Organisation (ILO) website.
www.unido.org	**:**	UN Industrial Development Organisation (UNIDO) website.

22.7 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) The Directorate of Economic and Statistics in the Department of Agriculture and Statistics and Animal Husbandry Statistics Division of the Department of Animal Husbandry.
2) See foot note no. 2 and 4.
3) See Sub-section 22.2.3 and foot note no. 6.

Check Your Progress 2

1) See Sub-section 22.2.4 and 22.2.5
2) See foot note no. 8.
3) Agricultural Statistics at a glance.
4) RBI Handbook of Statistics on Indian Economy 2005.

Check Your Progress 3

1) See Sub-section 22.3.1
2) See foot note no. 18
3) A system that enables the data uses to access specific data relating to ASI.
4) Derivation of important technical ratios for using the principal characteristics of factory sector.
5) Energy Statistics published by CSO.

Check Your Progress 4

1) Development commissioner for Small Scale Industries (DCSSI).
2) Directorate of Industries of State Government and Union Territory Administrations.
3) SIDBI, LIC, Handbook of statistics on the Indian Security Market.

UNIT 23 TRADE AND FINANCE

Structure

23.0 Objectives

23.1 Introduction

23.2 Trade

- 23.2.1 Merchandise Trade
- 23.2.2 Services Trade

23.3 Finance

- 23.3.1 Public Finances
- 23.3.2 Currency, Coinage, Money and Banking
- 23.3.3 Financial Markets

23.4 Let Us Sum Up

23.5 Exercises

23.6 Some Useful Books

23.7 Answers or Hints to Check Your Progress Exercises

23.0 OBJECTIVES

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

- describe the kind of data available in India in the area of trade and finance;
- state the various sources of data on trade and finance;
- explain the reasons for divergence between merchandise trade deficit/surplus data provided by DGCI&S and RBI's BOP data; and
- discuss which agencies are involved in compilation of data on trade and finance.

23.1 INTRODUCTION

Trade is the means of building up an enduring relationship between countries and the means available to any country for accessing goods and services not available locally for various reasons like the lack of technical know-how. It is also the means of earning foreign exchange through exports so that such foreign exchange could be utilized to finance essential imports and to seek the much-needed technical know-how from outside the country for the development of industrial and technical infrastructure to strengthen its production capabilities. Trade pacts or agreements between countries or groups of countries constitute one way of developing and expanding trade, as these provide easier and tariff-free access to goods from member countries. While efforts towards such an objective will be of help in expanding our trade, globalization and the emergence of World Trade Organization (WTO) have only sharpened the need to ensure efficiency in the production of goods and services to compete in international markets to improve our share of world merchandise trade and trade in services. Trade is also closely tied up with our development objectives. Trade deficit or surplus, made up of deficit/surplus in merchandise trade and trade in services, contributes to current account deficit

or surplus. Data on trade in merchandise and services would enable us to appreciate the trends and structure of trade and identify areas of strength and those with promise but need sustained attention.

While trade and finance have been closely bound up with each other ever since the time money replaced barter as the means of exchange, finance is the lifeline of all activities – economic, social and administrative activities. It flows from the public as taxes to Government, as savings to banking and financial institutions and as share capital or bonds or debentures to the entrepreneur. It then gets used for a variety of developmental and non-developmental activities through Government and other agencies and flows back to members of the public as incomes in various ways, as factor incomes. It would therefore, be of interest to know how funds get mobilized for various purposes and get used. This Unit looks at the kind of data available that could enable us to analyse this mobilization process and the flows of funds to different areas of activity. Let us begin to discuss the data on trade.

23.2 TRADE

23.2.1 Merchandise Trade

The Directorate General of Commercial Intelligence and Statistics (DGCI&S) collects and compiles statistics on imports and exports. It releases these data at regular intervals through their **publications** and through **CDs**. It prepares **"Quick Estimates"** on aggregate data of exports and imports and principal commodities within two weeks of the reference month and releases these in the monthly press release. It publishes

1) A monthly brochure **"Foreign Trade Statistics of India (Principal Commodities and Countries)"** containing provisional data issued to meet the urgent needs of the Ministry of Commerce, other government organizations, Commodity Brands (CBs), Export Promotion Councils (EPCs) and research organizations. It contains commodity-wise, country-wise and port-wise foreign trade information;

2) **"Monthly Statistics of Foreign Trade of India, Volume I (Exports including re-exports) & Volume II (Imports)** containing detailed data on foreign trade at the 8-digit level codes of the ITS (HTS) (see below);

3) **Quarterly publications :**

   a) **Statistics of the Foreign Trade of India by Countries – VolI (Exports including re-exports) and**

   b) **Statistics of the Foreign Trade of India by Countries – Vol II (Imports)**

4) **Annual publications:**

   a) **Statistics of the Customs and Excise Revenue Collections of the Indian Union,**

   b) **Statistics of the Inland Coasting Trade Consignment of India,**

   c) **Inter-State Movements/ Flows of Goods by Rail, River and Air,**

   d) **Statistics of Foreign and coastal Cargo Movements of India,**

   e) **Selected Statistics of the Foreign Trade of India**;

5) The **DGCI&S website (www.dgciskol.nic.in)** has two parts, one consists of static pages and the dynamic pages. The first contains the history and the activities of DGCI&S and summary data on principal commodities, and countries and is updated regularly. The dynamic pages are mainly for on-line data dissemination and provide access on free and payment basis. This contains at least 24 months final foreign trade data at 8-digit commodity and principal commodity level and updated regularly; and

6) **The Priced Information Service System (PISS)** provides information to private parties, EPCs, CBs, Foreign Embassies etc., on payment basis @ Re. 1 /- per unit record of information. It does not give the whole basket of 8-digit commodity-country data for any particular period for reasons of 'copyright' provision of the DGCI&S. It however, provides aggregate and detailed data to **Centre for Monitoring Indian Economy (CMIE), Mumbai** for an efficient trade intelligence service.

7) It also has a web-based data dissemination system for online data transaction through advance payment. The export and import trade data are available countrywise and economic regionwise. The detailed data on India's foreign trade in merchandise are made available after completion of the third month from a particular month. Here the commodities referred to in these data stand for the ones specified in the ITC(HS) against 8-digit codes which have been developed by the DGCI&S by sub-dividing the 6-digit codes of Harmonised System as internationally standardised. If you want to get data online, you can visit http://www.dgciskol.nic.in/new_ registration.asp and see the terms, conditions and procedures.

The DGCI&S data are also presented as time series data in the **Reserve Bank of India Handbook of Statistics on the Indian Economy**. As mentioned above the **CMIE** is another source – **their volume on Foreign Trade and BoP**. The **EPWRF (see its website** – google search with **EPWRF** will enable you to access it) also publishes time series data on foreign trade – it is one of the 35 sets of special statistics on which it publishes long time series data along with information on conceptual and methodological issues.

Foreign trade data published by DGCI&S relates to merchandise trade through all recognized seaports, airports, land customs stations and inland container depots located all over India. Data on exports include re-exports and relate to the free on board (f.o.b.) values and imports relate to cost, insurance and freight (C.I.F.) values. Exports and imports are based on a general system of recording. According to this, exports relate to Indian merchandise and re-export relates to foreign merchandise previously imported into India. Imports relate to foreign merchandise, whether these are intended for consumption in India, bonding or re-exportation.

The commodities comprising merchandise imports and exports are classified according to a standard classification. The trade classification in vogue in India between April, 1977 and March, 1987 was the Indian Trade Classification, Revision 2 (ITC – Rev. 2) – one that was based on the Standard International Trade Classification Revision 2 (SITC – Rev.2). A new system of commodity classification, known as Indian Trade Classification (based on the Harmonized Commodity Description and Coding System), or ITC(HS), has been adopted since April, 1987. ITC (HS) is an extended version of the International Classification System called "Harmonized Commodity Description and Coding

System" evolved by the World Customs Organization, Brussels[1]. These changes in the trade classification of commodities mean that time series data on export and import relating to some commodities may not be strictly comparable. Another element of non-comparability of time series data arises from changes in the definition of countries and/or groups. For example, data for Russia prior to 1993-94 relate to erstwhile USSR, with the exception of 1992-93, the data for which relate to the Commonwealth of Independent States (CIS) representing a group of 15 countries[2]. Similarly, Indian trade with Germany relates to Federal Republic of Germany (FRG) till 1989-90 and to the unified Germany thereafter. There are also changes in the membership of groups like the European Union, and Oil and Petroleum Exporting Countries (OPEC) from time to time.

What would we like to know about foreign trade? The volume of trade, that is, the volume of exports and imports, the size of export earnings, the expenditure on imports, the size of exports relative to imports, earning from exports compared to expenses incurred on imports since exports earn foreign exchange while imports imply outflow of foreign exchange. We should like to know about the trends in these variables. Besides looking at the trends in the quantum and value of imports and exports, it is important to analyse the growth in foreign trade both in terms of value and volume, since both are subject to changes over time. Exports and imports are made up of a large number of commodities and fluctuations in the export and imports of individual commodities contribute to overall fluctuations in the volume and value of exports and imports. We, therefore, need a **composite indicator of the** trends in trade. The index number of foreign trade of a country is a useful indicator of the temporal fluctuations in exports and imports of the country in the term of **value, quantum and unit price and** so on. Similarly, measures of the terms of trade could be derived from such indices relating to imports and exports.

The index number of foreign trade is computed and presented as Unit Value Index (UVI) and Quantum Index (QI). These are defined as follows:

$$\text{UVI} = [\sum P_1 Q_t] / [\sum P_0 Q_t] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

$$\text{QI} = [\sum P_0 Q_t] / [\sum P_0 Q_0] \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

where P_1 Is the unit value of an item in the current period and Q_t is the quantity of the same item in the current period, P_0 and Q_0 are the unit value and the quantity respectively of the same item during the base period and Σ is the summation over all commodities. These indices have been computed with 1979-80 as the base year.

Three types of Terms of Trade are computed from these indices:

1) Gross Terms of Trade (GTT) = 100 × [(QI of imports)/QI of exports)]
2) Net Terms of Trade (NTT) = 100 × [(UVI of exports)/(UVI of imports)]
3) Income Terms of Trade (ITT) = [NTT × QI of exports]/100
   = [UVI(exports)× QI(exports)]/ [UVI (imports)].

The existing index numbers have the base year 1979-80. Changes in the ITC since April, 1987 and the recasting of the basket of commodities to suit the new

[1] This organization was previously known as Customs Cooperation Council.

[2] These were Armenia, Azerbaijan, Belarus, Estonia, Georgia, Kazakhstan, Kyrgyz republic, Latvia, Lithuania, Moldova, Russia, Tajikistan, Turkmenistan, Ukraine and Uzbekistan.

classification system, the base year 1979-80 has become too old to serve the purpose of temporal comparability and the new ITC adopted in April, 1987. The DGCI&S, therefore, decided to construct new index numbers with the **base year 1998-99**. It also decided to bring out global monthly indices and for selected countries. The preparation of these index numbers is in progress.

What kinds of data are available on foreign trade? Time series on the following data are published in the **RBI Handbook of Statistics of the Indian Economy 2013-14, under the sub-heading "Trade and Balance of Payments"**:

a) India's foreign trade (in US $ and Indian Rupees) from 1975-76 to 2013-14, The exports and imports have been shown as total and in two groups, namely, oil and non-oil. The "Trade Balance", i.e., exports – imports is also given for each of these three groups. A negative figure meaning exports are less than imports. You can observe from this Table that India had a positive non-oil trade balance, albeit sporadically, till 2003-04. However, after that, trade balance for both oil and non-oil has become negative.

b) Exports (in US$ and Indian Rupees) of principal commodities, from 1998-99 to 2013-14[3];

c) Imports (in US$ and Indian Rupees) of Principal Commodities[4], from 1998-99 to 2013-14;

d) Exports (in US$ and Indian Rupees) of selected commodities to principal countries;

e) Direction of Foreign Trade (in US$ and Indian Rupees) showing exports and imports for each year by trade areas, groups of countries and countries; and within each group of countries or trade area, data are presented only for selected countries[5];

f) Year-wise indices, both UVI and QI, for imports and exports from 1976-77 to 2007-08 (base 1978-79 = 100) and 1999-2000 to 2013-14 (base 1999-2000 = 100) the three terms of trade measures; GTT, NTT and ITT;

---

[3] Commodities were classified into 4 groups (including total), namely, primary products, manufactured goods, petroleum products and others; sub-divided further in into 11 sub-groups and 44 items: A. Agriculture and allied Products (15 items), 2. Ores and Minerals (3 items), 3. Leather and manufactures, 4. Chemicals and related products (4 items), 5. Engineering goods (6 items), 6. Textile and textile products (11 items), 7. Gems and jewelry, 8. Handicrafts, excluding handmade carpet; 9. Other Manufactured Goods; 10. Petroleum products; and 11. Others.

[4] The classification of imports consisted of three broad groups, I Bulk Imports, II Non-bulk Imports and III Total Imports; Bulk Imports were further divided into three groups, namely, A. Petroleum, Crude and Products; B. Bulk Bulk Consumption Goods (consisting of 4 sub-items); C. Other Bulk Items (9 sub-items); Non-bulk Import were A. Capital Goods (8 sub-items), mainly export related items (4 sub-items) and Others (9 sub-items).

[5] The trade areas for which data are presented are: I OECD countries, II OPEC, III Eastern Europe, IV Developing Countries, V Others (Unspecified). Each of these trade areas is further divided into country-group and within each group figures are given only for some selected countries. Trade area I is divided into A. European Union (within which data for 6 selected countries are given). B. North America (2 selected countries). C. Asia & Oceania (2 selected countries) and D. Other OECD Countries (1 selected country). II and III have no further grouping of countries within them but figures for 6 selected countries under II and for 2 selected countries under III are presented. Trade area IV is divided into A. Asia, B. Africa and C. Latin America Countries. A gets further grouped into (a) SAARC (figures for all countries are given under this), (b) Other Asian Developing Countries (figures for 6 selected countries are given). Figures for 7 selected countries under Africa are given and for Latin America no further break up by countries are given.

g) Index numbers of Exports – QI and UVI for 9 commodity groups[6] from 2005-06 onwards (base 1999-2000 = 100).

h) Index numbers of Imports - QI and UVI for 9 commodity groups[7] from 2005-06 onwards (base 1999-2000 = 100) Similar type of data on imports.

i) India's overall Balance of Payment (in US $ and Indian Rupees) under: A current account, B capital account, C. errors and omissions, D. overall balance and E. Monetary movements from 2008-09 onwards.

j) Invisibles by category of transactions (in US $ and Indian Rupees) from 1994-95 onwards.

k) Exchange rate of Indian Rupee vis-à-vis US $, UK £, DM/€ and Japanese ¥ from 1975 onwards (calendar year – annual average, financial year – annual average and end year rates).

l) Real Effective Exchange Rate (REER) and Net Effective Exchange Rate (NEER) of Indian Rupee 36-country bilateral weights) with base 2004-05 from calendar year 2005 onwards.

m) Indices of REER and NEER.

n) External assistance in the form of loans and grants from 1985-86 onwards (in US $ and Indian Rupees), authorization, utilization and debt service payments.

o) NRI deposits outstanding (in US $ and Indian Rupees) from 1996 onwards in different types of accounts like NR (E)RA, FCNR(A), FCNR(B), NR(NR)RD, NRO, FC (B&O)D, FC(O)N and total. It has increased from US $17.446 billion in 1996 to US $103.844 in 2014.

p) Foreign investment inflows (in US $ and Indian Rupees) namely, gross investments, repatriation/ disinvestment, FDI by India, net foreign direct investment, net portfolio investment from 2000-01 onwards.

q) Foreign exchange reserves (in US$ and Indian Rupees) from 1956-57 onwards.

r) India's external debt (in US $ and Indian Rupees) from 1996 onwards[8], which, inter alia, providesconcessional debt as percentage of total debt, short term debt as % to total debt, debt stock-GDP ratio (%) and debt service ratio (%).

---

[6] The commodity groups are: I Food and live animals (6 sub-groups), II. Beverages and Tobacco, III. Crude materials, inedible, except fuel (4 sub-groups), IV. Mineral fuels, lubricants and related materials (3 sub-groups), V. Animal and vegetable oil, fats and waxes, VI. Chemicals and related products (6 sub-groups), VII. Manufactured goods classified chiefly by material (5 sub-groups), VIII. Machinery and transport equipment (9 sub-groups) and IX. Misc. manufactured articles (4 sub-groups). In a few cases, indices for specific commodities under a sub-group are also given.

[7] The commodity groups are: I Food and live animals (5 sub-groups), II. Beverages and Tobacco (2 sub-groups), III. Crude materials, inedible, except fuel (7 sub-groups), IV. Mineral fuels, lubricants and related materials (3 sub-groups), V. Animal and vegetable oil, fats and waxes, VI. Chemicals and related products (8 sub-groups), VII. Manufactured goods classified chiefly by material (7 sub-groups), VIII. Machinery and transport equipment (9 sub-groups) and IX. Misc. manufactured articles (3 sub-groups). In a few cases, indices for specific commodities under a sub-group are also given.

[8] The debts are classified into I. Multilateral, II. Bilateral, III. International Monetary Fund, IV. Trade credit, V. Commercial borrowing, VI. NRI & FC (B&O) deposits, VII. Rupee debt, VIII. Total long-term debt, IX. Short-term debt and X. Gross total debt. Most of the groups have sub-groups A. Government borrowing, and B. Non-Government borrowing, further disaggregated into concessional and non-concessional borrowing.

Some of these data are from the DGCI&S and the rest are from the **Balance of Payments (BoP)** data of the RBI. The **BoP** data reported by RBI show the value of merchandise imports on the debit side and that of exports on the credit side. It also shows trade balance – a trade deficit or a trade surplus – depending upon whether the difference 'export - imports' is negative or positive. These are all shown in the balance payment format as part of current account, which also shows another entity 'invisibles'. However, there is a divergence in trade deficit/surplus in merchandise trade shown by **DGCI&S data** and that shown by **RBI's BoP data**. This discrepancy between the two sources also affects data on current account deficit (CAD) or surplus (CDS), since current account deficit/surplus is the total of trade deficit/surplus and net invisibles (inflow of invisibles net of outflow in the category 'invisibles'[9]) For example, the **Economic Advisory Council to the Prime Minister noted, in its Report on Balance of Payments (BoP)** submitted to the Prime Minister recently, that the divergence between the two sources of data on trade was growing[10]. It noted that trade deficit is projected for the year 2005-06 at 5.2 per cent of GDP on the basis of trade data from DGCI&S, while it is 7.7 per cent according to RBI's BoP data – a difference of the size of 2.5 per cent of GDP. CAD based on trade data of DGCI&S is only 0.3 per cent of GDP while it is 2.9 per cent of GDP if the estimate of CAD is based on trade data from BoP. The reasons for the divergence in the data between the two sources are, as noted by the Council:

- DGCI&S tracks physical imports and exports while BoP data tracks payment transactions relating to merchandise trade;
- DGCI&S data do not capture Government imports, which are exempted from customs duty. Defence imports fall into this category; and
- DGCI&S data do not capture imports that do not cross customs boundary (for example, oil rigs and some aircrafts) while they are still paid for and get captured in BoP data.

23.2.2 Services Trade

Besides export and import of merchandise, a number of services, like transportation services, travel services, software, Information technology-Enabled Services (ITES), business services and professional services are exported and imported. These are captured by "non-factor services" included in the entry "Invisibles" in the **Tables on India's Overall BoP and on Key Components of India's BoP in the RBI Handbook**. The handbook has also a table (Table 140 in RBI's Handbook of 2013-14) that gives the distribution of 'invisibles' by transactions – credit, debit and net, under the Current Account, giving separately for a) Services: Travel, Transportation, Insurance, G.n.i.e, Miscellaneous (software services, business services, financial services and communication services); b) Transfers: Official, Private; c) Income: investment income, compensation of employees.

Check Your Progress 1

1) What kind of Trade Data is compiled by DGCI&S?

……

……

[9] See under the section on Finance later in this Unit.

[10] News report in the Hindu dated February, 23, 2006 Chennai edition – Business page.

2) What is the existing base year for constructing unit value Index and quantum index by DGCI&S?

……

……

……

……

……

……

3) What are the reasons for divergence between two sources of trade data?

……

……

……

……

……

……

4) List the various transactions covered in the 'net invisibles'.

……

……

……

……

……

23.3 FINANCE

The finance sector consists of public finances, the central bank or the Reserve Bank of India, the scheduled banks, urban and rural cooperative banks and related institutions. The financial market consists of the stock exchanges dealing with scripts like shares, bonds and other debt instrument, the primary and secondary markets, the foreign exchange market, the treasury bills market and the securities market where financial institutions, mutual funds, foreign institutional investors, market intermediaries, the market regulator the Securities Exchange Board of India (SEBI), the banking sector and the RBI all play important roles. It also has the insurance (life and general) and pension funds as well as their respective regulators, i.e., the Insurance Regulatory and Development Authority (IRDA) and the Pension Funds Regulatory and Development Authority (PFRDA). It has the holding companies, which invest in various subsidiaries and controls its operations through its stocks[11]. There is also the unorganized sector made up of financial operators like individual money-lenders and pawn shops, insurance agents, unregistered stock brokers/

[11] One example of a Holding Company is Tata Sons, whose subsidiaries include TISCO, TCS, Tata Motors, Tata Housing, etc.

sub-brokers, etc. In the System of National Accounts, 2008 (SNA 2008)[12], the financial sector has been sub-divided into nine sub-sectors, depending on its type of business.

23.3.1 Public Finances

What would we like to know about public finances? We would like to know how they are managed. What are the sources of such finances and how and on what are they spent? Does the Government restrict its expenditure within its means or does it spend beyond the resources available to it? Does it, in the process, borrow heavily to finance its expenditure? **The Budget documents** of the Central and State Governments, the pre-Budget **Economic Survey** of the ministry of finance and the Reserve Bank of India **(RBI) Handbook of Statistics on the Indian Economy 2005** and the **Monthly Bulletin of the RBI** provide a variety of data on public finances. The National Accounts Statistics of the CSO, MoSPI also provides the output, GVA, gross savings, etc. of the financial sector and its various sub-sectors. **EPWRF website** will also help to access data in a time series format on public finances.The Economic Survey, for instance, gives an overall summary of the budgetary transactions of the Central and State governments and Union Territory Administrators. This includes the internal and extra-budgetary resources of the public sector undertakings for their plans. It indicates the total outlay, the current revenues, the gap between the two, the manner in which the gap is financed by net internal and external capital receipts and finally, the overall budgetary deficit. It gives the break-up of the outlay into developmental and non-developmental outlays and the components of the latter, the components of current revenues-tax revenue and non-tax revenue – and sub-components of these and the components of internal and external capital receipts. The **RBI Handbook 2013-14** presents the following time series data in respect of public finances.

a) **Central Government Finances**

i) Key deficit indicators[13] of the Central Government – gross fiscal deficit, gross primary deficit, net primary deficit, revenue deficit, primary revenue deficit, drawdown of cash balances, net RBI credit from 1975-76 onwards;

ii) Major components of Central Govt. Receipts – tax revenue (direct and indirect taxes and their components) non-tax revenue (one of its important components is interest receipts) and capital receipts;

[12] Discussed in Unit 20 in the discussion on National Accounts Statistics.

[13] **GFD** = total expenditure including loans (net of recovery) – revenue receipts (including external grants and non-debt capital receipts); **NFD** = GFD – net lending of the Central Govt; **GPD** = GFD – interest payment; **NPD** = net interest payments; **RD** = revenue receipts – revenue expenditure; **PRD =** RD – interest payment; **BD (Conventional Deficit)** = all expenditure (revenue and capital) – all receipts (revenue and capital); with the discontinuation of *ad hoc* treasury bills and 91-day treasury bills, the concept of conventional budget deficit has lost its relevance since 1/4/97. The figures shown against BD from 1997-98, therefore, represent draw down of cash balances from RBI; **Net RBI Credit to Govt.** = the sum of variations in the RBI's holdings of (i) Central Govt. dated securities, (ii) treasury bills, (iii) Rupee coins, and (iv) loans and advances from RBI to the Central Govt. since 1/4/97 adjusted for changes in the Central Govt.'s cash balances with the RBI in the case of the Centre.

iii) Major[14] heads of capital receipts of the Central Government – market borrowings, small savings, provident funds[15], special deposits, recoveries of loans, disinvestments receipts[16], external loans (net);

iv) Major heads of Central Govt. expenditure – revenue expenditure and its important components (defence, interest payments and subsidies), capital expenditure including loans and advances and defence expenditure; also the breakup of expenditure into developmental and non-developmental heads and the shares of economic services and social services in developmental expenditure;

v) Centre's gross fiscal deficit and its financing – GFD receipts, GFD expenditure, Gross Fiscal Deficit, financing of GFD through external financing and internal financing (market borrowings, other borrowings, draw down of cash balances);

vi) Gross capital formation from budgetary resources of the Central Government – fixed assets, work stores, increase in stocks of foodgrains & fertilisers, gross financial assistance for capital formation to State Governments, Non-Departmental Commercial Undertakings (NDCUs)[17] and others;

vii) Public sector Plan outlay, its sectoral profile and the manner in which it is financed – from sources like own resources, domestic market borrowings and net capital inflow from abroad;

viii) Financing of Public Sector Plan – balance from current revenue, contribution of public enterprises, borrowings (including long and medium term borrowings), small savings, deficit financing, net capital inflow from abroad, central assistance and others.

b) **Finances of the State Governments**

i) Key deficit indicators of the State Governments – fiscal deficit, gross primary deficit, revenue deficit, primary revenue deficit, overall deficit, net RBI credit to States (annual variation) from 1975-76 onwards;

ii) Pattern of receipts of the State Government – total Revenue receipts: tax receipts (like sales tax and State excise duties), the share of Central taxes like income tax and union excise duties; and non-tax receipts like interest receipts, grants from the Centre, total capital receipts;

iii) Pattern of major capital receipts of the State Government – loans from centre, recovery of loans and advances, market loans, State provident fund (net), special securities issued to NSSF;

iv) Expenditure pattern – revenue and capital expenditure; capital outlay, loans and advances by State Government, developmental expenditure and the shares of economic services and social services in it and non-developmental expenditure and the shares of interest payments, administrative services and pension and miscellaneous general services;

---

[14] Adjusted for changes in classification effected in 1974-75 and 1987-88.

[15] Only Govt. provident funds, as the Public Provident Fund (PPF) is part of small savings since 1998-99.

[16] These are not to be treated as budgetary receipts as these are to be credited to a separate fund.

[17] NDCU means the Central and State Public Sector Units

v) States' Gross Fiscal Deficit and its financing – loans from Central Government, market borrowings, special securities issued to NSSF and others,

vi) Outstanding liabilities of the State Governments.

c) **Combined Finances of Central and State Governments**

i) Combined deficit;

ii) Receipts and Disbursements;

iii) Direct and Indirect tax revenues;

iv) Developmental and non-developmental expenditure;

v) Market borrowings;

vi) Interest rates on dated securities (Range and weighted averages) of Central and State Government;

vii) Outstanding liabilities;

viii) Ownership of Central and State Govt Securities – 11 categories like RBI, scheduled commercial banks, cooperative banks, primary dealers, insurance companies, financial institutions, mutual funds, provident funds, corporates, foreign institutional investors and others;

d) **Transactions with the Rest of the World**

We have enumerated the kind of data available on Central and State Government finances. But these relate mainly to domestic finances and transactions. What about our transactions with the rest of the world? We have looked at one of these, namely, trade in merchandise and services in the section on trade. There are a number of other areas in which India interacts with the rest of the world. Foreign exchange flows into the country as a result of exports from India, external assistance/aid/loans/borrowings, returns from Indian investments abroad, remittance and deposits from Non-Resident Indians (NRI) and foreign investment (foreign direct investment – FDI – and portfolio investment) in India. Foreign exchange reserves are used up for purposes like financing imports, retiring foreign debts and investment abroad. What is the net result of these transactions on the foreign exchange reserves? What are the trends in these flows and their components? What is the size of the current account imbalance relative to GDP and its composition? If it is a deficit is it investment-driven? What is the size of foreign exchange reserves relative to macro-aggregates like GDP, the size of imports and the size of the short-term external debt[18]?

[18] One can measure the size of the forex reserves as equivalent to so many weeks or months of imports. Alternatively, one can relate it to financial stability. The Guidotti-Greenspan Rule sets a standard for this. According to this rule, reserves should equal one year's short-term debt. Yet another way of examining the size of the reserves is to look at it is terms of its opportunity cost. The reserves earn a zero real return measured in domestic terms. If on the other hand, these are invested either domestically in infrastructure or in a fully diversified long-term way in global markets, substantial incremental benefits would accrue to the domestic economy. [see **the Hindu, April 3, 2006, page 17 "Intriguing Pattern of Global Capital Flows" excepts from the L.K. Jha Memorial Lecture delivered recently (March 24, 2006) in Mumbai by Prof. Lawrence Summers.** President of Harvard University and former Secretary of the Treasury in the Clinton Administration, at the invitation of the RBI.] The full text of the lecture may also be published in the **RBI Monthly Bulletin**. The lecture discusses the implications of the rising forex reserves of the developing countries and the opportunities and challenges that these present.

The RBI Handbook 2013-14 gives time series data (which are often in US$ as well as in Indian Rupees) on a number of these parameters:

i) India's overall BoP showing current account and capital account and key components of these like trade balance, invisible[19], types of foreign investment, net external assistance, net commercial borrowing, rupee debt service and net NRI deposits. It also shows monetary movements in terms of increase/decrease in forex/reserves, IMF and SDR allocation,

ii) Exchange rate of Indian Rupee vis-à-vis US $, UK £, DM/€ and Japanese ¥ from 1975 onwards (calendar year – annual average, financial year – annual average and end year rates),

iii) Real Effective Exchange Rate (REER) and Nominal Effective Exchange Rate (NEER) of Indian Rupee based on 36 country bilateral weights with base 2004-05 from calendar year 2005 onwards,

iv) Indices of REER and NEER,

v) External assistance in the form of loans and grants from 1985-86 onwards (in US $ and Indian Rupees), authorization, utilization and debt service payments,

vi) NRI deposits outstanding (in US $ and Indian Rupees) from 1996 onwards in different types of accounts like NR (E)RA, FCNR(A), FCNR(B), NR(NR)RD, NRO, FC (B&O)D, FC(O)N and total. It has increased from US $17.446 billion in 1996 to US $103.844 in 2014,

vii) Foreign investment inflows (in US $ and Indian Rupees) namely, gross investments, repatriation/ disinvestment, FDI by India, net foreign direct investment, net portfolio investment from 2000-01 onwards,

viii) Foreign exchange reserves (in US $ and Indian Rupees) from 1956-57 onwards,

ix) India's external debt (in US $ and Indian Rupees) from 1996 onwards[20], which, inter alia, provides concessional debt as percentage of total debt, short term debt as % to total debt, debt stock-GDP ratio (%) and debt service ratio (%).

FDI has assumed importance in the context of the country's efforts to meet the increasing need for investment capital for its expanding economy. Data on FDI are important. These are available from August, 1991 onwards. FDI data collected by **RBI** and the **Department of Industrial Policy & Promotion (DIPP)** in the Ministry of Commerce & Industry before 2000 related only to equity capital. The coverage of the data has since then has been expanded so that these are in accordance with the best international practices. FDI data collected by RBI now cover (a) equity capital, (b) reinvested earnings (retained earnings of FDI companies); and (c) other capital (inter-corporate debt transaction between related entities). FDI data from 2000-01 onwards are, therefore, not comparable with data of earlier years. The Department of

[19] The component transactions that make up 'invisible' have been enumerated in the section on Trade – Services.

[20] The debts are classified into I. Multilateral, II. Bilateral, III. International Monetary Fund, IV. Trade credit, V. Commercial borrowing, VI. NRI & FC (B&O) deposits, VII. Rupee debt, VIII. Total long-term debt, IX. Short-term debt and X. Gross total debt. Most of the groups have sub-groups A. Government borrowing, and B. Non-Government borrowing, further disaggregated into concessional and non-concessional borrowing.

Industrial Policy and Promotion of the Ministry of Commerce and Industries publishes monthly data on FDI inflow in India[21]. It presents time series data on the break-up of FDI flows by

1) Cumulative FDI inflows (equity inflows + reinvested earnings + other capital) from 2000 – 2014,

2) Cumulative amount of FDI equity inflows (excluding amount remitted through RBI's + NRI schemes),

3) FDI inflows during the financial year till that month,

4) FDI equity inflows (month-wise) during that financial year and that calendar year,

5) Share of top investing countries FDI equity inflows,

6) Sectors attracting highest FDI equity inflows, showing 10 different areas, like service sector, construction development, telecommunication, computer software and hardware, drugs and pharmaceuticals, automobile industry, chemicals (other than fertilisers), power, metallurgical industries, hotel and tourism,

7) Statement of RBI's regional offices (with State covered) received FDI equity inflows,

8) Country-wise FDI equity inflows from April 2000 onwards (from 140 countries, FIIs, NRI),

9) Sector-wise FDI equity inflows from April 2000 onwards (showing 63 industrial sectors).

The RBI handbook 2013-14 and **the SEBI Handbook of Statistics of Indian Securities Market**[22] present time series on FII. The SEBI handbook, gives

i) Foreign Investment Inflows,

ii) Trends in FII to portfolio investment,

iii) Trends in resource mobilization by mutual funds,

iv) International Securities Market (market capitalization, no. of listed companies, value of shares traded, no. of trading days,

v) International Fixed Income Market (no. of bonds listed, value. of bonds listed, value of bonds traded),

vi) International Derivatives market (stock options, stock futures, stock index options, stock index futures).

FDI data can also be accessed in the **RBI website www.rbi.org.in**.

[21] You can visit the site http://dipp.gov.in/English/Publications/FDI_Statistics/FDI_Statistics.aspx#MainContent to get the monthly FDI inflow.

[22] http://www.sebi.gov.in/sebiweb/home/list/4/32/0/0/Handbook%20of%20Statistics

Check Your Progress 2

1) Name the documents that provide different kinds of data on public finances.

……

……

……

……

……

2) What is the difference between Gross Fiscal Deficit (GFD) and Net Fiscal Deficit (NFD)?

……

……

……

……

……

……

3) Identify the transactions other than trade in merchandise that India has with the rest of world.

……

……

……

……

……

……

23.3.2 Currency, Coinage, Money and Banking

Economic transactions need a medium of exchange. We have come a long way from the days of barter and come to the use of money and equivalent financial instruments as the medium of exchange. Banks function as important financial intermediaries not only in this process but also in matters of resource mobilization and the deployment of such resources. The central bank of the country (the RBI in India) regulates the functioning of banking system. In addition, it issues currency notes and takes steps to regulate the money supply in the economy in order to achieve the objectives of ensuring adequate credit to development activities and at the same time to maintain stability in prices. We should, therefore, be interested in data on money supply or the stock of money and its structure and the factors that bring about changes in these, the kind of aggregate that need monitoring, the transactions in the banking system in pursuance of the nation's development objectives, the flow of credit to different activities, indicators of the health and efficiency of banks which are the custodians of the savings of the public. We should also be interested in data on prices as price level affects the purchasing power of money and indices of

prices appropriate for the purpose/group in question – consumer prices for producers and different groups of consumers.

Most of these of data are compiled by the RBI on the basis of its records and those of NABARD and returns that it receives from banks and can be found in **RBI Bulletins and RBI handbook**. The wholesale Price Index (WPI) is compiled by the Economic Advisor's Officer in the Ministry of Industry, the Consumer Price Index for Industrial Workers (CPI – IW) and CPI for Agricultural Labour (CPI - AL) by the Labour Bureau, Shimla, Ministry of Labour and CPI for rural and urban (CPI – Rural, CPI-Urban and CPI Rural+Urban) by the CSO. All these are published by the agencies concerned and also presented in the RBI and CSO publications mentioned above. The **Consumer Price Index released on 12th of each month by the CSO** also provides retail prices of selected commodities/services, separately for the rural and urban areas in India. Two other reports of the Reserve Bank of India published every year – the **Report on Currency and Finance** and the **Report on Trends in Banking** provide a wealth of information of use to analyse. The RBI also maintains an online database in searchable format, where one can select the parameters for which one need to use the data. You can also right click on the selected page, copy the data necessary for your analysis and paste it on a blank excel workbook to carry out your own analysis[23]. The **EPWRF website** also provides data on Banking, Money and Finance.

The **RBI Handbook 2013-14** presents time series data on

i) Liabilities and assets of the RBI, 1980 onwards. The liabilities are deposits from Central Government, State Governments, Scheduled Commercial Banks, Scheduled State Co-operative Banks, Non-Scheduled Co-operative Banks, other Banks and others. The assets are notes and coins, balances held abroad, loans and advances to Central and State Govts. banks and other agencies, bills purchased and discounted, investments and other assets;

ii) Components of money stock, namely, Reserve Money (M_0) made up of currency in circulation, other deposits with RBI and bankers' deposits with RBI, currency with the public (= currency in circulation – cash with bank), Narrow Money (M_1) consisting of currency with the public, other deposits with RBI and demand deposits and Broad Money (M_3) comprising Narrow Money and time deposits;

iii) Sources of money stock consisting of net RBI credit to Central Government, net RBI credit to State Government, other banks' investments in Government securities, RBI credit to commercial sector, Other banks' credit to commercial sector, net foreign exchange assets of the RBI, net foreign exchange assets of other banks, Government's currency liabilities to public, net non-monetary liabilities of RBI, net non-monetary liabilities of other banks, RBI's gross claims on banks;

iv) Average monetary aggregates, like currency with the public, demand deposits, time deposits, other deposits with the RBI, reserve money, narrow money, broad money, net bank credit to Government, bank credit to governmental sector, net foreign exchange assets of the banking sector,

[23] See the link http://dbie.rbi.org.in/DBIE/dbie.rbi?site=home

Government's currency liabilities to the public, banking sector's net non-monetary liabilities;

v) Major monetary policy rates and reserve requirements: bank rate, LAF (REPO, reverse REPO, and MSF) rates, CRR and SLR;

vi) Monthly data series on the detailed composition of each of the components of the money stock (the "C" components) and the composition of individual sources of change in the money stock; also defining new monetary aggregates[24] NM_2 and NM_3, these being respectively equal to "M_1 + short-term time deposits" and "[NM_2+ long-term deposits + call/term funding from Financial Institutions (FIs)] = [domestic credit + Government's currency liability to the public + net forex assets of the banking sector – capital account – other items (net)];

vii) Monthly data series on Liquidity Aggregates[25], namely, L_1= NM_3 + Postal deposits[26]; L_2 = L_1 + liabilities[27] of FIs; L_3 = L_2 + public deposits with NBFCs[28] (L_3 is compiled on a quarterly basis);

viii) Monthly average price of gold and silver in domestic (Mumbai) and foreign markets;

ix) Selected Aggregates of Scheduled Commercial Banks (SCBs) like outstanding demand and time deposits, investment in Govt. and other securities, bank credit (food and non-food), cash in hand and balance with the RBI.

x) Deployment of non-food credit to priority sector and its sub-sectors (like agriculture, small scale industries), industry and its groups, whole sale trade and export credit; and short and long term direct and indirect institutional credit to agriculture and allied activities and to farmers by size of holdings;

xi) Consolidated balance sheets of SCBs; Gross and net Non-Performing Assets (NPAs) of SCBs by bank groups[29]; Distribution of SCBs and different sub-group of SCBs by Capital to Risk-weighted Assets Ratio (CRAR);[30] and

xii) Important banking indicators of Regional Rural Banks (RRBs), State Cooperative Banks, Primary Agricultural Coop. Societies (PACS), State Coop. Agricultural and Rural Development Banks and Primary Coop. (A&RD) Banks;

---

[24] See footnote 19.

[25] The methodology for compiling liquidity aggregates is available in the "New Monetary and Liquidity Aggregates" in the RBI Bulletin of November, 2000. The acronyms $NM_2$and NM_3 are used to distinguish the new monetary aggregates from the existing monetary aggregates.

[26] Post office SB deposits + PO time deposits + PO recurring deposits + other deposits + PO Cumulative Time Deposits.

[27] Term money borrowings + CDs + term deposits.

[28] Non-Banking Financial Companies. Estimates of public deposits are generated on the basis of a sample study of more than 1000 NBFCs with public deposits of Rs. 20 crores or more.

[29] 1) public sector banks, 2) old private sector banks, 3) new private sector banks, and 4) foreign banks in India.

[30] The detailed instructions issued by the RBI on CAR (Capital Adequacy Ratio) or CRAR can be seen at http://rbi.org.in/scripts/BS_ViewMasCirculardetails.aspx?Id=8133&Mode=0. SCBs had to comply with a minimum CRAR of 8 per cent up to the end of March, 1999 and 9 per cent from the end of March, 2000. The sub-group 1 in the footnote 25 is split into two sub-groups – SBI group and nationalized banks.

23.3.2 Financial Markets

What would we like to know about the financial market and its functioning? We should like to know about the ways in which financial resources can be accessed and at what cost. What are the prevailing interest rates payable for funds to meet short-term or long-term requirements? How do new ventures access the large amount of resources that are need for the new ventures? How do term lending institutions access funds required for their operations? What are the sources of funds?

The RBI, which regulates banking operations and the operations of the NBFCs and FIs and the Securities Exchange Board of India (SEBI), which regulates the capital market, and the Department of Company Affairs that administers the Companies Act, are the major sources of data on financial markets. The **RBI Handbook of Statistics of the India Economy – 2013-14**[31] and the **Handbook of statistics on the Indian Securities Market - 2014**[32] published by **SEBI** contain comprehensive data on the financial market. The two together provide annual time series data on several aspects of the financial market:

i) The structure of interest rates – call/notice money rates, commercial bank rate, lending rates of banks, prime lending rates (PLR) of term lending institutions like IDBI, dividend and yields of the units of UTI, annual gross redemption yields of Govt. Securities and average annual price and yield rate of Central Govt. securities (SGL transactions);

ii) Financial assistance sanctioned and disbursed and financing of project cost of companies by FIs, loans sanctioned by HDFC and NABARD, and refinancing operations of National Housing Bank (NHB);

iii) Aggregate deposits of NBFCs and Non-Banking Non-Financial Companies (NBNFCs)[33] ;

iv) Taxable and tax-free bonds issued by public sector undertakings – both public issue of bonds and privately placed bonds;

v) Resource mobilization in the Private Placement Market – financial and non-financial institutions in the public and private sector;

vi) Net resources mobilized by mutual funds (MFs) – MFs sponsored by banks, financial institutions (FIs), UTI and the private sector;

vii) New capital issues (number and amount mobilized) by non-govt. public Ltd. Companies;

---

[31] Also available on the RBI website http://www.rbi.org.in . In fact, the entire Handbook is available on the website. The web version can be downloaded from the website. For some of the financial market data longer time series are provided in the CD ROM and web versions. The CD ROM now incorporates intelligent search features that allows for searches across tables, enabling users to select one or more of any data series for any selected time period from any table in the handbook in a user-friendly manner. The data can be downloaded in the form of a user-defined spreadsheet table that can be read by most standard econometric software.

[32] Copies can be had from Research Deptt., SEBI, World Trade Centre I, 29th floor, Cuffe Parade, Mumbai-400 005. This is also available on the website of **SEBI – www.sebi.gov.in**and can be downloaded from the website.

[33] After the new regulatory framework for NBFCs came into force in 1998, the NBFCs and Residuary NBCs.

viii) Absorption of private capital issues – the no. of issuing companies, the number of shares and amount subscribed by promoters etc., and Govt., FLs etc., and the number of shares and amount subscribed by public, other than underwriters and other groups;

ix) Investments of LIC by sector and instrument and of UTI by instrument;

x) Assets and liabilities of institutions like IDBI, NABARD, EXIM Bank, NHB and SIDBI;

xi) Annual averages share price indices – BSE SENSEX (base 1978-79 =100), BSE National (base 1983-84 = 100) and RBI Index (base 1980-81 = 100) and Market Capitalisation[34];

xii) Market intermediaries like stock exchanges (cash and derivatives market), brokers, corporate brokers, sub-brokers, custodians, FIIs, depositories, merchant bankers, bankers to issues and underwriters registered with SEBI; and registered brokers by stock exchanges and by ownership categories – proprietary, partnership and corporate; (SEBI);

xiii) Long-term capital raised during 1957-90 (pre-reform period) through shares, debt and loans; (SEBI);

xiv) Annual and monthly data series on resources raised by the corporate sector through (i) equity issues and (ii) debt issues (public issues and private placement) and the share of private placement in total debt and total resource mobilization and the share of debt in total resource mobilization; (SEBI);

xv) Pattern of funding for non-govt, non-financial public limited companies – I. Internal sources [(i) reserves and surplus, and (ii) depreciation], and II. External sources [(paid up capital through new issues and premium), borrowings (debentures, from loans and from FIs) and trade dues and other current liabilities]; (SEBI);

xvi) Annual and monthly series on resources mobilized, instrument wise, from the primary market – number and amount mobilized by category of issue (public issue and rights issue); by type of issues [by listed companies and initial Public Offerings (IPO)]; by equities at par and equities at a premium; cumulative convertible preference shares (CCPS); bonds; and others; (SEBI);

xvii) Annual and monthly series of data on capital raised by (i) industrial (economic activity) classification (banks/FIs), industries like electronics, and engineering, entertainment, finance etc.; (ii) size of capital raised; (iii) sector (public and private); and (iv) region (north, east, south and west); (SEBI);

xviii) Annual and monthly data series on the number and quantum of Euro Issues; (SEBI);

[34] The compilation of the RBI Index was discontinued from 1999-00. The compilation of market capitalization – all India was discontinued by BSE since 1999-00. The SEBI Handbook provides data on market capitalization – all India from 1999-00 as given in the publication of the National Stock Exchange (NSE).

xix) Annual and monthly data series on transactions of MFs on the Stock Exchanges – gross purchases and sales and net purchase/sales in (a) equity, and (b) debt; (SEBI);

xx) Trends on trading on stock exchanges – the number of shares traded and the number and value of shares delivered; (SEBI);

xxi) Indicators of liquidity – Market capitalization – GDP ratio (BSE), market capitalization – GDP ratio (NSE), turnover ratio – BSE, traded value ratio – BSE, traded value ratio – NSE; (SEBI);

xxii) Trends in foreign investment flows – direct and portfolio investment;[35] (SEBI);

xxiii) Annual and monthly series on trends in FII investment – gross purchases and sales and net investment; (SEBI);

xxiv) Comparative evaluation of Indices through Price to Earnings Ratio and Price to Book Ratio (these are monthly averages of closing values) for BSE SENSEX, BSE 100 Index, S&P CNX NIFTY, CNK NIFTY Junior; (SEBI); and

xxv) Survey of investor households – a joint effort of SEBI and the National Council of Applied Economic Research – result giving an estimate of investor and non-investor households by type of investment, household by type of instruments invested in and so on.

Check Your Progress 3

1) Which document does contain the methodology for compiling liquidity aggregates?

……

……

……

……

……

2) List the kind of time series data available in RBI Handbook 2013-14.

……

……

……

……

……

……

[35] Figures from 1995-96 include acquisition of shares of Indian companies by non-residents under S/6 of FEMA, 1999. Those from 2000-01 have been revised with expanded coverage to approach international best practices and are, therefore, not comparable with earlier data.

3) Identify 3 different sub-sectors of financial market and name a few major sources of data for each of these sub-sectors of financial markets.

……

……

……

……

……

……

……

23.4 LET US SUM UP

We have, in this Unit, surveyed the data available in the area of trade and finance. We noted that data on the volume of merchandise trade by commodities and countries, direction of trade, trade balance, quantity and unit value index numbers of imports and exports and indices depicting different measures of the terms of trade are available from the DGCI&S and are also presented as a time series in the **RBI Handbook 2013-14**, besides the **Handbook of Statistics of the SEBI**. Data on trade in services and on merchandise are available as part of BoP data of the RBI. There is a divergence between merchandise trade deficit/surplus data provided by DGCI&S and that shown by BoP data of RBI and the reasons for such a discrepancy are related to the manner in which the two sources collect the basic data.

Data are available in the RBI Handbook (long time series) on public finances such as receipts and expenditure of the Central and the State Governments, the manner in which these Governments mobilize resources through taxes and non-tax revenue borrowings, the patterns of their expenditure – developmental outlays, debt servicing, economic and social services and so on – and the various types of (fiscal) deficits that they run and the manner in which these are financed. Data on balance of payments, inflow of foreign investment, foreign aid/borrowings, assistance and the size of foreign debt are also available in the RBI Handbook. So are the data on the money stock, the different monetary aggregates like M_0, M_1 and M_3, and the new monetary aggregates NM_2 and NM_3 and "C" components of the money stock, the sources of change in the money stock and their "S" components and liquidity aggregates L_1, L_2 and L_3. The role of the banks and financial institutions in the mobilization of savings and the deployment of credit to different economic activities, their functional and operational efficiency in terms of indicators like NPA and CRAR can be examined well with the time series data available in the RBI Handbook and such an effort can be supplemented with information on profitability provided by the RBI's Report on Trends in Banking. Similarly, the functioning of the financial market in terms of the structure of interest rates in different markets, resources mobilized through different modes like capital issues, private placement, equities with and without premium, preference shares and **CCPS**, total resources raised from the primary market, trends in trading in stock exchanges, share price indices and market capitalization, indicators of liquidity like traded value ratio and market capitalization to GDP ratio and measures of comparative evaluation of indices (BSE SENSEX, BSE 100 Index, etc.,) like Price to Earnings Ratio and Price to Book Ratio, can be analysed adequately

with the data available on these aspects in the RBI **Handbook and the Handbook of Statistics**, published by SEBI. The RBI Handbook, especially with the special features incorporated in its website version, stand out as an invaluable source of data on finance and also trade.

23.5 EXERCISES

1) How can you analyse trend in foreign trade? Discuss the role of unit value index and quantum index in this regard.

2) Discuss the kind of time series data on money and banking compiled by RBI. How this data can be useful for research?

3) List the types of data on financial market compiled by SEBI. To what extent is it adequate to analyse the financial markets?

23.6 SOME USEFUL BOOKS

Ministry of Finance, Govt. of India : Economic Survey – 2005-06 and earlier years, Ministry of Finance, Govt. of India, New Delhi.

Budget Documents, Ministry of Finance, Govt. of India, New Delhi.

Reserve Bank of India : Report on Currency and Finance – 2005, RBI, Mumbai.

Report on Trends in Banking – 2005, RBI, Mumbai.

World Trade Organization : International Trade Statistics – 2005, WTO, Geneva.

Also accessible at http://www.wto.org/english/res_e/statis_e.htm

23.7 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Sub-section 23.2.1

2) 1998-99.

3) See Sub-section 23.2.1

4) See Sub-section 23.2.2

Check Your Progress 2

1) The budget documents of the central and the state governments, the pre-budget economic survey, RBI Indian Economy 2005, Monthly Bulletin of the RBI.

2) The excess of total expenditure including loans over revenue receipts – net lending of the central government.

3) See Sub-section 23.3.1 (sub head and ‘d’)

Check Your Progress 3

1) RBI Bulletin, Nov. 2000 under the title 'New Monetary and Liquidity aggregates'.

2) See Sub-section 23.3.2

3) The RBI Handbook of Statistics of Indian Economy 2005 and the Handbook of Statistics on Indian Securities Market 2005 provide comprehensive data on the financial market.

UNIT 24 SOCIAL SECTOR

Structure

24.0 Objectives

24.1 Introduction

24.2 Employment, Unemployment and Labour Force

- 24.2.1 Magnitude of Employment and Unemployment
- 24.2.2 Quality/Adequacy of Employment
- 24.2.3 Labour Welfare

24.3 Education

- 24.3.1 Educational Infrastructure
- 24.3.2 Infrastructure Utilization and Access to Educational Opportunities

24.4 Health

- 24.4.1 Health Infrastructure
- 24.4.2 Public Health, Morbidity and Mortality Indicators
- 24.4.3 National Family Health Survey (NFHS)

24.5 Shelter and Amenities

24.6 Social Consequences of Development

24.7 Environment

24.8 Quality of Life

24.9 Let Us Sum Up

24.10 Some Useful Books

24.11 Answers or Hints to Check Your Progress Exercises

24.0 OBJECTIVES

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

- know the sources of data on various aspects of employment and unemployment, labour welfare, education, health, shelter, safe drinking water, etc., which determine the quality of life of the people;
- Identify the various organization/agencies involved in the compilation of data on different aspects of social sector;
- describe the different concepts used in collection of data on social sector; and
- explain the kind of data/information available in the various publications containing data on social sector.

24.1 INTRODUCTION

Social Sector consists of education, health, employment, shelter, sanitation, other housing amenities, environment, and adverse consequences of development and levels of living or quality of life in general. Investments in

this sector pay rich dividends in terms of rising productivity, distributed growth, reduction in social and economic inequalities and levels of poverty although after a relatively longer time span than in the case of investment in physical sectors. Let us look at the kind of data available in this sector.

24.2 EMPLOYMENT, UNEMPLOYMENT AND LABOUR FORCE

Employment is the means to participate in the development process. Creation of employment opportunities is an important instrument for tackling poverty and to empower people, especially women. Regarding data on employment, firstly, we should know how many are employed and how many are ready to work but are unable to gain access to employment opportunities. How do women fare in these matters? Or, for that matter, what is the experience of men and women belonging to different social/religious/disadvantaged groups? Are children employed in any economic activity that is not only hazardous to their health but which also adversely impacts on our dream of a golden future for them through efforts to ensure their mental and physical well-being? What is the quality of employment opportunities available to the work force? What are the conditions in which people work? Let us first look at data on the magnitude of employment and unemployment.

24.2.1 Magnitude of Employment and Unemployment

Data on employment in selected sectors are available from several sources. Data on Employment in the organized sector of the economy is available from the **Employment Market Information (EMI) programme of the Directorate General of Employment & Training (DGE&T), in the Ministry of Labour**. This is based on statutory returns submitted[1] to the employment exchange of the area every quarter by non-agricultural establishments in the private sector employing at least 25 persons and all public sector establishments (excluding defense forces and Indian missions posted abroad) irrespective of their size. Smaller establishments, that is, those employing 10 to 24 persons submit such returns on a voluntary basis. Data on employment levels in the organized sector of the economy (as defined above) is thus available every quarter at district levels through **Quarterly Employment Reviews** released by the DGE&T and the Directorates of Employment (DEs) in the State Governments and Union Territory Administrations. The **Annual Employment Reviews** released by the DGE&T and the DEs provide more detailed data on employment at the three levels of NIC code. The returns also provide data on the number of vacancies occurring in the establishment during the quarter and the number filled. This is the only source of employment data on the organized sector of the economy that is available at quarterly intervals. This is, however, subject to several limitations. First of all, it excludes the entire agricultural sector, self employed persons and part-time workers. In addition, adjustments are to be made to take care of non-response in the submission of the returns. Incompleteness of the employers' register is also an impediment. Establishments in the organized sector of the economy also submit statutory biennial returns providing the occupational distribution of their employees and the educational profile of those employed in selected occupations. These returns provide the basis for the preparation of biennial reports on the occupational pattern of employment and

[1] These statutory returns are submitted under the Employment Exchanges (Compulsory Notification of Vacancies) Act, 1959 and the Rules of 1960 made there under.

the educational profile of selected occupations in the public and private sectors. However, non-response in the submission of these returns has over the years deteriorated so much that the data on the occupational/education pattern has now lost its utility.

The **DGE&T and the DEs** in the State Governments and Union Territory administrations also provide data on the number of job seekers by age, sex and education qualifications, and the socially and physically challenged and their distribution by sex and education on the live register of employment exchanges, the number of vacancies notified to the employment exchanges under the Act referred to in the preceding paragraph and the number of job seekers placed in jobs by the employment exchange. It would not at all be fair to measure the efficiency of the employment exchanges by comparing the number of placements with the number of vacancies notified to the exchanges or with the number on the live register for several reasons. First, vacancies being filled through public service commissions, other commissions, recruitment boards and the like and those filled through competitive examinations need not be notified to employment exchanges. Second, public sector undertakings need to fill from the Employment Exchanges only vacancies carrying salaries below a certain level specified by Government. Finally, establishments in the private sector need to only notify their vacancies to the employment exchanges. They are not obliged to confine their choice of candidates for recruitment to the nominees of the exchanges and are free to fill their vacancies through the open market. Sample surveys of those on the live register at any point of time have shown that a sizeable proportion of the job seekers are already employed, some are unemployed and some have registered at more than one exchange. At the same time, surveys of employment and unemployment like the NSSO quinquennial surveys (see below) have shown that not all the unemployed are registered in the employment exchanges. Thus, the size of the live register can overestimate unemployment and, from another angle, can also underestimate unemployment. These factors may not also cancel each other and it is, therefore, difficult to consider the size of the live register as a reasonably accurate estimate of the level of unemployment. Nevertheless, it does represent the extent of pressure in the job market, especially for government and public sector jobs.

The second source is the **Economic Census**. This has been conducted in 1977, 1980, 1990, 1998, 2003 and 2013 covering all economic enterprises in the country except those engaged in crop production and plantation and provides data on employment in these enterprises[2]. These thus include also the unorganized sector outside crop production, animal husbandry and plantations. The third source is the **Annual Survey of Industries (ASI)**, which gives us estimates of factory sector employment, i.e., all factories registered under the Factories Act, 1948 and the Bidi and Cigar Workers Condition of Employment Act, 1966[3]. The ASI presents estimates of levels of employment by industries at the three-digit level of NIC 2008 codes along with industry level aggregates like investment, output, inputs, value added and so on[4] so that one can derive useful technical ratios to facilitate analysis of the role of the factors of production in

[2] **The first report on the Economic Census 2013** is available on the website of the Ministry of Statistics and Programme Implementation (MoSPI) www.mospi.gov.in .

[3] See the Sub-section 21.3.3 titled "Factory Sector – Annual Survey of Industries" in Unit 21 on Agricultural and Industrial Data.

[4] See the list of principal characteristics given in footnote 22 in the section referred to in the preceding footnote. Sub-section 21.3.3 referred to therein also gives an idea of the kind of data on employment available from **ASI**.

industry. The quality of ASI data is thus tied to the completeness of the frame of factories, which, in turn depends on the quality of the enforcement of the Factories Act, 1948 and the *Bidi* and Cigar Workers (Conditions of Employment) Act, 1966. Data on employment in Railways is available from the **Railway Board, Ministry of Railways**, those on employment in the Banking Sector from the **Reserve Bank of India (RBI)**, those on employment in posts and telecommunications respectively by the **Department of Posts** and the **Ministry of Telecommunications** and those on employment in insurance by the **Ministry of Finance**. The reports on the **Census of Central Government Employees** conducted by the DGE&T every year and the **Census of Central Government Employees** conducted by the respective State Governments provide a time series data of employment in Government sector.

The publications[5] of the Labour Bureau (Ministry of Labour), Shimla present estimates of average daily number of workers in Government and local fund factories and other factories – male and female adults (those who have completed 18 years of age), male and female adolescents (those who are aged between 15 to 18) and boys and girls (those who are not yet 15 but not below 14) – and employment in factories by industry groups. These also provide estimates of average daily employment in mines and quarries for different minerals. The first is based on the statutory returns received from factories through the State Governments under the Factories Act, 1948 and the second on similar returns received by the Director General of Mines Safety (DGMS) under the Mines Act, 1952 as amended in 1983. The third set of data that they provide is employment in shops, commercial establishments and restaurants, theatres, etc., based on the returns submitted to the State Governments under the Shops & Establishments Act and Weekly Holidays Act, 1942. All suffer from inadequacy of response in the matter of submitting statutory returns. Further, the last one suffers from an added limitation arising from the fact that these Acts are in force only in certain urban areas, towns and cantonments. Of late, similar returns are being collected by some of the Gram Panchayats, who work under the Department of Rural Development of the respective State Governments. Lastly, the publications give data on employment in tea, coffee and rubber plantations on the basis of data received from the Tea, Coffee and Rubber Boards and the Ministry of Agriculture. It must be mentioned here that employment in these specific sectors will also be covered by the **EMI programme** to the extent establishments in these sectors are covered by the Act governing the EMI programme and the extent to which these establishments submit the relevant statutory returns.

Comprehensive data on employment and unemployment covering the entire country at regular intervals are available from two sources, namely, the **Population Census** conducted by the **Office of the Registrar General and Census Commissioner, India** every ten years and the **quinquennial sample surveys on employment and unemployment** conducted by the **National Sample Survey Office (NSSO)**. The former is based on a complete enumeration of the population. Workers in the Census are enumerated as main workers and marginal workers. The Census 2011 presents data, at the village/urban block level (for slums in all statutory towns also, which have been treated as separate enumeration blocks in **Census 2011**), on the number employed (male and female) or the size of the male and female workforce

[5] Indian Labour Yearbooks (the latest is for 2009 and 2010) and Annual Employment Unemployment Survey (the 4th report of this pertaining to 2012-13 is the latest) Labour Bureau (Ministry of Labour and Employment), Shimla.

(main workers and marginal workers) and its distribution by four economic activity categories – (i) cultivators, (ii) agricultural labourers, (iii) household industries and (iv) other activities. The rest of the population, namely, those who are neither main workers nor marginal workers, are categorized as non-workers. All other data on employment and unemployment are available at the district level upwards for rural and urban areas and the urban data by urban agglomerations and size classes of cities and towns. In Census 2011, till December 2014, the following data has been published:

a) The Primary Census Abstract (PCA), published for each village and urban ward, provide, among others, the number of female and male main workers, marginal workers and non-workers for overall, SC and ST for each of the four-fold classification of economic activity, i.e., cultivators, agricultural labourers, household industry and other workers;

b) The number of female and male main workers, marginal workers and non-workers in each single year of age, classified further by four-fold classification of economic activity and social group disaggregated at India, State and district level.

The kinds of data on employment and unemployment available from the Census 2001, and to be published in Census 2011 for rural (R) and urban (U) and (R+U) is shown below:[6]

a) The distributions of the male and female main workers by economic activity (the first digit level of the NIC code) – ten categories of economic activities[7] and classified further by age groups and educational level (for the all India table) and by educational level[8] for the table on India, States and Union Territories;

b) Similar distributions for male and female marginal workers;

c) Main workers, marginal workers, non workers, those marginal workers that are available for or are seeking work (unemployed) and those non-workers that are available for or are seeking work (unemployed), by sex, social group and educational level;

---

[6] **Census** tables released so far are available on **CDs** for users on payment of the prices indicated in the **Census website http://www.census.nic.in**. Most of the India and State/UT level tables can be viewed in / downloaded from the census website. Highlights of and introductions to tables, concepts and definitions and an advance release calendar are also in the website. Tables are also available at State/Union Territory/District City level on **CDs**.

[7] In Census 2001, the NIC 1998 was used for dissemination of the results. In NIC '98, the first level had the following categories: a) Agriculture, Hunting & Forestry, b) Fishing, c) Mining & Quarrying, d) Manufacturing, e) Electricity, Gas and Water Supply, f) Construction, g) Wholesale and Retail Trade, h) Hotels and Restaurants, i) Transport, Storage and Communications, j) Financial Intermediation, k) Real Estates and Business Activities, l) Public Administration and Defence, and Compulsory Social Security, m) Education, n) Health & Social Work, o) Other Community, Social and Personal Service Activities, p) Private Households with Employed Persons, q) Extra-Territorial Organizations and Bodies. In Census 2001 tables, Categories A& B were grouped together, J & K were grouped together and L to Q were grouped together.

[8] 1)Illiterate, 2) literate but below metric, 3) metric/secondary but below graduate, 4) technical diploma or certificate not equal to degree, 5) graduate and above other than technical degree, and 6) technical degree or diploma equal to degree or post graduate degree and a further sub-classification of item 6 into engineering and technology, medicine, agriculture and others.

d) Similar distributions for unemployed SC population, ST populations and further classification of the distribution for unemployed general population by religious communities;

e) Distributions of (i) main workers, (ii) marginal workers, and (iii) non-workers by main activity, educational levels, age and sex; and

f) Similar distributions for SC population, ST population, further classification of the distribution in item "e" above for the general population by religious communities;

g) The number of main workers, marginal workers and non-workers among the disabled by type of disability, age and sex; and

h) The male and female workforce by detailed economic activities at the three or four digit levels, distribution by industrial categories (according to **National Industrial Classification – 1998 (NIC 1998**) and a separate table by occupations (as per the **National Classification of Occupations – 1986 (NCO 1968)** and educational profile and;

i) The number of marginal and non-workers who are seeking/available for work;

j) The major non-economic activity of the non-workers, namely, student, pensioner, household duties, etc. by age and sex. This category also provide information on vulnerable sections of the society like beggars, etc.

The **NSSO quinquennial surveys on employment and unemployment** constitute the other major source of data on employment and unemployment. The last survey for which results have been published is the **68th Round survey** conducted during July, 2011 and June, 2012. The earlier surveys related to October, 1972 to September, 1973, July, 1977 to June, 1978, January – December, 1983, July 1987 to June, 1988 July, 1993 to June, 1994, July 1999 – June 2000 and July 2004- June 2005. These measure employment/unemployment/labour force status on the basis of the Usual Status and Current Status and arrived at four measures of employment/unemployment/labour force, namely, the Usual Principal Status (UPS), the Usual Principal and Subsidiary Status (UPSS), the Current Weekly Status (CWS) and the Current Daily Status (CDS) employment/ unemployment/labour force status. These concepts and the measures have been discussed in the **Unit on "Employment and Unemployment: Policy Implications" in Block 1 of MEC – 005** on Indian Economic Policy in your first year course. These surveys provide the per thousand distribution of variables by characteristics along with estimated number of persons and/or the sample number of persons for each column and row characteristic. These can be used to estimate new proportions – like for example the worker participation rate or the unemployment rate among the poor (those below a certain income class) and the non-poor – and also estimates of the magnitude of employment or unemployment using the estimated sample proportions and the relevant population projection for the midpoint of the survey period. The surveys[9]

[9] Some of the reports relating to employment and unemployment on the 68th round are **Report No. 554 on "Employment and Unemployment Situation in India" and Report No. 557 on "Informal Sector and Conditions of Employment in India".** All these can be downloaded from the **MoSPI website: www.mospi.gov.in by becoming a registered user, which is free of charge.**

provide the following type of data[10] for Rural (R), Urban (U) and (R+U) at national and State levels[11]:

a) Distribution of male and female UPS/UPSS/CWS/CDS workforce by employment status – self-employed, regular wage/salaried employment and casual labour – and by age group, educational level and primary, secondary and tertiary economic activity[12];

b) Distribution of UPS/UPSS male and female workforce by economic activity up to the three-digit code of NIC; and similar distribution by occupation up to the three-digit code of NCO 1968;

c) Distribution of male and female UPS/UPSS/CWS/CDS workforce by employment status and by primary, secondary and tertiary economic activity and by MPCE classes;

d) Distribution of UPS workers in each employment status by usual subsidiary economic activities;

e) Distribution of UPSS employed by place of work (same village/town, another village/town) and by distance of the place of work from the place of residence;

f) Proportion of persons who are more or less regularly employed and the distribution of persons who are not regularly employed by duration for which they are available for work; (*underemployment*)

g) Distribution of UPS employed by "CWS employed", "CWS unemployed", "out of the labour force by the CWS criterion", showing *underemployment* among the UPS employed;

h) Similarly distribution of the CWS "employed", "unemployed," and "out of the labour force" on each half-day of the reference week giving the *underemployment* among the CWS employed;

i) Distribution person-day of employment by (i) principal [13] economic activity of the household, (ii) household type[14], and (iii) household land cultivated class, [distribution at (iii) only for rural areas];

j) Distribution of households and female-headed households by the number of male and female adults (aged 15+) employed under the UPS criterion – households with no adult UPS employed, number with one male adult UPS employed, number with one female adult employed, number with one male adult and one female adult employed and so on;

---

[10] The **unit record data,** that is the basic data collected in the surveyed households, are available in **CDs** from the Computer Centre of the MoSPI on signing an undertaking and making a payment. The details can be seen at http://mospi.nic.in/Mospi_New/upload/nsso/ratelist_UnitData.pdf.

[11] Technically, data should be available by 70 and odd NSS regions being made up of one or more contiguous districts. As the **unit record data are available on floppies from NSSO on payment,** one can make any kind of tabulation for any region or regions. However, one should be checking the sample size while preparing such estimates, as that affects the reliability of an estimate.

[12] In some cases, only by agricultural sector and the non-agricultural sector.

[13] The economic activity that accounted for the maximum of the income of the household in the 365 days preceding the date of the survey.

[14] The nature and type of work from which a household derives its major income is an important indicator of the activity pattern of its members. The classification of households by household type is based on the major economic activity of the household during the 365 days preceding the date of the survey.

k) UPS/UPSS/CWS/CDS unemployment rates (can be easily estimated from the tables on distribution of 1000 persons by various characteristics, as indicated in the main paragraph) for all groups together and for different groups like gender, age group, educational level, MPCE classes, household type, household land cultivated class and principal household economic activity, etc.;

l) Average daily salary/wage earnings of regular wage/salaried employees aged 15-59 by sex, sector of work and education; a similar distribution by occupational groups and education; and

m) Average daily wage/earnings per day received by casual labour by sex, age group, type of operation, and sub-round[15].

Data on employment are thus available from several sources. Estimates of employment derived from different sources for the same sector or sub-sector will differ. While this may be frustrating to a lay person, it opens up an exciting and challenging opportunity to an analyst to unravel the factors responsible for such variations and to find ways and means of tackling such factors. Such factors could be differences in the concepts, definitions and mode of collection of data (sample survey, census or administrative/statutory reporting systems) used by the different sources. There may be other factors causing the divergence in estimates. Some steps could be taken to remove at least a few of these factors. Data collecting agencies should adopt the same concepts and definitions for enquiries on the same or similar subjects, as far as possible. The use of trained personnel in data collection does have a favourable impact on the quality of data and should receive consideration. Estimates based on sample surveys should accompanied by estimates of the standard error to which the estimates from the survey are subject[16].

24.2.2 Quality/Adequacy of Employment

NSSO survey data (items above) can be used to analyse the quality and adequacy of employment. Besides underemployment among the employed, other aspects of quality can also be looked at. One is the proportion of the employed by employment status, especially the proportion of the work force in 'casual employment'. The second is a comparison of the average daily earnings of male and female, regular and casual workers in different sectors and operations with the prevailing statutory minimum wage and the poverty line. The third is a look at the pattern of employment status of the workers belonging to poor households. These will throw light on the quality of employment enjoyed by the workforce, quality in terms of the intermittent nature of work, tenure of employment, employment security and low wages and above all, the prevailing gender differences in these aspects of quality. Gender differences in access to employment in different occupations and sectors of economic activity would be clear from a perusal of the tables listed in the preceding section. Even aspects of the question of the invisibility of the female worker, leading to underestimation of the female workforce can be examined[17].

---

[15] One year long round of the Survey has four sub-rounds to take note of the seasonal effect on the variables under investigation.

[16] Reports on NSSO surveys do provide information on the concepts, definitions, coverage, sampling design and estimation procedures adopted and the standard errors of estimates presented in the reports.

[17] **NSSO Report No. 554** referred to in an earlier footnote. See also **"Gender and Employment in India", T.S. Papola & Alakh N. Sharma (Ed), published by Indian Society of Labour Economics and Institute of Economics Growth in association with Vikas Publishing House Ltd, New Delhi. (1999).**

The Labour Bureau, Shimla under the Ministry of Labour also publishes[18] data on wage levels in the organized sector and on the welfare of labour. Data on total earnings of factory workers are collected through the statutory returns under the Payment of Wages Act, 1936 from establishments defined as factories under the Factories Act, 1948. The Act was applicable to employees with earnings[19] up to Rs. 200. This limit has been increased from time to time to 1600 in 1982. These returns are, however, not received from all factories. The Bureau also conducts Occupational Wage Surveys in selected industries at regular intervals (called first round, second round and so on) and publishes reports[20] on these surveys) that provide the distribution of workers by levels of earnings and by occupations for each industry. These reports facilitate analysis of variation in levels of earnings within occupations, across gender within occupations, across regions and industries and to judge the adequacy of such employment opportunities in the light of statutory minimum wages and the poverty line. We have seen in the Unit on Agricultural and Industrial Data that the ASI gives information on the total wages, total emoluments and employee compensation along with mandays of workers and mandays of employees in different industries in the factory sector. These thus help in deriving the average wage per worker-manday and the average emoluments per employee-manday and the average employee compensation per employee-manday.

As for the unregistered sector and the unorganized part of other sectors, the **DE, NDE and OAE Surveys[21] and other Establishment Surveys of the CSO** in different sectors (till the early 1990s) and the **unorganized sector surveys of the NSSO[22]** provide data on average annual earnings for men, women and children in such establishments. **Regular reports like "Wage Rates in Rural India"** (the latest published relates to 2004-05) and the **Report on the Working of the Minimum Wages Act, 1948** (the latest is for 2003), of **Labour Bureau** enable an analysis of rural/unorganized sector wage levels *vis-à-vis* statutory minimum wages and poverty line.

The prevalence of child labour, which depicts an exploitative dimension of the economic system, can be seen from the Census and NSSO tables on the distribution of the work force by age groups and activity, the former down to the district level and the latter up to the State level (technically down to NSS regions due to availability of the unit record data on floppies/CDs). Such data may not fully reflect the ground level realities because surveys may not be able to extract information on employment of children, for various reasons. Further, such data are also dated. These can help only in drawing attention to the areas where child labour is prevalent so that the authorities concerned can initiate further action. It is usually the Non-Government Organisations (NGOs) active in the field of the rights of children that succeed in locating establishments employing children and press the Government to take further action.

---

[18] **Yearbooks of Labour Statistics, (various years), Ministry of Labour, Shimla.**

[19] Earnings include basic wages, dearness allowance, money value of concessions, annual or prepaid bonus, and arrears.

[20] **Reports on Occupational Wage Surveys** (Successive Rounds – the latest publications relate to the sixth round), Ministry of Labour, Labour Bureau, Shimla.

[21] Directory Establishments, Non Directory Establishments and Own Account Enterprises. The first two employ at least one hired worker and a total of 6 or more persons and 1 to 5 persons respectively. In Own Account Enterprises, no hired worker works on a regular basis. These surveys are now being conducted by the NSSO. The of **NSS Report 549, for the year 2010-11 is the latest report on unorganized non agricultural enterprises (excluding construction), which provide data on DE, NDE and OAE.**

[22] Besides the reports listed in Sub-section 21.3.4 of unit 21.

24.2.3 Labour Welfare

The Labour Bureau publishes[23] data on several aspects of labour welfare – data on industrial injuries, injuries in mines, compensation to workers for injuries and death, industrial disputes, health insurance, provident fund and trade unions of employers and workers. Statistics on industrial injuries are collected through statutory returns under the Factories Act, 1948 that provides that industrial accidents due to which the affected persons are prevented from attending to work for at least 48 hours should be reported to the inspector of factories. These depend on returns, and the sizeable non-response in the submission of these mars the quality of this source of data on incidence of fatal and non-fatal injuries in factories in mines under the Mines Act, 1952, as amended in 1983. Initially, data on the number of serious accidents and the accident rate were classified as 'fatal' and 'serious'. Subsequently, from 1984, accidents and the accident rate (accidents per 100 employees) were classified as 'fatal' 'spot serious'[24] and 'reportable serious'[25]. Complete statistics are not available for the same reasons mentioned earlier. Statistics on 'compensated injuries' and the amount of compensation paid, both classified as (resulting in) 'death' and (leading to) 'disability' – 'permanent' and 'temporary', are collected under the Workmen's Compensation (WC) Act, 1923 on the basis of annual returns received from the State Governments, Posts & Telegraph Departments and the Railway Board for different Zonal Railways under the Act. Compensation is payable to workers employed in 'scheduled employments'[26] for injuries due to accidents resulting in death or disablement for more than three days, provided that it is not caused through the fault of the worker himself. The number of injuries reported in the returns does not reflect the total number of injuries that occur, as all the injuries are not compensated for. Further, many of the establishments covered by the Act fail to submit returns and, therefore, the information received by the State Governments about the compensated injuries and the amount of compensation paid is incomplete. Compensation for injuries in establishments covered by the Employees State Insurance (ESI) Act, 1948 are paid under the ESI Act and not under WC Act. These limitations affect the trends and comparability over time of compensated accidents.

Statistics on the number of (a) industrial disputes[27], (b) workers involved in the disputes, (c) mandays lost and (d) causes of disputes – (i) wages and allowances, (ii) bonus, (iii) personnel matters, (iv) retrenchment, (v) leave and hours of work/shift working, (vi) indiscipline and violence, (vii) others, and (viii) cause not known. Completeness of these data would depend on the extent to which outside authorities are involved in the resolution of the disputes. How are the workers and the employers, the two pillars of industrial progress, organized to fight for their rights? The Trade Unions Act, 1926 regulates the formation and functioning of such organizations. The annual returns under the

[23] Yearbooks of Labour Statistics referred to in an earlier footnote.

[24] This is one where one or more persons have received serious bodily injury. This means any injury which involves, or in all probability will involve, (a) the permanent loss of any part or section of the body or use of any part or section of the body or (b) the permanent loss of, or injury to, sight or hearing or (c) any permanent incapacity or (d) the fracture of any bone or one or more joints or bones of any phalanges of hand or foot.

[25] This is one where one or more persons have received "reportable injury". This means any injury other than a serious bodily injury, which involves, or in all probability will involve, the enforced absence of the injured person from work for a period of 72 hours or more.

[26] Employments specified in the Schedule appended to the Act under reference.

[27] Disputes resulting in work-stoppages involving 10 or more workers.

Act received from the State Governments relate only to those organizations that have been registered with the State Governments under the Act. And it is not obligatory on these organizations to secure registration under the Act. The response rate even among the registered ones is less than 50 per cent. The data provided in the **Yearbooks of Labour Statistics** relates to the number of (i) workers' organizations on the register (ii) those submitting returns, (iii) percentage response, (iv) membership at the end of the year, (v) income including balance carried over from the previous year, (vi) expenditure, and (vii) balance of funds at the close of the year and similar data for employers' organizations. Information on the membership of the employers' and workers' organizations by NIC 1987 is available in the Yearbooks. Statistics relating to the working of various welfare funds like the Coal, Mica etc., Mines Welfare Funds and the ESI Corporation and the Employees Provident Fund Organisation (EPFO) – number of beneficiaries/members, etc., - are also provided by Yearbooks and the **annual reports of the ESIC and EPFO**[28].

As for the unorganized sector, the **Labour Bureau's** reports on their ongoing programme of surveys on (i) the working and living conditions of Scheduled Caste/Tribe workers, (ii) living conditions of unorganized workers, and (iii) contract labour provide information on the qualitative aspects of employment in terms of variables like wage levels, working and living conditions, work place safety and amenities at the work place of these classes of unorganized workers in rural and urban areas.

Check Your Progress 1

1) Name the sources that provide data on employment.

……

……

……

……

……

2) Explain the kind of data on employment and unemployment available from the population census 2001.

……

……

……

……

3) State the different measures of employment/unemployment/labour force used by NSSO in different quinquennial surveys.

……

……

……

……

[28] See also the Section on Health in this Unit.

4) Identify the different aspects of labour welfare on which data are compiled by the Labour Bureau.

……

……

……

……

……

24.3 EDUCATION

Education is an important instrument for empowering people, especially women. Education nurtures and develops their innate talents and capabilities and enables them to contribute effectively to the development process and reap its benefits. It is also an effective instrument for reducing social and economic inequalities. We have built up over the years a vast educational system in an attempt to provide education for all, to ensure that the skill and expertise needs of a developing economy are met satisfactorily and at the same time, to monitor the functioning of the educational system as an effective instrument for tackling inequalities. We would, therefore, like to look at data on different aspects of the educational and training system such as its size and structure, the physical inputs available to it for its effective functioning, its geographical spread, the type of skills and expertise it seek to generate, the access of different sections of society and areas of the country to it and the progress made towards the goals like 'total literacy', 'universalisation of secondary education', 'education for all' and 'removal of inequalities'.

The United Nations, in the year 2000, has put achieving of universal primary education as one of the eight Millennium Development Goals, to be achieved by 2015. Education plays an important role in shaping the post 2015 development agenda as well. The Government of India has enacted the Right to Education Act in August 2009. Therefore, the importance of research on the progress of the nation towards achieving these goals and its effects on the other social and economic goals is increasing by day.

The Department of School Education and the Department of Higher Education of the Ministry of Human Resources Development (MHRD), the National University of Educational Planning and Administration (NUEPA), National Council for Educational Research and Training (NCERT) and the University Grants Commission (UGC) collect and publish educational statistics, conduct research studies and surveys in the area of education. The International Standard Classification of Education (ISCED) is used worldwide to compile and compare cross-country statistics of education. In 2014, for the first time, the MHRD has developed the Indian Standard Classification of Education (InSCED), as a part of collection, dissemination and presentation of statistical data on education. Educational activities have been first classified into 16 broad levels from A through O and X. Levels A to E pertain to School Education, levels F to L pertains to Higher Education, levels M, N and O pertain to Certificate Courses, In-service training and Adult Education. The detailed structure can be accessed at http://mhrd.gov.in/sites/upload_files/mhrd/files/statistics/InSCED2014_0.pdf.

The annual publication **"Educational Statistics at a Glance"** provide a host of information on school education, like number of educational institutions, level and Stage-wise enrollment in school and higher education, teacher-student ratio, drop-out rates, examination results, public expenditure on education, etc. The MHRD also publishes a compendium on Universities in India, providing University-wise data on student enrollment, number of teachers, staff quarters, hostels, etc. The Annual Report of the MHRD also provides information on school and higher education in India. All these reports can be accessed free of charge at the website www.mhrd.gov.in in the sections documents and reports and the section statistics.

The NUEPA has developed a **District Information System of Education (DISE, can be accessed at www.dise.in)** to record data from the schools and junior colleges in all districts of India. Analytical reports based on this data, like school and facility related indicators, enrollment based indicators, teacher related indicators and EDI and analytical tables can be accessed at http://www.dise.in/AR.htm. These reports are available from the period 200-02 onwards, although the time series varies among topics. Provisional results of the **"Report of the eighth AIES", (http://www.aises.nic.in/surveyoutputs)** published by the NCERT, for the year 2009, in spite of being a bit dated, provides important data on school education statistics. While the data in the publication of MHRD are based on returns from the State Governments and Union Territory Administrations, the last one is based on a field survey. The University Grants Commission and MHRD provide data on university level courses in professional and technical courses. The MHRD has set up a **National Technical Manpower Information System (NTMIS) –** (seehttp://www.iamrindia.gov.in/_aboutNTMIS.htm) with lead centre at the Institute of Applied Manpower Research (IAMR), New Delhi and 21 nodal centres located at different States. It provides data on technical manpower – intake in and output from institutions and the utilization patterns of such output in detail through Tracer Studies[29] and other studies. These are disseminated through reports released from time to time by IAMR. **The Employment-Unemployment surveys of the NSSO and Education surveys conducted in the 52nd (1995-96), 64th (2007-08) and 71st (2014) Round of NSSO and the social and cultural tables of the population Census (ORGI)** also provide useful data on literacy and educational composition of the population or stocks of educated manpower of different levels of education and their utilization patterns – snapshots, at specific points of time, of the impact of the efforts in the area of education.

The kind of data available in these publications and sources are indicated below:

24.3.1 Educational Infrastructure

a) Levels of literacy and progress of efforts under literacy campaigns; (MHRD)

b) The number of institutions – colleges (including universities, deemed universities and institutions of national importance) and all types of general, technical and professional schools; (MHRD & UGC)

[29] Tracer Studies are, as the name suggests, trace or follow up specific cohorts or batches of students passing out of institutions to find out their present activity status. This helps in evaluating the trends over time in (a) the capacity of the economy to absorb the specific category of manpower in employment, (b) the waiting time for employment for the category of manpower, and (c) market value, in terms of salary levels in first employment, of the manpower category. These surveys are also called Cohort Studies.

c) The number of universities, deemed universities and institutions of national importance, the distribution of number of colleges by categories like those teaching arts, science & commerce, oriental learning and different categories of professional subjects like law, agriculture, engineering & technology and medicine and the distribution of schools in the area of general education (and stages like pre-primary and primary etc.,) and different categories such as those dealing with vocational, professional, special education etc. (MHRD & UGC);

d) Teacher (male and female and those belonging to SC/SCcommunities), the proportion trained among the, terms and conditions of their employment and attrition of the stock of teachers by cause. (AIES);

e) Patterns of management of colleges including professional colleges and deemed universities, recognition by appropriate bodies, availability of teachers and facilities like laboratory and equipment etc. (UGC);

f) Patterns of management (Government, local bodies, private management, religious and linguistic minority trusts/organizations and so on), medium of instruction, type of school buildings (*pucca, kutcha,* thatched hut etc.,), crowding (number of rooms in the building and the number used for instructional purposes), availability and adequacy of facilities like drinking water, labatories and urinals (and whether these are available separately for girls). (Sixth AIES);

g) Teachers in professional and technical institutions and the number of technical teacher training institutions. (MHRD & UGC);

h) The number of Industrial Training Institutions (ITIs) and Advanced Training Institutes, and the training capacity for apprenticeship training in industry under the Apprenticeship Act, 1961 for those passing out of ITIs and the vocational stream of schools. (MoLE and MHRD);

i) The number of institutions for training instructors for it is. (MoLE)[30];

j) The number of Vocational Rehabilitation Centres (VRCs) for the Physically Handicapped[31] set up by the DGE&T all over the country for giving adjustment training and placement in suitable employment through the Special Employment Exchanges for this group of people. (DGE&T[32]);

k) Expenditure on education by programmes like the *Sarva Siksha Abhiyan*, the Total Literacy Campaign and Adult Education; direct and indirect expenditure on recognized institutions of education. (MHRD).

24.3.2 Infrastructure Utilisation and Access to Educational Opportunities

a) Literacy rates – overall and age-specific, among SC/ST and adults; (Census – Social & Cultural Tables; NSSO & MHRD);

b) The number of students and the number of female students in the institutions specified in item 'b' above;

[30] Annual Reports of the Ministry of Labour, the part relating to DGE&T.

[31] The phrase "physically challenged" instead of the phrase "physically handicapped" is being brought into use nowadays in public discussions on the subject.

[32] See the preceding footnote.

c) The number of students and the number of girl students by courses and stages of education in recognized institutions – from the nursery class to the high/higher secondary level and in schools for vocational & professional education and special education and similar data for rural areas. (MHRD);

d) The number of male and female workers and marginal workers in the age group 5-9 and 10-14 – working children; (Census and NSSO); a comparison of this with the population in these age groups and data on school enrolment would lead to an assessment of the number of children who are neither in school nor in the workforce;

e) Enrolment and output in institutions specified in items 'h', 'I' and 'j' above;

f) Enrolment (and enrolment of female students) in university level general education courses by stages (degree, post graduate degree, diploma/certificate, research) and university level professional and technical education courses by faculty. (MHRD & UGC);

g) Similar information of the kind mentioned above for Scheduled Castes and Tribes (SC/ST) and females belonging to such sections of society. (MHRD & UGC);

h) Teacher-Pupil ratios at different levels of education. (MHRD);

i) Intake and output of graduates and post graduates in different disciplines and faculties. (MHRD & UGC);

j) Coverage of population in appropriate age groups by different stages of education. (MHRD);

k) Dropout rates at different stages of education. (MHRD);

l) Distribution of population attending educational institution by age, sex and type of educational institution for general and SC/ST population;

m) Access of the general and SC/ST population to (or availability of) facilities for different levels of education in rural habitations and urban settlements belonging to different population slabs, in terms of distance of the habitation from the facility; similar information regarding non-formal education (NFE) centers. (Sixth AIES);

n) Availability of special institutions suited to different types of disability of children in different villages and towns and the enrolment of disabled children in such institutions. (Sixth AIES);

o) Utilization patterns of different categories of professional and technical manpower in general and trends in waiting time for employment, utilization patterns and bargaining power for employees' compensation (salary etc.,) through tracer studies. (IAMR);

p) The disabled among main workers, marginal workers and non-workers by type of disability, age and sex. (Census);

q) Stocks of educated manpower and also selected categories of technical manpower for the general population, SC/ST and by religious communities;

r) The distribution of population by different levels of educational attainment (including illiterates, literates without any educational attainment) for population groups like SC/ST/Backward Classes (BC)/general households and those belonging to different religions, household types like (a) those with different sizes of land holdings; (b) households self-employed in agriculture, households self-employed outside agriculture, agricultural labour households, other labour households and households depending largely on regular wage/salaried employment outside agriculture; and (c) households belonging to different monthly per capita consumption expenditure (MPCE) classes. (NSSO and Census)[33].

One should not forget to mention in this context the **National Human Development Reports (NHDR)** prepared by the Planning Commission (now rechristened as NitiAyog) and the HDRs prepared by different State Governments as very useful sources for the manner in which available data on education, health, and other areas have been utilized to compute human development indicators and also as important sources for basic data on these sectors.

Check Your Progress 2

1) Name the organizations/institutions that collect and publish the data on different aspects of education.

……

……

……

……

……

2) What do you understand by the term 'Tracer Studies'? How are these useful?

……

……

……

……

……

3) Prepare a chart stating the kind of educational data available in the various documents published by MHRD and UGC.

……

……

……

……

[33] See also the sub-section (b) on education in the Section on 'Level of Inequality on Non-income Aspects of Life in Unit 4 on Poverty & Inequality: Policy Implications in Block 1 of MEC – 005 on Indian Economic Policy in your first year course.

24.4 HEALTH

One of the important dimensions of quality of life is health. A healthy individual can contribute effectively to production of goods and services. Investment in health is, therefore, an essential instrument of raising the quality of life of people and the productivity of the labour force. What is the health status of the population? What are the challenges to the health of the population and how are these being tackled? What kind of data is available about these aspects of the population, the health infrastructure and the efforts being made to deal with problems of health? What is the impact of these on the health situation, especially of women and children? The World Health Organisation (WHO), India website (http://www.who.int/countries/ind/en/) provides country statistics on health profile, nutrition, mortality and burden of disease and risk factors like those of alcohol and tobacco. The Central Bureau of Health Intelligence of the Ministry of Health and Family Welfare publishes the **"National Health Profile (NHP) of India"** every year. The publications **"Sample Registration System (SRS): Statistical Reports", "SRS Compendium of India's Fertility & Mortality Indicators, 1971-2001", Mortality Statistical and Cause of Death", SRS Bulletin** and the **Social & Cultural Tables (C Series Tables)** of Census 2001 of the Office of the Registrar General of India (RGI), Ministry of Home Affairs and the **"Report on the National Family Health Survey (NFHS-4) – 2014-15"** of the Ministry of Health and Family Welfare contain a large amount of information on these aspects of health, at the State level. For district level information, you can see the reports of the Annual Health Survey (AHS) of the ORGI and the District Level Health Survey (DLHS) of the MoH&FW. The Government of India launched the National Rural Health Mission in 2005, which has now been expanded to cover urban areas and called the National Health Mission. From 2005, the MoH&FW publishes, as an offshoot of the NHM, the Rural Health Statistics (RHS) report each year.

24.4.1 Health Infrastructure

The RHS of the MoH&FW provide the following types of data:

i) Demographic indicators – state-wise number of villages, rural and urban population, population growth rates and density, estimates of crude birth rates, crude death rates and infant mortality rates;

ii) Rural health infrastructure – number of health sub-centres in each 5-year plan period, number of Primary Health Centres, number of community Health Centres, number of sub-centres, PHCs and CHCs functioning at the end of the year, number of sub-divisional hospitals, district hospitals and mobile medical teams, shortfall in health infrastructure, building position of sub-centres, PHCs and CHCs;

iii) Health manpower in rural areas – Health workers (female)/ANM at sub centre, PHCs, Health workers (male) at sub centre, number of sub-centres/ PHCs without ANMs/Health worker (Male), health assistants/LHVs at PHCs, doctors at PHCs, number of PHCs without doctors/ lab technicians/ pharmacists, number of PHCs with AYUSH facility, surgeons at CHCs, obstrreticians and gynaecologists at CHCs, physicians at CHCs, paediatricians at CHCs, general dutty medical officers at CHCs, radiographers at CHCs, pharmacists at CHCs and PHCs, lab technicians at CHCs, nursing staff at PHCs and CHCs, etc.;

iv) Facilities available at sub centre – number of sub centres functioning, with ANM quarter, ANMs living in sub centre quarters, number of sub centres functioning as per IPHS norms, number of sub centres with regular water supply, with electricity, with all-weather motorable approach road, number of PHCs functioning, with labour room, with operation theatre, with at least 4 beds, regular water supply, with electricity, with all-weather motorable approach road, with telephone, with computer, with referral transport, registered RKS, functioning as per IPHS (Indian Public Health Standard) norms;

v) Facilities available at CHCs - number of CHCs functioning, with all four specialists, with computer/ statistical assistant for MIS/ Accountant, with functional laboratory, with functional OT, with functional labour room, with at least 30 beds, with functional X-ray machine, with quarters for specialist doctors, with referral transport, registered RKS, functioning as per IPHS norms, number of CHCs having a regular supply of allopathic drugs for common ailments, with AYUSH drugs for common ailments;

vi) Training of medical and paramedical personnel –ANM/ HW (F) training schools funded by Government of India, established by Government of India, health and family welfare training centres, MPW (M) training centres;

vii) Rural health care – some parameters of achievement – classification of States/UTs by average rural population covered by a sub centre, by a PHC, by a CHC, average radial distance covered by PHCs, ratio of LHV/ health assistant training schools to LHV/ health assistant, etc.;

viii) Expenditure on health and family welfare – overall and on individual programmes like malaria control, filarial control and national leprosy eradication programme.

24.4.2 Public Health, Morbidity and Mortality Indicators

i) Progress of the programme for vaccination of children and pregnant women;

ii) Time series on the number of notified/reported cases of and deaths due to diseases like cholera, small pox, acute diarrhoeal diseases, malaria, *kalaazar*, Japanese encephalitis and meningitis;

iii) The number of cases detected, treated and discharged in respect of diseases like leprosy and tuberculosis;

iv) The number of patients treated, discharged and deaths due to (a) different types of cancer in specialized cancer hospitals, (b) different types of mental diseases, and (c) communicable diseases like diphtheria, poliomyelitis, tetanus (neonatal and others), hepatitis and rabies;

v) Progress in the National Aids Control Programme and other National Control/Eradication programmes;

vi) Utilization by beneficiaries of facilities provided by CGHS, ESI scheme, Control/Eradication programmes;

vii) Incidence of morbidity and mortality by causes in zonal Railway hospitals;

viii) Incidence of morbidity by diseases in the ESI scheme;

ix) Causes of death statistics: distribution of deaths due to (a) selected diseases by age and sex, (b) specific causes under the group of diseases peculiar to infancy, and (c) causes related to childbirth, (d) causes related to childbirth and pregnancy (maternal mortality) and (e) accidents due to different types of natural and other causes (RGI);

x) Medical certification of cause of death – distribution of such deaths by age, sex and major cause groups (18 groups of causes) and the extent of coverage of such certification to total deaths in each of the 18 groups (RGI);

xi) Birth Rates (BRs), Death Rates (DRs) and natural growth rates and State-wise BRs and DRs (from SRS, RGI);

xii) Mortality Indicators like the Crude Death Rate (CDR), Infant Mortality Rate (IMR)[34], Nano-Natal Rate (NNR)[35], Post-Natal Rate (PNR)[36], Still Birth Rate (SBR)[37], Age-Specific Death Rates (ASDRs) and Maternal Mortality Rates (MMR) (from SRS, RGI);

xiii) Fertility data – 2001 (F series Tables) (census 2001);

xiv) Disabled population in India by type of disability, sex and age in general/SC/ST population and a further classification by marital status for the general population (from Census 2001); and

xv) Expectation of Life at Birth (e_0) [38], expectation of life at ages 10, 20, 30, 40, 50 and 60 for males, females and persons (from RGI).

24.4.3 National Family Health Survey (NFHS)

What is the impact of the efforts made to expand the size and reach and also enhance the quality of the health services on the health status of the family, especially of women and children? **The first National Family Health Survey (NFHS-1)** conducted in 1992-93 succeeded in answering this question and also in building up an important demographic and health database in India. This success paved the way for the conduct of the **second National Family Health Survey in 1998-99 (NFHS-2), third NFHS in 2005-06 and NFHS-4 in 2014-15,** to strengthen this database further and facilitate implementation and monitoring of population and health programmes in the country. **NFHS – 4 is covering, for the first time, all the 35 States and UTs as per Census 2011.** The survey is based on a representative sample of 5,60,000 households, up from about 1,90,000 households of NFHS-3. The survey provides State-level estimates of demographic and health parameters and also data on various socio-economic and programmatic factors that are crucial for bringing about desired changes in India's demographic and health situation.NFHS-1 and NFHS-2 were funded by the United States Agency for International Development, with supplemental funding from UNICEF. NFHS-3 funding was provided by the United States Agency for International Development, the Department for

[34] Relates to the number of children who die before their first birthday.
[35] Relates to the number of children who die in the first month of their life.
[36] Relates to children who die after the first month of life but before their first birthday.
[37] Relates to children who are born dead.
[38] The number of years that a person born today will, on the average live. Similary, e_i., (i=10, 20, 30, 40, 50, 60 ….) is the number of years that a person aged i years today will, on an average live.

International Development (United Kingdom), the Bill and Melinda Gates Foundation, UNICEF, the United Nations Population Fund, and the Government of India. Assistance for the HIV component of the NFHS-3 survey was provided by the National AIDS Control Organisation and the National AIDS Research Institute[39]. The field data collection in the NFHS is conducted by different research organisations. A total of 18 such organisations collected data in NFHS-3 during December 2005 to August 2006. It has been planned to conduct the NFHS-4 through 11 reputed field agencies like the AMS, DRS, EHI, GFK, GIM, IIHMR, NIELSON, SPVM, RDI, SPYM and VIMARSSH. **NFHS-3** provides

a) urban and rural estimates for most States, (b) separate estimates for 8 cities, and (c) estimates for the slum areas. **Besides a national report, reports are prepared for the States**. The NFHS reports can be downloaded free from the site http://www.rchiips.org/nfhs/sub_report.shtml. The kind of data on health and nutrition status of women and children and estimates of parameters related to these presented by the NFHS, are indicated in brief below:

i) Educational level of the household population, school attendance of boys and girls and reasons for not attending school;

ii) Distribution of housing by availability of basic amenities[40], ownership of agricultural land, house and livestock and durable goods; standard of living indicators based on these and habits like drinking, tobacco and smoking;

iii) Distance of households from the nearest health facility and distribution of rural residents living in villages that have selected facilities and services;

iv) Age at first marriage of (women) respondents, their exposure to mass media, their employment status and aspects of their empowerment or lack of it, including domestic violence;

v) Current fertility, variation in current fertility by various factors, fertility trends, pattern of outcome of pregnancy, median number of children ever born and living to ever married women, patterns of birth order and birth intervals, median age of women at the first birth and the last birth of child, patterns of fertility preference and sex (of the child) preference;

vi) Knowledge of use and time of first use of contraception, reasons for discontinuation of, or for not using contraception;

vii) Estimates of age-specific death rates, crude death rates, infant and child mortality[41];

viii) Morbidity by selected diseases and its variation over States;

ix) Vaccination of children, vitamin A supplementation for children, prevalence of acute respiratory infection, fever and diarrhoea, treatment of diarroea and awareness of treatment like ORS packets;

---

[39] See the website http://www.rchiips.org/nfhs/nfhs3.shtml which provides details of all the 4 NFHS surveys conducted so far.

[40] Electricity, source of drinking water, time taken to get water, method of purifying water, sanitation facility, fuel for cooking, type of house (*pucca*, semi-*pucca* and *kutcha*) and persons per room.

[41] Child mortality relates to children who die between their first and 4th birthday. Under-five mortality relates to children who die between their 4th and 5th birthdays.

x) Knowledge about AIDS and ways of avoiding it;

xi) Food consumption of women, nutritional status of women, anaemia among them, iodisation of salt consumed in households and body mass index[42] - an indicator of the heath status of women;

xii) Median duration of breastfeeding – overall and in different States, type of food consumed by children, nutritional status of children, anaemia among them and indicators of acute and chronic malnutrition among children – weight for age index, height for age index and weight for height index[43];

xiii) Health problems of pregnancy, antenatal care, assistance during delivery, place of delivery, post-*partum* care, and care and treatment of reproductive health problems; and

xiv) Couple protection rate[44].

It is necessary to make a reference here again to the NHDR 2001 of the Planning Commission and the HDRs of the State Governments for their importance in the field of health, education and other sectors.

Check Your Progress 3

1) Indicate the publications which contain the data on different aspects of health.

……

……

……

……

……

2) What type of database is provided by National Family Health Survey?

……

……

……

……

[42] "Body mass index (BMI) = [weight in kgms.] / [height in metres)(height in metres)]. See Unit 4 on Poverty & Inequality – Policy Implications" (section on human development indicators) in Block 1 of MEC – 005 on Indian Economic Policy in your first year course.

[43] The National Institute of Nutrition, Hyderabad has recommended that the nutritional status of the international reference population recommended by the WHO could be used as the standard for India. The **Weight for Age Index** is a composite measure that takes into account both chronic and acute malnutrition. Children who are more than 2 standard deviations (s.d.) below the median value of the index for the reference population are "underweight". The **Height for Age Index** measures linear growth retardation. Children who are more than 2 s.d. below the median value of the index for the reference population are considered short for their age, or are "stunted". This is chronic undernutrition. The **Weight for Height Index** examines body mass in relation to body height. Children who are more than 2 s.d. below the median value of the index for the reference population are considered too thin or "wasted". This is acute undernutrition. (NFHS – II). See the preceding footnote on BMI also.

[44] The percentage of currently married women aged 15 to 49 years using family planning methods at the time of the survey.

……

3) What do you mean by the term 'weight for age index'?

……

……

……

……

……

24.5 SHELTER AND AMENITIES

Another important dimension of the quality of life is shelter and access to amenities like safe drinking water, toilet, adequate lighting and safe fuel for cooking. We have already discussed this in sub-section (d) under the section on "Levels of Inequality in Non-income Aspects of Life" in Unit 4 on "Poverty and Inequality: Policy Implications" in Block 1 of MEC – 005 on "Indian Economic Policy" in your first year course. The population Census, from 1991 onwards, have collected data on housing and amenities of the population during the house-listing and housing census phase, i.e., in the years 1990, 2000 and 2010. From Census 2001 onwards, results on parameters related to this aspect is disseminated for the urban slums as well. While Census 2001 results on urban slums relate to the towns with 20,000 or more population, all statutory towns were considered for identifying the urban slums and disseminating results for them. The results based on Census 2011 on housing stock, amenities and assets, at district level can be downloaded from http://www.censusindia.gov.in/2011census/hlo/HLO_Tables.html. This site alsso provides such tables for the urban slums. While the Population Census provides data at a much disaggregated level, a higher quantum of information is provided by the surveys conducted by the NSSO in their surveys on housing condition and condition of urban slums. The NSS survey reports based on surveys conducted in the 49th round (1993), 58th round (2002) and 69th round (2012)provide results at State/UT level on these parameters. The 69th round of NSS survey has published three reports – (i) "**Key indicators of drinking water, sanitation, hygiene and housing condition in India"**, (ii) **"Drinking water, sanitation, hygiene and housing condition in India"** and (iii) **"Housing condition and amenities in India"** – contain the latest data on housing conditions including those of dwelling in slums. The **NFHS** also gives some data on housing and amenities. These sources of data provide the following kind of information (data presented for rural/urban/urban slum/other urban areas in the case of items a to f):

a) Distribution of households by type of structure –*pucca*, semi – *pucca* and *kutchha*;

b) Access to toilet facility – SC/ST/other households – within the premises or not; type of latrine;

c) Access to safe drinking water – SC/ST/other households, major source of drinking water;

d) Use of electricity for lighting – SC/ST/other households;

e) SC/ST/other households with **no** access to electricity, safe drinking water and toilet;

f) SC/ST/other households with access to electricity, safe drinking water and toilet;

g) Percentage of villages connected by roads[45];

h) Type of slum – notified or non-notified -, availability of infrastructure like sewerage system, drainage system, means of garbage disposal and *pucca* road(s) within and approaching the slum, how has the slum been there, and whether the land on which the slum is located is owned by Local Bodies, State Government, etc.;

i) Duration of residence of the household in the slum, reasons for coming to the slum, the place from which the household came, whether the household had any document of identification and whether it had at any time tried to leave the slum;

j) Possession of durable goods and rent being paid for the premises; and

k) Workforce participation rates, female workforce participation rates, workers and non-workers and workers by broad categories of economic activity, including household, industry; and demographic features of the slum population like sex ratio – overall and for children in the age group 0-6, etc.

24.6 SOCIAL CONSEQUENCES OF DEVELOPMENT

The development experience of the last few decades show that the cost of development is not shared equally by all sections of the society. Often the burden of the cost falls almost entirely on the poor and the voiceless. The best example of inequitable sharing of the burden of the cost of development programmes is provided by the current controversy over the *Sardar Sarovar Project (SSP)* over the river *Narmada* and the struggle of the *Narmada Bachao Andolan* (NBA) on behalf of the people displaced by the construction of the dam and the consequent submerging of their villages and their lands or are likely to be displaced with each installment of rise in the height of the dam. The other recent example that invited wide attention is the acquisition of the lands of tribals in Orissa for the industries in Kalinganagar in the State. The problem of people ousted from the land submerged by the water filling the reservoirs created by the dams or, for that matter, people whose lands are acquired by the State for purposes of development (building factories, dams etc.,), called "oustees", has been coming up ever since the first development projects were taken up in 1951. The State promised jobs to the oustees in the industries that came up in their lands in addition to compensation for the land acquired or alternative land to prevent loss of their livelihood. The question that has continued to agitate the minds of most development analysts is about the extent to which the rehabilitation of the displaced or the "oustees" has been successfully completed. What data are there on this vital question?

It may be mentioned that "oustees" were one of the priority categories for registration as jobseekers in the employment exchanges and placement in employment. The DGE&T, the organization concerned with employment exchange at the national level, is one source where some information on (i)

[45] Roads are all categories of roads surfaced or unsurfaced, district road, highways and rural roads. (**Basic Road Statistics of India, Ministry of Surface Transport, New Delhi**.)

how many of them were registered as job seekers under these categories, (ii) how many of these were actually placed in a job and (iii) how many languished on the "live registers" of employment exchanges for a long time before passing into oblivion, because if he or she did not renew her registration at specified intervals he or she would have ceased to be on the "live" register of job seekers. Past data with the DGE&T in the Ministry of Labour might show the proportion of those who just ceased to be on the register. This source of data, however, would have represented a very small part of whatever rehabilitation was done, because the kind of vacancies that can be filled through employment exchanges is restricted to jobs in Government not filled through competitive examinations, public service commission and other commissions/recruitment boards and jobs up to a specified salary level in public sector undertakings. The private sector is not obliged to fill its vacancies through the employment exchanges. Other than this source, Government – the Central or the State Governments – do not seem to have any data on the number of families which were displaced and the number rehabilitated. Data maintained by the State Governments concerned with the implementation of rehabilitation under, for example, the *Sardar Sarovar Project*, do not seem to reflect the actual situation on the ground, as has been shown by the Report of the Group of Ministers which toured some of the affected areas in Madhya Pradesh in April, 2006[46].

It was in 1994 that a **study supported by the Indian Council of Social Science Research (ICSSR)** attempted estimates of the extent of displacement of families between 1951 and 1990 due to projects like mines, dams, industries and wild life sanctuaries. It also estimated the proportion of those displaced in the Fifties and the Sixties who were resettled till 1980. It pointed out that there was practically no improvement in the Eighties. A **study on development-induced displacement** in West Bengal, between 1947 and 2000, conducted by a team led by **Walter Fernandes** has made estimates of (a) the number of people adversely affected by projects, (b) the number physically displaced, (c) the number resettled by the projects, (d) the rest that had been left to fend for itself following displacement, and (e) the proportion of those in (d) who were *dalits* and tribals. Yet another **study published in Economic and Political Weekly by Manipadma Jena** quotes official figures (of the Government of Orissa) of families displaced due to development projects between 1950 and 1993 and of the extent of land that had to be acquired by the Government for the purpose. Jena's study also presents data on (a) the number of villages affected by the Hirakud Dam that was constructed between 1948 and 1957, (b) the number of families and people whose livelihood was disrupted due to displacement, (c) the number of families (all belonging to Scheduled Castes/Tribes) which were displaced forcibly by police, and (d) the number of families resettled in rehabilitation camps[47].

This aspect of development projects will cast greater responsibilities on the respective project authorities in these matters and closer monitoring of the implementation of resettlement plans. These will necessarily require the setting

---

[46] *The Hindu*, Chennai edition of Monday, the 17th April, 2006 carried the full text of '**the Group of Ministers' (GoM's) Report** on the OP-ED page.

[47] The studies referred to in this paragraph and their contents are taken from the article **"Creating Dispensable Citizens" by Usha Ramanathan, in the Chennai edition of *The Hindu of Friday April 14, 2006***.UshaRamanathan is an Honorary Fellow of the Centre for the Study of Developing Societies, New Delhi. The estimates made by the studies are given in the article. These have not been indicated here as the purpose of this Unit is only to indicate the very few sources of data that are available on this important social aspect of development and not to discuss the problem.

up of a reliable and transparent system of collection, compilation and analysis of statistics relating to displacement of people due to the project, their demographic, social, cultural and economic profiles, their resettlement in alternative livelihood, number absorbed in employment in the industrial establishment(s) for which the lands of the oustees were acquired, details relating to the levels of living of the oustees in the area of resettlement in terms of parameters like incomes, access to basic needs of life like water, shelter, education for their children and health and medicare at least until a reasonable number of years after resettlement, the time taken and the expenditure incurred for resettlement and so on.

24.7 ENVIRONMENT

The process of development adversely affects the environment and through it the quality of life of society. For instance the excessive use of fertilizers and pesticides rob the soil of its nutrients. Letting sewers and drainage and industrial effluents without prior treatment into rivers and water bodies pollute these water bodies causing destruction of aquatic life and endanger the health of people using such polluted water. The recent outcry in Tirupur, near Coimbatore, a place well known for garment exports, against untreated effluents from garment factories being let into the river used for drinking purposes is a case in point. The exhaust fumes containing Carbon Monoxide (CO) and lead (Pb) particles let in to the air we breathe by vehicles using petrol or diesel is an example of air pollution. The best example of industrial pollution through insufficient safety measures is the Bhopal gas disaster where lethal gases leaking from a factory's storage cylinder killed many people immediately and maimed many others for life. The forest cover of the country is continuously getting reduced due to indiscriminate felling of trees leading to reduction in rainfall and changes in rainfall pattern, besides climatic changes. The destruction of mangroves along seacoasts for housing/tourism development often leads to soil erosion along the coast by the sea. The adverse effects of current models of development on environment and the realization of the need to take note of the cost to development represented by such effects have now led to the development of environmental economics as a new discipline in economics.

The system of national accounts currently in use throughout the world, suffers from extreme narrowness. Vast quantities of information relevant for economic evaluation do not appear in them. Some don't because the appropriate data are hard, even impossible, to collect; but others don't because until recently the theory and practice of economic evaluation didn't ask for them. The demand for green national accounts has arisen because of a growing recognition that contemporary national accounts are an unsatisfactory basis for economic evaluation. The qualifier "green" signals that we should be especially concerned about the absence of information on society's use of the natural environment. As an extension of national accounting, the development of green accounting framework and related researches are therefore being undertaken around the world. The CSO, for the first time in India, have published the Report "Green National Accounts in India" in the year 2013. You can download a copy of this report from the Social Statistics Division of the CSO.

The Central and State Pollution Control Boards and the Ministry of Environment and Forests (MOEF) evolve and, monitor implementation of policies to protect the environment. Statistics on environment are collected through this process by the agencies mentioned above and the CSO. **The**

annual reports of the MOEF[48] and the Compendium on Environment Statistics, India 2003 published by the **CSO** from time to time are excellent sources of data on environment. The latter especially is very comprehensive and includes a very informative write up. The Compendium (and the annual report of MOEF) can be accessed in the respective **website of the two organizations**. The type of data on environment available from these publications are mentioned below by way of illustration:

a) Ambient Air Quality Status [concentration of Sulphur di-oxide, Nitogen di-oxide and Solid Particulate Matter[49] (SPM) in air] in major cities of India;

b) Percentage of petrol-driven two-wheelers, three wheelers and four-wheelers meeting CO emission standards; and

c) Water quality of Yamuna river (in the Delhi stretch) in respect of selective physio-chemical parameters between April, 1998 and March, 1999 – dissolved oxygen (milligrams/litre), Biological Oxygen Demand (BOD) (mg/l), faecal coliforms[50] (number/100ml), total coliforms (number/100ml) and ammonical nitrogen (mg/l).

24.8 QUALITY OF LIFE

We have already looked at several of the factors determining the quality of life of the people – education, health, employment, shelter and amenities and environment. We have also referred to the plight of displaced people uprooted from their normal life by development projects. One other factor, an important one, is the level of income or consumption. We have already looked at the data available on this aspect of life in the Unit on "Poverty and Inequality: Policy Implications" in Block 1 of MEC-005 on "Indian Economic Policy" – A compulsory course in first year. The relevant data are available from successive quiquennial surveys of the NSSO on consumer expenditure, namely, those on levels of consumption of different MPCE classes. That Unit also discusses dimensions of poverty and inequalities in income (consumption) and non-income aspects of life and about HDIs measuring shortfalls in human development in the population, Gender Development Indices (GDI) measuring gender discrimination, BMIs evaluating the health status of women and the measures, Weight for Age Index Height for Age Index and Weight for Height Index gauging the nutritional status of children. All these measures are also available from these sources, namely, NSSO, the Planning commission and the National Human Development Report 2001 and those of the State Governments for judging the quality of life of the Scheduled Castes and Tribes. The reports (prepared every year) of the Commission for Scheduled Castes and Tribes provide (review) data on:

i) the progress in education of these sections of society;

ii) Progress in the utilization of the reservation quotas in different services of the Governments and the public sector undertakings and the efficacy of measures taken by Government in enabling members of the weaker sections of society in gaining access to these opportunities;

[48] The latest report of 2013-14 can be accessed at http://www.moef.nic.in/sites/default/files/annual_report/AR-2013-14-Eng.pdf.

[49] SPM consists of metallic oxide silicon, calcium and other deleterious metals.

[50] The most common contamination in water is from disease-bearing human wastes which is usually detected by measuring faecal coliform levels.

iii) enforcement of laws like the Untouchability Act; and

iv) difficulties of these sections of society – ill-treatment, practice of untouchability, discrimination and lack of access to financial and technical assistance to entrepreneurs belonging to these classes.

The Ministry of Social Justice and Empowerment, the Ministry of Tribal Affairs, the National Commission of Backward Classes (NCBC), the State Backward Class Commissions, the Minorities Commission, the National Commission on Women and the State Commissions on Women also conduct surveys, bring together and review a large amount of data on the progress of or the lack of progress in curbing exploitation of, and violence against these sections of society, raising the shares in education and employment of these communities and groups and in empowering and bringing these sections into the mainstream of life. These are comprehensive sources of data on the efforts being made and need to be made to empower these sections of society and elevate their status in society. The Commission for the Aged (Senior Citizens) and the Commission for Children are sources of data on these sections gathered at one place from various primary sources[51] and analysed to highlight issues that need attention. The Ministry of Social Justice and Empowerment (Annual Report of the Ministry), the nodal agency that is entrusted with the responsibility for the development and empowerment of the physically and mentally challenged, the weaker sections of society and women and the development of children and the welfare of the aged, constitute another source of data on the status of these vulnerable groups in society.

Check Your Progress 4

1) What kind of data/information is provided by NSSO on housing conditions in India?

……

……

……

……

……

2) Do you think that the data on registrants with the employment exchanges provide adequate information about the persons displayed by development projects? Give reasons in support of your answer.

……

……

……

……

[51] For instance, for the aged, important sources of data are (a) the NSSO Report No. 446 (52nd Round) "The Aged in India – A Socio-economic Profile: 1995-96". The Census 2001 is a mine of information on the Aged as it presents a number of social, economic and cultural characteristics by age groups and gender, one such group being 60+. Data on women and children and the disabled also flow from the Census and other sources of data for these grops have already been referred to elsewhere in this unit.

3) Name the documents, which contain the statistics on environment.

……

……

……

……

……

4) Which department/ministry has been assigned the responsibility to provide information data on development and empowerment of weaker sections of the society?

……

……

……

……

……

24.9 LET US SUM UP

We have tried to enumerate the various sources of data in the social sector and looked at the kind of data available in different sub-sectors of the social sector. We have seen that there are a number of sources of data on employment covering parts of the economy. There are two sources that are comprehensive and cover the entire economy. These enable an analysis of trends in employment and unemployment over the years, the industrial, occupational and educational composition of the employed and their age profile, the educational profile and the age structure of the unemployed, the quality and adequacy of employment in terms of wages, tenure of employment and employment security of workers. Similar analysis is possible for several subsections of the economy and society that are relevant to the country's objectives of economic and social policies. The database on welfare of labour is somewhat incomplete in some areas, calling for stricter monitoring of the submission of statutory returns and the implementation of existing legislation. Enormous amount of data on educational infrastructure, the utilization of such facilities and the output of the educational system are available and periodic comprehensive surveys add to the richness of the database. However, there is a need to go into such discrepancies as may be there between data collected by the reporting system in the educational administration and the data thrown up by the surveys, to initiate corrective action to tone up the large database built up over the years. Another important need is to reduce the time lag in the availability of data for policy makers as well as researchers. Data on health are being built up from more than one source, namely, the Registrar General of India on aspects like vital rates such as birth and death rates, IMR, MMR and morbidity rates and life expectancy, while the CBHI collects and brings together data from the different wings of the Ministry of Health and the State Health Departments and the different health professional councils. Health and Medicare and health situation data are comprehensive in terms of item coverage but time lags need to be reduced so that data relating a particular time point is complete in terms of coverage of all States and Union Territories. Data from NFHS is a valuable

addition to the database on health, especially from the point of view of health and nutrition status of women and children. The database on shelter and amenities has been updated recently by the Census 2011 and the NSSO survey of the 69th round (July – December, 2012) and also includes specific data on urban slums. A database on environment – air pollution, soil degradation and noise pollution – is getting built up steadily. Data on levels of living, both in terms of income and non-income aspects of life and for the general population as well as for the weaker sections of society can be derived from existing data and used as a basis for planning programmes to reduce poverty and inequalities.

We have also noted a grey area in the data systems we have considered. There is no data worth the name on people displaced from their livelihoods by development projects and the number among these who have been resettled in alternative livelihood. There is a growing realization, at least among some sections of development analysts and thinkers that the burden of the costs of development is invariably borne mostly by the poor and the voiceless. There is an urgent need to build up a transparent database on people displaced by development projects and their resettlement and the basic amenities made available to them.

24.10 SOME USEFUL BOOKS

Ministry of Environment and Forest (2009) : State of Environment Report, India

Ministry of Statistics &Programme Implementation (2013) : Green National Accounts in India – A Framework

Ministry of Labour and Employment (2010) : Indian Labour Yearbook 2009 and 2010

24.11 ANSWERS OR HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) i) Employment Market Information (EMI) Programme of Directorate General of Employment and Training (DGE&T), Ministry of Labour
   ii) State Directorates of Employments (DES)
   iii) Economic Census
   iv) Annual Survey of Industries
   v) Population Census
   vi) Quininquennial Sample Surveys on employment and unemployment conducted by NSSO.
2) See Sub-section 24.2.1
3) Usual Principal Status (UPS), Usual Principal and Subsidiary Status (UPSS), Current Weekly Status (CWS) and the Current Daily Status (CDS).
4) See Sub-section 24.2.3

Check Your Progress 2

1) The department of School Education and higher education of Ministry of Human Resource Development, NCAER, UGC, Institute of Applied Manpower Research etc.

2) See footnote no. 29

3) Do yourself

Check Your Progress 3

1) See Section 24.4

2) State Level estimates of demographic and health parameters and the data on various socio-economic factors crucial for bringing changes in demographic and health situations.

3) Composite measure that takes into account both chronic and acute malnutrition.

Check Your Progress 4

1) See Section 24.5

2) See Section 24.6

3) The annual Reports of Ministry of Environment and Forest and the compendium of environment statistics.

4) Ministry of Social Justice and Empowerment (annual reports of the Ministry).