
UNIT 1 QUANTITATIVE DATA COLLECTION METHODS AND DEVICES

Structure

- 1.1 Introduction
- 1.2 Primary Data Collection: Meaning and Methods
- 1.3 Questionnaire Method of Data collection
- 1.4 Interview Schedule
- 1.5 Secondary Data Collection Methods
- 1.6 Let Us Sum Up
- 1.7 References and Selected Readings
- 1.8 Check Your Progress - Possible Answers

1.1 INTRODUCTION

There are two types of primary research: one is done through quantitative data collection and the other, through qualitative data collection. Customarily, quantitative data collection means using numbers to assess information. As you are aware, some kinds of information are numerical in nature, for example, a person's age, or annual income. The answers to these questions are in numbers.

Quantitative data is used for testing of a hypothesis and drawing inferences. Quantitative data is collected by using the following two set of data resources:

- i) Primary data
- ii) Secondary data.

In this unit, we will discuss in detail, methods of collecting primary and secondary data, along with the advantages and disadvantages of the methods.

After reading this unit, you should be able to

- explain the primary data collection methods
- discuss the questionnaire and interview methods of data collection
- describe secondary methods of data collection

1.2 PRIMARY DATA COLLECTION: MEANING AND METHODS

Data which are originally collected by the investigators are called primary data, while the secondary data are collected through some other sources. For example, information collected by an investigator from a student regarding his class, caste, family background, etc., is called primary data. On the other hand, if the same information collected about the student from the school record and register, then it is called secondary data.

However, the difference between primary and secondary data is largely of degree, and there is hardly any watertight difference between them. The data collected through primary sources by one investigator may be secondary in the hands of

others. For example, field data collected by an investigator for writing his thesis is primary to him, and when the same data is used by another investigator, for reference purposes, then it became secondary data. Let us discuss the methods that are used to collect primary and secondary data.

There are various types of quantitative primary data gathering tools, but the important ones among them are:

- The Questionnaire
- The Interview Schedule

1.3 QUESTIONNAIRE METHOD OF DATA COLLECTION

Questionnaires are a popular method of data collection. Although they look easy, it is difficult to design a good questionnaire. Careful design of a questionnaire is vital for the collection of required facts and figures. Any frivolous attempt in framing a questionnaire will lead to either shortage, or, collection of unnecessary information neither of which will be useful to your research. Questionnaire design depends on whom, and, from where information has to be collected; what facts and figures need to be collected; and the calibre of the informants.

The questionnaire can be broadly categorized into two types:

- i) structured questionnaire
 - ii) unstructured questionnaire.
- i) **Structured questionnaires** are prepared in advance. They contain definite and concrete questions. The structured questionnaire may contain close ended questions and open ended responses. In the close ended questionnaire, the question setter gives alternative options for which the respondent has to give definite response. The best example of the close ended questionnaire format is the one that leads respondents to the “Yes” or “No”/ “True” or “False” answers.
 - ii) **Unstructured questionnaires** are those that are not structured in advance, and the investigators may adjust questions according to their needs during an interview.

1.3.1 Methods of Data Collection Using Questionnaires

Questionnaire methods are conducted in different ways. A few important methods are outlined here.

- i) **Personal Interview**

In personal interviews, the interviewer or investigator personally approaches the interviewee and administer questions. This method is largely followed in research and the accuracy of data is very high. However, it is an expensive method.

- ii) **Mail Questionnaire**

In this method, the investigator mails the questionnaire to respondents and respondents are requested to fill it up and return it to the investigator. In many cases, a self addressed stamped envelope is sent along with the

questionnaire to facilitate the return of the questionnaire mail immediately. This method is usually adopted where the respondents are widespread and the investigator has limited resources to approach them. The success of this method depends of the literacy level of the respondents and the accuracy of the address database. One of the drawbacks of this method is that, sometimes, the respondents do not take the questionnaire seriously, and, as a result the answer may not be accurate.

Implementing a Mail Survey

- Design a written questionnaire with identification number.
- Pretest questionnaire to assure validity and reliability.
- Select sample population.
- Two weeks before mailing the survey, send an advance letter
- Mail the questionnaire including a cover letter and a stamped, self-addressed envelope
- Send a postcard a week or so later, thanking those who responded and reminding those who did not return the questionnaires
- Three weeks after mailing the first questionnaire, send a follow-up letter stating that a response has not been received, including a replacement questionnaire and a stamped, self-addressed envelope.
- In developing the mailing schedule avoid holidays.
- For most purposes, a 60 to 90 percent return rate is considered satisfactory.

(Source : Suvedi et.al., 2008)

iii) Telephone

In this method, the investigator administers a questionnaire by seeking responses from the respondent over the telephone. It is largely administered to the urban respondents where telephone facilities are widely available. However, the success of this method depends on the availability of telephone with the respondents. It is also expensive as well.

Implementing a Telephone Survey

- Arrange the facilities for survey.
- Identify the sample and their telephone numbers.
- Send an advance letter if addresses are available with information on when you will be likely to contact respondents, during working or non-working hours and how much time you need.
- Prepare well on the background information about the survey to answer respondents questions, if any.
- Develop an interview schedule.
- Decide on the number of calls to make to each number. In local surveys six to seven calls are customary.
- Decide how to handle refusals.
- Stick to the time schedule

Sample Call Sheet for Telephone Interviews

A call-sheet is used for each number chosen from the sampling frame. The interviewer records information that allows the supervisor to decide what to do with each number that has been processed. Call sheets are attached to questionnaires after an interview is completed.

Telephone Interview Call Sheet				
Survey title :				
Questionnaire identification number				
Area code & number () &				
Contact attempts	Date	Time	Result code & comments	Interviewer I.D.
1				
2				
3				
4				
5				
6				
Additional comments				
Code	Result Codes			
	No answer after seven rings			
	Busy, after one immediate redial			
	Answering machine (residence)			
	Household language barrier			
	Answered by nonresident			
	Household refusal			
	Disconnected or other non-working number			
	Temporarily disconnected			
	Business or other non-residence			
	No one meeting eligibility requirement			
	Contact only			
	Selected respondent temporarily unavailable			
	Selected respondent unavailable during field period			

	Selected respondent unavailable because of physical/mental handicap
	Language barrier with selected respondent
	Refusal by selected respondent
	Partial interview
	Respondent contacted - completed interview
	Other
Sample Help Sheet for Interviewers	
Name of sponsoring agency:	
Purpose of study:	
Contact person for survey:	
Size of survey:	
Identity of interviewer:	
How respondents name was obtained:	
Issues of confidentiality:	
How to get a copy of results:	
How will results be used:	
(Source : Suvedi et.al., 2008)	

iv) E-Mail

With the IT revolution, nowadays, questionnaires are attached to the e- mails and sent to respondents who send an answer through return e-mail. The success of this method depends on the availability of internet facilities.

1.3.2 Qualities of a Good Questionnaire

Questionnaire framing is the most arduous task in social science research. Careful framing of questionnaires is essential to obtain reliable data,. Some of the principles that need to be taken into consideration while framing a questionnaire follow.

- i) **It must be simple:** the questions must be simple and straightforward. They must also be short, which could be easily answered.
- ii) **Begin with a covering letter:** the front page of the questionnaire must contain an introduction to investigator or institution collecting data, and the purpose of the quest. If the questionnaires are to be returned by mail, then, the address to which they are to be sent must be clearly mentioned.
- iii) **The number of questions must be kept to a minimum:** the questions asked in the questionnaire must be kept to a minimum and restricted to the subject and topic of the study. Any questions which do not have direct bearing on the problem must be avoided.

- iv) **Minimum use of Technical Terms:** try to avoid the technical terms as far as possible. If abbreviations are used, they need to be explained with illustrations, either separately or in the questionnaire itself. However, the investigator should be conversant about those technical terms.
- v) **Questions must be logically arranged:** here lies the acumen of the investigator or question setter. He, or, she must arrange the questions in such a way that such questions should flow naturally from the answer to the previous question.
- vi) **Avoid asking controversial questions:** do not include questions which are controversial in nature, or, are too personal or specific to community sentiments. Hypothetical questions, too, need to be avoided.
- vii) **Pre testing of questionnaire:** before final administration, questionnaire need to be pre tested among a small number of respondents. This will give an opportunity to the investigator to rectify the problems, and, if required, any addition and deletion of questions.

1.3.3 Physical form of Questionnaire

While designing the questionnaire, the physical form of the questionnaire may be meticulously prepared. The following factors needed to be taken into consideration.

- i) **Size:** the size of the questionnaire depends on the scope of the study. Adequate space should be provided for recording the comments and suggestions of the respondents. However, a single space is needed provide for recording the response. The Coding of questionnaire will reduced the need for space. Taking all these factors into consideration, the size of the questionnaire can be fixed, accordingly.
- ii) **Quality of the paper:** good quality paper should be used in the question so that it lasts for a longer period. Except for the front page, white papers may be used in other pages.
- iii) **Covering Letter:** Every questionnaire must have a covering letter. The purpose of the questionnaire must be clearly mentioned. Assurance should be given that the information gathered will be used only for research purpose, and be given confidential treatment.

1.3.4 Advantages of Questionnaires

The advantages of questionnaires are:

- i) they are less expensive compared to the interview schedule and can be administered over a large number of respondents.
- ii) they are less time consuming.
- iii) since the interviewer is not present during the administration of a questionnaire, respondents may feel freer and have greater confidence in answering questions
- iv) one of the advantages of the questionnaire method is that once it is standardized, then, the information collected from the respondents becomes more uniform.

1.3.5 Disadvantages of Questionnaire

Some disadvantages of questionnaires are :

- i) In a questionnaire, there is no personal contact between the investigator and the respondents because of which clarifications on responses, if needed, cannot be sought.
- ii) a questionnaire is not a suitable mode when a spontaneous answer is required through probing
- iii) it is possible that the investigator may not get a response for all questions. Sometimes the responses may be vague and provide incorrect information
- iv) there is the chance that information may be manipulated.

Sample Questionnaire	
Indira Gandhi National Open University School of Extension and Development Studies PG Diploma in Urban Planning and Development	
Title: Functioning of Primary School in Municipality	
1)	Name of the State
2)	Name of the District
3)	Name of the Block
4)	Name of the Municipality
5)	Name of the Teacher (Respondent)
6)	Sex: Male/Female7. Age
8)	Educational Qualification
9)	Caste 10. Marital Status
11)	Years of Teaching
12)	Training received, if any
13)	If yes, write the subjects taught in the training programme
14)	Subject you are teaching
	Mathematics
	Science
	Literature
	Any other; specify
15)	Medium in which you are teaching
	English
	Hindi
	Any other, specify

16) In your opinion, which students were performing better in the class:			
General Caste			
SC			
ST			
Girls			
Boys			
17) Your interaction with the			
Categories	Frequent	Occasional	Not at all
Parents			
Father			
Mother			
18) The role of municipality in your school management			
Good			
Average			
Poor			
19) Functioning of the Education Committee			
Good			
Average			
Poor			
Areas of their involvement			
20) In your opinion, who are the real beneficiaries of education?			
Economically Poor			
Girl Children			
Socially backward			
All			
21) Write the main problems of your School			
1			
2			
3			
22) What are your suggestions for improvement of the school conditions?			
1			
2			
3			

In this section, you studied about quantitative data collection and the questionnaire method of data collection. Now, answer the questions given in *Check Your Progress-1*.

Check Your Progress 1

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) What is primary data?

.....

.....

.....

.....

.....

2) What are the advantages of a questionnaire?

.....

.....

.....

.....

.....

1.4 INTERVIEW SCHEDULE

Interview schedule is a common practice in research. The schedule puts the whole thing in a structured form, so that the tabulation and analysis of data become easier. The basic difference between the schedule and the questionnaire is that in case of former, the presence of a field investigator or interviewer is a must, while the same may not be mandatory in the latter case. In other words, in the case of the interview schedule, the field investigator is an essential component. In other aspects, there is not much difference between the interview schedule and the questionnaire.

Some important aspects need to be taken into consideration in the preparation and execution of interview schedule.

- i) **Selection of Respondents:** The selection of respondents is the key to interview schedule administration. The respondents are selected through various sampling methods and their names and addresses are noted down. The field investigator approaches them for data collection by filling up the interview schedule.
- ii) **Training of the Field Staff:** Before sending field investigators for data collection, try to give them proper training on the interview schedule. If possible, some orientation on various aspects of the problem may be given. It will enable the field investigator to effectively interact with the respondents. Nowadays, the NFHS (National Family Health Survey), RCH (Reproductive and Child Health) surveys and many base line surveys conducted by various agencies spend a lot of money in training of the field investigator before sending them for data collection.

- iii) Method of Conducting an Interview:** The field investigator must be practiced in conducting interviews; otherwise, respondents sometimes may not allow them to take the interviews. He must approach the respondents politely, introduce himself/herself and tell them the purpose of interview and the confidentiality involved in it. The respondents must be approached by the field investigator according to their convenience. Getting correct information from the informants depends on the skill of the field investigator.
- iv) Editing of the Interview Schedule:** Editing of the interview schedule is a must before sending it for tabulation and analysis. While checking the schedule, one must notice the number of cases allotted, number of cases contacted, and number of cases lost due to refusal. The field supervisor must check the schedule filled up by the field investigator. If information is missing from any schedule then the field investigators could be sent again for data collection. If the schedules have codes for different alternative responses, these should also be checked, and, if any contradiction exist must be shared before sending it for final data entry and tabulation.

Sample Interview Schedule	
Title: Socio-Economic Study of Households in an Urban Slum (To be filled by the Head of the Household Or Any Adult Member of the Family)	
Name of the Respondent:	
I) Identification of slum:	
1) Name of the slum	
II) Identification of Household:	
1) Household Survey No	
(marked by the survey team)	
2) Name of the Head of the Household	
3) Religion	
4) Caste/Sub Caste, Specify	
III) Housing Condition:	
1) Type of House;	
i) Kacha	(.....)
ii) Pacca	(.....)
iii) Semi Pacca	(.....)
iv) Any other	(.....)
2) Main source of drinking water;	
i) Government/own hand pump	(.....)
ii) Tap/pipe	(.....)
iii) Canal/river	(.....)
iv) River/well/pond	(.....)
v) Any others, specify	

3) Toilet facilities;

- i) Flush (.....)
- ii) Pit (.....)
- iii) Open field (.....)
- iv) Any other (.....)

4A) Main source of light;

- i) Electricity (.....)
- ii) Kerosene/Oil (.....)
- iii) Gas (.....)
- iv) Any other (.....)

4B) Main source of cooking;

- i) Traditional *chulha* (.....)
- ii) Bio-gas/gas (.....)
- iii) Kerosene/electric stove (.....)
- iv) Any other (.....)

5) Communication Media;

- i) Radio/transistor/tape recorder (.....)
- ii) Television (.....)
- iii) Newspaper/magazine (.....)
- iv) Any other (.....)

6) Agricultural land owned by the household in their village

- i) 1-10 Bigha (.....)
- ii) 11-20 Bigha (.....)
- iii) 21-30 Bigha (.....)
- iv) 31-40 Bigha (.....)
- v) 41 Bigha and above (.....)

(Note : One hectare is equivalent to approximately 12 *Bigha*)

7) Average monthly income of household (Rs. per month);

- i) Rs.1-1000 (.....)
- ii) Rs.1001-2000 (.....)
- iii) Rs.2001-3000 (.....)
- iv) Rs.3001-4000 (.....)
- v) Rs.4001-5000 (.....)
- vi) Rs.5001 and above (.....)

(Kindly mention actual income of the household.....)

IV) Household Profile

Sr. No.	Members of the household (start from Head of the household)	Relationship	Sex	Age	Marital Status	Education	Occupation	Diseases
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								

- Note:** 1) If there are more than 15 members in a household, use another household schedule to fill the relevant data of the household profile.
- 2) If age of any member, under (Col.5) is below 5 years then Col. 6, Col. 7 and Col. 8 are not to be filled.

Col. (3) Relationship:

(i) Head of household-01; (ii) Spouse (wife/husband)-02; (iii) Son-03; (iv) Daughter-04 (v) Grandson-05; (vi) Granddaughter-06; (vii) Father-07; (viii) Mother-08; (ix) Grandfather-09; (x) Grandmother-10; (xi) Son in law-11; (xii) Daughter in law-12; (xiii) Brother-13; (xiv) Sister-14; (xv) Brother in law-15; (xvi) Sister in law-16; (xvii) Uncle-17; (xviii) Auntie-18; (xix) Nephew-19; (xx) Niece-20; (xxi) Servant-21; (xxii) Other household member specify.....-22. (It should be clear that persons sharing their meals in a single kitchen, are considered as family members)

Col. (6) Marital Status:

(i) Currently married-01; (ii) Separated-02; (iii) Widow-03; (iv) Widower-04; (v) Divorced-05; (vi) Never married-06

Col. (7) Education:

(i) Illiterate-01; (ii) Literate (non-formal)-02; (iii) Primary-03; (iv) Middle-04; (v) High School-05; (vi) Higher Secondary or Intermediate-06; (vii) Graduate-07; (viii) Post Graduate and above-08; (ix) Professional or Technical Education-09; (x) Any other, specify.....-10.

Col. (8) Occupation:

(i) Cultivator-01; (ii) Agricultural/casual labourer-02; (iii) Self-employed - 03; (iv) Private service-04; (v) Government service-05; (vi) Household/ domestic activities-06; (vii) Student-07; (viii) Unemployed-08; (ix) Any other, specify.....-10.

Col. (9) Diseases:

(i) Tuberculosis (T.B.)-01; (ii) Asthma-02; (iii) Cataract/Blindness-03; (iv) Leprosy-04; (v) Physical impairment-05; (vi) Malaria during last 3 months-06; (vii) Diabetes-07; (viii) Hypertension-08; (ix) Heart Problem-09; (x) Any other specify.....-10.

1.4.1 Advantages of Interview Schedule

- i) the interview probes the problem in detail which gives scope to gather detailed information
- ii) there is a personal touch between the investigator and the respondents and, therefore, detailed and exhaustive information can be collected
- iii) there is greater accuracy in getting the information
- iv) the interview method is particularly suitable for illiterate respondents.

1.4.2 Disadvantages of Interview

- i) it is a time consuming and expensive method, compared to the questionnaire method
- ii) lack of objectivity is a common lacuna of the interview method
- iii) the interview method sometimes leaves investigators at the mercy of respondents.

Now that you have read about interview schedule, try and answer the following questions in *Check Your Progress 2*.

Check Your Progress 2

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

- 1) What are the advantages of the interview schedule method?

.....

.....

.....

.....

.....

.....

.....

.....

2) What steps need to be taken while preparing an interview schedule?

.....

.....

.....

.....

.....

.....

1.5 SECONDARY DATA COLLECTION METHODS

Secondary data are collected by investigators from sources other than primary respondents. Secondary data are collected from both, published and unpublished sources. The main resources of secondary data are given below.

- i) Official publications of the Central Government such as Census, NSSO Report, Human Development Report, SRS report, etc.
- ii) Research and study reports of bilateral and multilateral organizations such as WHO, World Bank, IMF, UNESCO, UNICEF, etc. A few of them include, the World Development Report, World Development Indicators, Human Development Report, etc.
- iii) Reports brought out by committees and commissions such as the Mandal Commission Report, the National Planning Commission Report, the Human Rights Commission Report, the Population Commission Report, etc.
- iv) Policy documents of the Central and State Government such as National Population Policies, National Education Policies, National Health Policies, etc.
- v) Publications brought out by the research institutes, universities and organizations.
- vi) Publications of data sources on different national and international journals such as Economic and Political Weekly, Indian Economic Journal, etc.
- vii) Books and articles published on various subjects.
- viii) Official publications of the Reserve Bank of India, State Bank of India, Association of Indian Banking, etc.
- ix) Information available in year books and encyclopaedias.
- x) Statistical abstracts published both by the Central and State Governments.
- xi) Information published in the directories and bulletins of various institutions such as the Indian Council of Social Sciences Research, the Indian Council of Agricultural Research, etc.
- xii) Abstracts and index of reports and articles published by various research, teaching, and related organizations.

1.5.1 Precautions in the use of Secondary data

While using secondary data for the study, users have to be careful. Sometimes, the data published by an individual researcher may be full of errors and even drawn from an inadequate sample. Some factors to keep in mind while using the data from secondary sources are listed below.

Adequacy –sometimes data available from the secondary, sources are not adequate for the investigation. Data may either be from a different time period, or partially fulfil the requirement of the study. Therefore, adequacy of the data must be ensured before conducting the study.

Reliability – before using secondary data, its reliability must be taken into consideration. For example, the reliability on sample size and the sampling method used in the collection of data may be taken into consideration. Besides, the investigator has also to know the degree of bias in collection of data.

Suitability –the investigator has to check whether the data is suitable for the purpose of the research study. Sometimes, the secondary data may be suitable for tabular presentation, but, unsuitable for statistical analysis.

The investigator has to take all these factors into consideration before using the secondary data.

In this section you read about the secondary data methods. Now try and answer the questions in *Check Your Progress-3*.

Check Your Progress 3

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) What are few sources of secondary data?

.....

.....

.....

.....

.....

.....

2) What precautions have to be taken while using secondary data?

.....

.....

.....

.....

.....

.....

1.6 LET US SUM UP

This unit describes in detail the various sources and methods of collection of quantitative data. It also deals with the methods of collection of primary data and secondary data. The advantages and disadvantages of various methods of data collection have also been envisaged in the unit. The unit also narrates the precautions one has to take care up while collecting the primary and secondary data.

1.7 REFERENCES AND SELECTED READINGS

Abramson JH (1990), Survey Methods in Community Medicine. London: Churchill- Livingstone.

Cohen, L., Manion, L. and Morrison, K. (2000), Research Methods in Education 5th Edition. London: Routledge Falmer

Dillon, W.R., Madden, T.. and Firtle, N. H., (1994) Marketing Research in a Research Environment, 3rd edition, Irwin.

Green, P.E. Tull, D.S. and Albaum G (1993) Research methods for marketing decisions, 5th edition, Prentice Hall, p.136.

Joselyn, R. W. (1977) Designing the marketing research, Petrocellis/Charter, New York, p.15.

Moser CA, Kalton G (1989), Survey Methods in Social Investigation. Hants, UK: Gower Publishing Company.

Patton MQ (1990, 2nd ed.) Qualitative Evaluation and Research Methods. Newbury Park, USA: Sage Publications.

Pretty JN, Guyt I, Thompson J, Scones I (1995) Participatory Learning & Action. A Trainer's Guide. London: International Institute for environment and Development (IIED).

Suvedi,M., Singh,B., Vijayaraghavan,K., Padaria,R.N., and Wason,M. (2008). Evaluation Capacity Building in Rural Resource Management : A Manual, IARI, New Delhi.

1.8 CHECK YOUR PROGRESS-POSSIBLE ANSWERS

Check Your Progress 1

- 1) What do you mean by primary data?

Primary data are those data which are originally collected by the investigators, while the secondary data are used by investigator collected through some other sources. For example, information collected by an investigator from a student regarding his class, caste, family background, etc., is called primary data. On the other hand, if the same information is collected about the student from the school record and register, then it is called secondary data.

2) What are advantages of a questionnaire?

The advantages of a questionnaire are

- a) it is less expensive compared to interview schedule and can be administered over a large number of sample
- b) it is less time consuming.
- c) the interviewer is not present during the administration of questionnaire, the respondents have greater freedom and confidence in answering questions
- d) one of the advantages of the questionnaire is that once it is standardized, then there is greater uniformity in getting responses.

Check Your Progress 2

1) What are the advantages of the interview schedule?

The advantages of the interview method are as follows:

- a) the interview probes the problem in detail which gives scope to gathering detailed information
- b) there is a personal touch between the investigator and the respondents, and, therefore, detailed and exhaustive information can be collected
- c) there is greater accuracy in getting the information
- d) the interview method is particularly suitable for illiterate respondents.

2) What precautions have to be taken care of while preparing an interview schedule?

The precautions that have to be taken care of during the preparation of the interview schedule are:

- Selection of Respondents: this is the key to interview schedule administration.
- Training of the Field Staff: before sending the field investigator for data collection, try to give them proper training on the interview schedule.
- Method of Conducting an Interview: the field investigator must be apt in conducting the interview, otherwise respondents sometimes may not allow them to take interview
- Editing of the Interview Schedule: the investigator has to do proper editing of the interview schedule before sending it for tabulation and analysis.

Check Your Progress 3

1) What are a few sources of secondary data?

- 1) Official publication of the Central Government such as Census, NSSO Report, Human Development Report, SRS report, etc.
- 2) Research and Study Reports of bilateral and multilateral organizations such as WHO, World Bank, IMF, UNESCO, UNICEF, etc. A few of them like World Development Report, World Development Indicators, Human Development Report, etc.

- 3) Reports brought out by Committees and Commissions such as Mandal Commission Report, National Planning Commission Report, Human Rights Commission Report, Population Commission Report etc.
- 2) What precautions have to be taken care of while using secondary data?

Some of the factors to be kept in mind while using the data from secondary sources are:

Adequacy –sometimes data that is available from secondary sources are not adequate for the research project. Data may either be from a different time period or partially fulfil the requirement of the study.

Reliability – before using the secondary data, its reliability must be taken into consideration. The investigator has also to know the degree of bias in the collection of data.

Suitability –the investigator has to see whether the data is suitable to his or her study. Sometimes, the secondary data available may be suitable for tabular presentation, but unsuitable for statistical analysis.



UNIT 2 QUALITATIVE DATA COLLECTION METHODS AND DEVICES

Structure

- 2.1 Introduction
- 2.2 Qualitative Data - Meaning and Concept
- 2.3 Methods and Techniques of Qualitative Data Collection
- 2.4 Features of Qualitative and Quantitative Research
- 2.5 Let Us Sum Up
- 2.6 Keywords
- 2.7 References and Selected Readings
- 2.8 Check Your Progress – Possible Answers

2.1 INTRODUCTION

Data Collection is an important aspect of any type of research study. Data collection techniques allow us to systematically collect information about the subject of our study (people, objects, phenomena), and about the environment. In the collection of data we have to be systematic. If data are collected haphazardly, it will be difficult to answer our research questions in a conclusive way. Inaccurate data collection can impact the results of a study and ultimately lead to invalid results.

After studying this unit, you should be able to:

- discuss the meaning and concept of qualitative data.
- describe the features of various methods and devices used for qualitative data collection.
- state the uses and limitations of various qualitative data collection methods.

2.2 QUALITATIVE DATA - MEANING AND CONCEPT

Qualitative research is grounded in the assumption that individuals construct social reality in the form of meanings and interpretations, and that these constructions tend to be transitory and situational. Qualitative research typically involves qualitative data, i.e., data obtained through methods such as interviews, on-site observations, and focus groups that is in narrative rather than numerical form. Such data are analyzed by looking for themes and patterns. It involves reading, re-reading, and exploring the data. How the data are gathered will greatly affect the ease of analysis and utility of findings.

Qualitative data are descriptive in nature and can be statistically analyzed only after processing and after having them classified into some appropriate categories. Qualitative data can, however, facilitate in-depth analysis of a social situation. There are certain situations where qualitative research alone can provide the researcher with all insights needed to make decisions and take actions; while in some other cases quantitative research might be needed as well.

2.3 METHODS AND TECHNIQUES OF QUALITATIVE DATA COLLECTION

Qualitative methods are ways of collecting data which are concerned with describing meaning, rather than with drawing statistical inferences. They provide in-depth and rich descriptions. In this section, a detailed description and comparison of the most commonly used qualitative methods employed in social science research is given. These include observations, in-depth interviews, and focus groups.

2.3.1 Observation Method

In our daily life we observe many things and events around us, but this sort of observation is not scientific. Observational techniques are methods by which an individual or individuals gather first hand data on programs, processes, or behaviours being studied. They provide evaluators with an opportunity to collect data on a wide range of behaviours, to capture a great variety of interactions, and to openly explore the evaluation topic. By directly observing operations and activities, the evaluator can develop a holistic perspective, i.e., an understanding of the context within which the project operates. This may be especially important where it is not the event that is of interest, but rather how that event may fit into, or be impacted by, a sequence of events.

Scientific observation is a methodical way of recognizing and noting a fact or occurrence, often involving some sort of measurement. Scientific observations should be specific, and recorded immediately. Understanding the culture of the people and the ability to interact with them are essential for good observation. Researches may be based solely on observation, but in most cases observation precedes other methods of data collection.

When to use observations: Observations can be useful during both the formative and summative phases of evaluation. For example, during the formative phase, observations can be useful in determining whether or not the project is being delivered and operated as planned. In the hypothetical project, observations could be used to describe the faculty development sessions, examining the extent to which participants understand the concepts, ask the right questions, and are engaged in appropriate interactions. Such formative observations could also provide valuable insights into the teaching styles of the presenters and how they are covering the material.

Advantages

- i) Subjective bias may be eliminated, if observation is done accurately>
- ii) Information relates to current state of affairs ; and it is
- iii) Independent of respondents' willingness or capability to respond.

Limitations

- i) A time consuming and expensive method
- ii) A limited amount of information may be available; and
- iii) Extraneous factors may interfere with the task of observation.

Types of observation:

Structured and unstructured observation: in case the observation is characterized by a careful definition of the units to be observed, the manner of recording the observed information, standardized conditions of observation, and the selection of pertinent data of observation, then the observation is called as structured observation. But, when the observation is conducted without these features thought out in advance, the same is termed an unstructured observation. Structured observation is considered appropriate in descriptive studies, whereas, in an exploratory study, the observational procedure should be relatively unstructured.

Participant and non-participant observation: this depends on the degree of involvement of the researcher with the situation being observed. In participant observation, the researcher who may be an outsider, while observing the group, also plays the role of a group member. It is necessary to have the observation in an unbiased way, without getting emotionally involved in the affairs of the group or the community. The main advantage of participant observation is that, it helps the observer to get an intimate knowledge of the group or the community being observed, under natural condition. For example, if one wants to study the fishermen community, reliable information may be obtained through the method of participant observation. Participant observation, however, requires longer time, greater resources, and there may be loss of objectivity, if not properly done. In non-participant observation, the researcher observes the group or the community, while maintaining physical and psychological isolation from them. This ensures collection of information in an objective way.

Controlled and uncontrolled observation: when observation takes place according to a definite pre-arranged plan involving experimental procedure, the same is termed a controlled observation. The aim of a controlled observation is to check any bias due to faulty perception, incomplete information and effect of external stimuli on a specific situation. An uncontrolled observation on the other hand, is one where the researcher observes the behaviour and activities of a group under natural conditions (as they are) without any stimulation from the outside. This method provides a wide range of information and helps in developing an insight about the group or community. Care should, however, be taken against subjective interpretation of observed phenomenon.

Recording Observational Data

Observations are carried out using a carefully developed set of steps and instruments. The observer is more than just an onlooker, but, rather, comes to the scene with a set of target concepts, definitions, and criteria for describing events. While, in some studies, observers may simply record and describe, in the majority of evaluations, their descriptions are, or eventually will be, judged against a continuum of expectations.

Observations usually are guided by a structured protocol. The protocol can take a variety of forms, ranging from the request for a narrative, describing events seen to a checklist or a rating scale of specific behaviours/activities that address the evaluation question of interest. The use of a protocol helps assure that all observers are gathering the pertinent information and, with appropriate training, applying the same criteria in the evaluation. For example, an observational

approach is selected to gather data on the faculty training sessions, the instrument developed would explicitly guide the observer to examine the kinds of activities in which participants were interacting, the role(s) of the trainers and the participants, the types of materials provided and used, the opportunity for hands-on interaction, etc.

Field notes are frequently used to provide more in-depth background or to help the observer remember salient events if some forms are not completed at the time of observation. Field notes contain the description of what has been observed. The descriptions must be factual, accurate, and thorough without being judgmental and cluttered by trivia. The date and time of the observation should be recorded, and everything that the observer believes to be worth noting should be included. No information should be trusted to future recall.

Technological tools, such as a battery-operated tape recorder or a Dictaphone, laptop computer, camera, and video camera, can make the collection of field notes more efficient and the notes themselves more comprehensive. Informed consent must be obtained from participants before any observational data are gathered.

2.3.2 Interview/Questionnaire Method

In these methods, the data are collected by presenting stimuli to the respondents in the form of questions for eliciting appropriate responses from them. The questions may be presented to the respondents in a face-to-face situation as oral-verbal stimuli, and the researcher or personnel trained for the purpose (interviewers, enumerators) note down their oral-verbal responses. This method is known as the interview method, and the set of questions is known as the interview schedule. In another method, the questions are delivered (generally mailed) to the respondents, who note down their responses on it and send it back to the researcher. This method is known as the questionnaire method, and the set of questions is known as a questionnaire.

In both, answers to some systematically organised questions, relevant to the objectives of the study are sought. The questions should be accurate and clearly understood by the respondents, so that the responses are accurate. Both the methods have some advantages and limitations. The success of the questionnaire method depends more on the quality of the questionnaire itself, but in case of interview method much depends upon the honesty and competency of the enumerators.

Types of interview

Interviews may be of different types according to the needs of the situation.

Structured interview: For this purpose an interview schedule is used which is well structured with specific questions to be asked. The questions are precisely worded and systematically organised, and are prepared in advance after requisite pre-testing. The interviewer is not expected to make any change while interviewing the respondents. The data received are comparable and are more amenable to statistical analyses. The structured interview is also known as standardized, controlled or guided interview.

Unstructured interview: Here the interviewer proceeds with some well thought out themes or guidelines to be inquired into, and brings out the required

information from the respondents through the process of conversation. The situation is free and informal and no interview schedule is used. This provides more flexibility and freedom, but at the same time demands deep knowledge and greater skill on the part of the interviewer. The process may yield good amount of information, but the data lack comparability and are less amenable to statistical analysis. Unstructured interview is suitable for exploratory or formulative research studies.

Focused interview: In focused interviews, some specific issue, occurrence, experience, or event is taken into consideration instead of general aspects of a situation. The interviewer has the freedom to decide the manner and sequence in which the questions would be asked, and, has also the freedom to explore reasons and motives. The main task of the interviewer, however, is to confine the discussion to the specific issue under investigation. Such interviews are convenient for development of hypotheses, action research etc. and constitute a major type of unstructured interviews.

In-depth interview: An in-depth interview is a dialogue between a skilled interviewer and an interviewee. Its goal is to elicit rich, detailed material that can be used in analysis. These interviews are designed to discover motives and desires, and, are often used in motivational research. Such interviews are held to explore needs, desires, and feelings of respondents. Such interviews are best conducted face to face, although in some situations telephone interviewing can be successful.

In-depth interviews are characterized by extensive probing and open-ended questions. Typically, the researcher prepares an interview guide that includes a list of questions or issues that are to be explored and suggested probes for following up on key topics. The guide helps the interviewer pace the interview and makes interviewing more systematic and comprehensive.

The dynamics of interviewing are similar to a guided conversation. The interviewer becomes an attentive listener who shapes the process into a familiar and comfortable form of social engagement - a conversation - and the quality of the information obtained is largely dependent on the interviewer's skills and personality. In contrast to a good conversation, however, an in-depth interview is not intended to be a two-way form of communication and sharing. The key to being a good interviewer is being a good listener and questioner. Tempting as it may be, it is not the role of the interviewer to put forth his or her opinions, perceptions, or feelings. Interviewers should be trained individuals who are sensitive, empathetic, and able to establish a non- threatening environment in which participants feel comfortable. They should be selected during a process that weighs personal characteristics that will make them acceptable to the individuals being interviewed; clearly, age, sex, profession, race/ethnicity, and appearance may be key characteristics. Thorough training, including familiarization with the research problem and its goals, is important.

Specific circumstances in which in-depth interviews are particularly appropriate include

- complex subject matter
- detailed information sought
- busy, high-status respondents
- highly sensitive subject matter.

2.3.3 Case Study Method

The case study method is a very popular form of qualitative analysis and involves a careful and complete observation of a social unit, be that unit a person, a family, an institution, a cultural group, or, even the entire community. It is a method of study in depth rather than breadth. The case study places more emphasis on the full analysis of a limited number of events or conditions and their interrelations. The case study deals with the processes that take place and their interrelationship. Thus, a case study is essentially an intensive investigation of the particular unit under consideration. The object of the case study method is to locate the factors that account for the behaviour patterns of the given unit as an integrated totality.

Pauline V. Young describes case study as “a comprehensive study of a social unit be that unit a person, a group, a social institution, a district or a community.” In brief, we can say that the case study method is a form of qualitative analysis where careful and complete observation of an individual, situation, or an institution is done; efforts are made to study each and every aspect of the concerned unit in minute detail, and then, from case data generalizations and inferences are drawn.

Characteristics: the important characteristics of the case study method are listed below.

- i) In this method, the researcher can take a single social unit or more such units for his study purpose.
- ii) Here the selected unit is studied intensively, i.e., it is studied in minute detail. Generally, the study extends over a long period of time to ascertain the natural history of the unit so as to obtain enough information for drawing correct inferences.
- iii) In the context of this method we make a complete study of the social unit covering all facets. Through this method we try to understand the complex of factors that are operative within a social unit as an integrated totality.
- iv) Using this method, the approach happens to be qualitative and not quantitative. Mere quantitative information is not collected. Every possible effort is made to collect information concerning all aspects of life. As such, the case study method deepens our perception and gives us a clear insight into life. For instance, in the case study method, we not only study how many crimes a man has committed, but we peep into the factors that forced him to commit crimes when we are making a case study of a man who is a criminal. The objective of the study may be to suggest ways to reform the criminal.
- v) In respect of the case study method, an effort is made to know the mutual inter-relationship of causal factors.
- vi) We study behaviour pattern of the concerned unit directly, and not by an indirect and abstract approach.
- vii) The case study method results in fruitful hypotheses, along with the data which may be helpful in testing them, and, thus, this method enables generalized knowledge to get richer and richer. In its absence, generalized social science may get handicapped.

Assumptions: the case study method is based on several assumptions. The important assumptions may be listed as follows.

- i) The assumption of uniformity in basic human nature, in spite of the fact that human behaviour may vary according to situations.
- ii) The assumption of studying the natural history of the unit concerned.
- iii) The assumption of comprehensive study of the unit concerned.

Major phases involved

- i) Recognition and determination of the status of the phenomenon to be investigated or the unit of attention.
- ii) Collection of data, examination, and history of the given phenomenon.
- iii) Diagnosis and identification of causal factors as a basis for remedial developmental treatment.
- iv) Application of remedial measures, i.e., treatment and therapy (this phase is often characterized as case work).
- v) Follow-up programme to determine effectiveness of the treatment applied.

Check Your Progress 1

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit.

- 1) What is the difference between the interview method and the questionnaire method?

.....

.....

.....

.....

.....

- 2) List out the important assumptions of the case study method.

.....

.....

.....

.....

.....

2.3.4 Focus Groups

Focus groups combine elements of both interviewing and participant observation. The focus group session is an interview, not a discussion group, a problem solving session, or a decision making group. At the same time, focus groups capitalize on group dynamics. The hallmark of focus groups is the explicit use of group interaction to generate data and insights that would be unlikely to emerge without

the interaction found in a group. The technique inherently allows observation of group dynamics, discussion, and firsthand insights into the respondents' behaviours, attitudes, language, etc.

Focus groups are a gathering of 8 to 12 people who share some characteristics relevant to the evaluation. Originally used as a market research tool to investigate the appeal of various products, the focus group technique has been adopted by other fields, such as education, as a tool for data gathering on a given topic. Focus groups, conducted by experts, generally take place in a focus group facility that includes recording apparatus (audio and/or visual) and an attached room with a one-way mirror for observation. There is an official recorder who may or may not be in the room. Participants are paid for attendance and provided with refreshments. As the focus group technique has been adopted by fields outside of marketing, some of these features, such as payment or refreshment, have been eliminated.

When to use focus groups: When conducting evaluations, focus groups are useful in answering the same type of questions as in-depth interviews. Specific applications of the focus group method in evaluations include

- identifying and defining problems in project implementation
- identifying project strengths, weaknesses, and recommendations
- assisting with interpretation of quantitative findings
- obtaining perceptions of project outcomes and impacts
- generating new ideas.

Although focus groups and in-depth interviews share many characteristics, they should not be used interchangeably.

Developing a Focus Group

An important aspect of conducting focus groups is the topic guide. The topic guide, a list of topics or question areas, serves as a summary statement of the issues and objectives to be covered by the focus group. The topic guide also serves as a road map and as a memory aid for the focus group leader, called a moderator. The topic guide also provides the initial outline for the report of findings.

Focus group participants are typically asked to reflect on the questions asked by the moderator. Participants are permitted to hear each other's responses and to make additional comments beyond their own original responses as they hear what other people have to say. It is not necessary for the group to reach any kind of consensus, nor is it necessary for people to disagree. The moderator must keep the discussion flowing and make sure that one or two persons do not dominate the discussion. As a rule, the focus group session should not last longer than 1 ½ hours to 2 hours. When very specific information is required, the session may be as short as 40 minutes. The objective is to get high quality data in a social context where people can consider their own views in the context of the views of others, and, where new ideas and perspectives can be introduced.

2.3.5 Content Analysis

Content analysis consists of analyzing the contents of documents, such as books, magazines, newspapers, and the contents of all other verbal materials, either

spoken or printed. Content analysis prior to 1940's was mostly quantitative documentary materials concerning certain characteristics that can be identified and counted. But since the 1950's, content analysis is mostly qualitative analysis, concerning the general importance of the existing documents.

The analysis of content is a central activity whenever one is concerned with the nature of the verbal materials. A review of research in any area, for instance, involves of the contents of research articles that have been published. The analysis may be at a simple level, or, it may be a subtle one. It is at a simple level when we pursue it on the basis of certain characteristics of the document, or, verbal materials that can be identified and counted (such as on the basis of major scientific concepts in a book). It is at a subtle level when researcher uncovers the attitude, say of the press towards education by feature writers.

2.3.6 Other Qualitative Data Collection Methods

The last section outlines less common but, nonetheless, potentially useful qualitative methods for project evaluation. These methods include document studies, key informants and alternative (authentic) assessment.

- i) **Document studies:** Existing records often provide insights into a setting and/or group of people that cannot be observed or noted in another way. This information can be found in document form. A document can be defined as "any written or recorded material" not prepared for the purposes of the evaluation, or, at the request of the inquirer. Documents can be divided into two major categories: public records, and personal documents.
- ii) **Public records:** are materials created and kept for the purpose of "attesting to an event or providing an accounting". Public records can be collected from outside (external) or within (internal) the setting in which the evaluation is taking place. Examples of external records are census and vital statistics reports, county office records, newspaper archives, and local business records that can assist an evaluator in gathering information about the larger community and relevant trends. Such materials can be helpful in better understanding the project participants and making comparisons between groups/communities.

For the evaluation of educational innovations, internal records include documents such as student transcripts and records, historical accounts, institutional mission statements, annual reports, budgets, grade and standardized test reports, minutes of meetings, internal memoranda, policy manuals, institutional histories, college/university catalogues, faculty and student handbooks, official correspondence, demographic material, mass media reports and presentations, and descriptions of program development and evaluation. They are particularly useful in describing institutional characteristics, such as the backgrounds and academic performance of students, and in identifying institutional strengths and weaknesses. They can help the evaluator understand the institution's resources, values, processes, priorities, and concerns. Furthermore, they provide a record or history that is not subject to recall bias.

- iii) **Personal documents:** are first-person accounts of events and experiences. These "documents of life" include diaries, portfolios, photographs, artwork,

schedules, scrapbooks, poetry, letters to the paper, etc. Personal documents can help the evaluator understand how the participant sees the world and what she or he wants to communicate to an audience. And, unlike other sources of qualitative data, collecting data from documents is relatively invisible to, and requires minimal cooperation from, persons within the setting being studied.

The usefulness of existing sources varies depending on whether they are accessible and accurate. In the hypothetical project, documents can provide the evaluator with useful information about the culture of the institution and participants involved in the project, which in turn can assist in the development of evaluation questions. Information from documents also can be used to generate interview questions or to identify events to be observed. Furthermore, existing records can be useful for making comparisons (e.g., comparing project participants to project applicants, project proposal to implementation records, or documentation of institutional policies and program descriptions prior to, and, following the implementation of project interventions and activities).

- iv) **Key informant:** A key informant is a person (or, group of persons) who has unique skills or professional background related to the issue/intervention being evaluated, is knowledgeable about the project participants, or has access to other information of interest to the evaluator. A key informant can also be someone who has a way of communicating that represents, or, captures the essence of what the participants say and do. Key informants can help the evaluation team better understand the issue being evaluated, as well as the project participants, their backgrounds, behaviours, and attitudes, and, any language or ethnic considerations. They can offer expertise beyond the evaluation team. They are also very useful for assisting with the evaluation of curricula and other educational materials. Key informants can be surveyed or interviewed individually or through focus groups.

In the hypothetical project, key informants (i.e., expert faculty on main campus, deans, and department chairs) can assist with: (1) developing evaluation questions, and; (2) answering formative and summative evaluation questions.

- v) **Performance assessment:** the performance assessment movement is impacting education from pre-schools to professional schools. At the heart of this upheaval is the belief that for all of their virtues - particularly efficiency and economy - traditional objective, norm-referenced tests may fail to tell us what we most want to know about student achievement. In addition, these same tests exert a powerful and, in the eyes of many educators, detrimental influence on curriculum and instruction. The search for alternatives to traditional tests has generated a number of new approaches to assessment under such names as alternative assessment, performance assessment, holistic assessment, and authentic assessment. While each label suggests slightly different emphases, they all imply a movement toward assessment that supports exemplary teaching. Performance assessment appears to be the most popular term because it emphasizes the development of assessment tools that involve students in tasks that are worthwhile, significant, and meaningful. Such tasks involve higher order thinking skills and the coordination of a broad range of knowledge.

Performance assessment may involve qualitative activities such as oral interviews, group problem-solving tasks, portfolios, or personal documents/creations (poetry, artwork, stories). The quality of this product is assessed (at least, before and after training) in light of the goal of the professional development program. The actual performance of students on the assessment measures provides additional information on impact.

2.4 FEATURES OF QUALITATIVE AND QUANTITATIVE RESEARCH

In unit 1, we discussed various aspects of quantitative data collection methods. Let us now see the main differences between qualitative and quantitative methods.

Qualitative research	Quantitative research
Mainly for exploratory purposes and to generate hypotheses	Used to obtain descriptive data
Usual purpose is to generate a range and variety of data	Usual purpose is to consolidate the data and obtain a clear picture of the situation
The methods of inquiry are informal and flexible	All methods are , carefully planned tightly controlled
The researcher usually starts with only a broad indication of the information objectives of the project, but with clear understanding of the overall purpose of the research	The research is confined to a list of research objectives which set out what information is required
The researcher usually works from a list of the topics to be covered, but the course of each 'interview' will be influenced by the respondent	The interviewer uses a questionnaire, which must be followed exactly as instructed in every interview
Based on small numbers of respondents who take part individually or in small groups	Based on larger numbers of respondents; data are collected from each person individually
It cannot be known how true the findings are of the population from which the respondents are drawn	May be possible to estimate how reliable the findings are. It depends on which sampling method is used
Data collection is usually handled by research professionals	Usually done by trained interviewers or through self-completion questionnaires
A qualitative project cannot be repeated exactly, because every data collection event in a project is different	Can usually be replicated, because every interview in the project follows the same procedure
The findings can rarely be expressed in statistical form	Findings are expressed in number and can be analysed using statistical techniques
Analysis and conclusion rely heavily on the researcher's perceptions and interpretation skills	Because statistical procedures are used the analysis is less likely to be disputed

Source: John Boyce, Marketing Research, MacGraw Hill, Australia Pvt Ltd, 2005.

Check Your Progress2

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) What do you mean by focus group?

.....

.....

.....

.....

.....

.....

.....

.....

.....

2.5 LET US SUM UP

In this unit, we discussed the meaning and concept of qualitative data collection and found that the selection of appropriate method for data collection and research design depends on nature, scope and objective of enquiry. Each method of data collection, however, does not suit all categories of research design. The selection and preparation of tools for collecting data depends upon the types of data to be collected.

The researcher must decide in advance of collection and analysis of data as to which design would prove to be more appropriate for his research project. He/she must give due weight to various points, such as the type of universe and its nature, the objective of the study, the resource list or the sampling frame, the desired standard of accuracy, and the like, when taking a decision in respect of the design for the research project.

Qualitative data are descriptive in nature and can be statistically analyzed only after processing, and, after having them classified into some appropriate categories. Qualitative data can facilitate in-depth analysis of a social situation.

2.6 KEYWORDS

Key Informant : A key informant is a person (or group of persons) who has unique skills or professional background related to the issue/intervention being evaluated, is knowledgeable about the project participants, or has access to other information of interest to the evaluator.

- Content analysis** : Content analysis consists of analyzing the contents of document, such as books, magazines, newspapers and the contents of all other verbal materials v either spoken or printed.
- Sociometry** : Sociometry is a technique for describing the social relationships among individuals in a group.
- Case study** : The case study method is a very popular form of qualitative analysis and involves a careful and complete observation of a social unit, be that unit a person, a family, an institution, a cultural group, or even the entire community. It is a method of study in depth rather than breadth.
- Pre-testing** : Pre-testing means testing the interview schedule/questionnaire in advance to find out whether it is capable of eliciting appropriate responses from respondents.

2.7 REFERENCES AND SELECTED READINGS

Festinger L. and Katn D. 1953. Research Methods in Behavioural Sciences. Holt, Rinehart and Winston Inc., New York.

Goode W.J. and Hatt P.K. 1981. Methods in Social Research. McGraw- Hill Book Company, Singapore.

Kothari C.R. 1996. Research Methodology: Methods and Techniques. Wishwa prakashan, New Delhi.

Mulay Sumati and Sabarathanam V.E. 1980. Research Methods in Extension Education. Manasayan, New Delhi.

Young P.V. 1996. Scientific Social Surveys and Research. Prentice –Hall of India Pvt. Ltd., New Delhi.

2.8 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress 1

- 1) What is the difference between the interview method and the questionnaire method?

In the interview method, the questions are presented to the respondents in a face-to-face situation as oral-verbal stimuli, and the researcher, or personnel trained for the purpose (interviewers, enumerators) note down oral-verbal responses. In the questionnaire method, the questions are delivered (generally mailed) to the respondents, who note down their responses on it and send them back to the researcher.

Data Collection and Analysis

- 2) List the important assumptions of the case study method.
 - i) The assumption of uniformity in basic human nature, in spite of the fact that human behaviour may vary according to situations.
 - ii) The assumption of studying the natural history of the unit concerned.
 - iii) The assumption of comprehensive study of the unit concerned.

Check Your Progress 2

- 1) Focus group combine elements of both interviewing and participant observation capitalizing on group dynamics.



UNIT 3 STATISTICAL TOOLS

Structure

- 3.1 Introduction
- 3.2 Data: Meaning and Types
- 3.3 Variables and Tests
- 3.4 Measures of Central Tendency
- 3.5 Measures of Dispersion
- 3.6 Correlation and Regression
- 3.7 Hypothesis Testing and Inferential Statistics
- 3.8 Statistical Tests
- 3.9 Let Us Sum Up
- 3.10 Keywords
- 3.11 References and Selected Readings
- 3.12 Check Your Progress – Possible Answers

3.1 INTRODUCTION

Why a learner of urban planning and development needs to know about statistical tests is simply because statistical tests will help him/her in analyzing data and drawing up inferences about the data. Those who are middle as well as at the decision making level need some understanding of statistical analysis to understand the strengths and weaknesses of published data to take decisions on whether to apply it in decision making. With the availability of several user friendly software, use of statistical tests has now become a reality, even for non-statisticians, provided they are computer literate and understand the basic principles of statistical analysis. This unit will help you to acquire knowledge about some basic statistical tools which you can use in data analysis.

After reading this unit you will be able to:

- define data, types of data and variables
- explain measures of central tendency
- calculate measures of dispersion, correlation and regression
- describe various inferential statistical tools.

3.2 DATA: MEANING AND TYPES

You know that some basic statistical tools need to be applied for the analysis of data while writing a report of urban development studies. Before describing the meaning of data, let us, know what we mean by statistics. According to Netter and Wasserman “statistics refers to the body of technique or methodology which has been developed for the collection, presentation and analysis of quantitative data and for the use of such data in decision-making”. Statistical tools and techniques are used by the researchers to analyse and interpret data. Thus ‘data’ is a fundamental requirement for any decision making. Data is generally defined

as the evidence of fact which describes a group or a situation and from which conclusion is drawn. The data is the plural from the word 'datum', which means fact? Data is usually classified into two types:

- i) Primary data and Secondary data
 - ii) Discrete data and continuous data
- i) **Primary data and Secondary data:** Primary data is the first hand information gathered by an investigator or observer regarding a situation. Researcher collects primary data keeping problems in mind. According to P.V. Young, there are two types of sources of primary data i.e. direct primary sources and indirect primary sources. In direct primary sources, researchers have direct interaction of first hand field work observation through interview schedule and questionnaire. While in indirect primary sources, he uses the medium of radio broadcasting, television appeal and other valuable documents for gathering information. Some of the advantages of primary data are: (i) Flexibility in collecting data; (ii) more appropriate for large area.

The secondary data are gathered from personal or public documents. The various sources of secondary data are books, journals, reports, letters and diaries etc.

- ii) **Discrete Data and Continuous Data:** Discrete data can take only a discrete value, that can be divided into categories or group such as male and female, white and black; boys and girls, etc.

On the other hand, the continuous data can take any value including decimal. This is a type of data usually associated with some sort of physical measurement. The height of trees in a nursery is an example of continuous data.

3.3 VARIABLES AND TESTS

While dealing with the statistical tools and data you have to acquire knowledge about two important concepts i.e. variables and tests. Let us discuss them one by one.

- i) **Variables:** Variables represents persons or objects which can be manipulated, controlled or measured for the sake of research. There are two types of variable in research such as independent and dependent variables.

The independent variable is the variable that is varied or manipulated by the researcher. On the other hand, dependent variable is the response that is measured. In other words, an independent variable is the presumed cause; where as the dependent variable is the presumed effect. For example diseases among children are the independent variable, while infant mortality is the dependent variable.

- ii) **Statistical Test:** Generally these are two types of tests applicable for statistical interpretation of data for testing hypothesis and drawing inferences, i.e. parametric test and non-parametric test. A parametric test is a test whose model specifies certain conditions about the parameters of the parent

population from which the sample was drawn. On the other hand, non-parametric test is a test whose model does not specify conditions about the parameters of the parent population from which sample was drawn.

In this session you read about data, variables and statistical tests, now answer the questions given in Check Your Progress-1

Check Your Progress 1

Note: a) Write your answer in about 50 words.

b) check your progress with possible answers given at the end of the unit.

1) What are the important types of data?

.....

.....

.....

.....

.....

2) What do you understand by non-parametric test?

.....

.....

.....

.....

.....

3.4 MEASURES OF CENTRAL TENDENCY

Measures of central tendency help the researcher to provide quantitative description of objects and events. Here numbers are assigned as per the rules and after the assignment of numbers, the individual score are compared with the average score to know the position of the individual in the group. Here average is called as the “Central Value”. The score which represents the average performance of a group is known as central tendency. The main benefits to study the measures of central tendency are: (i) to get a single value that describe the characteristics of the entire group; (ii) to get a clear idea about the entire data; and (iii) lastly, it facilitates comparison.

There are three common measures of central tendency:

- i) Mean or Arithmetic mean
- ii) Median and
- iii) Mode

3.4.1 Mean

Generally mean of a distribution is called as arithmetic mean. It is the average value of the group. Mean is the sum of the scores divided by the number of

scores. It is defined as the point on the scale of measurement obtained by dividing sum of all scores by the number of scores.

Mean is calculated from two types of data. (i) Ungrouped Data and (ii) Grouped Data

- i) **Calculation of Mean from Ungrouped Data:** the formula for calculation of mean for ungrouped data is:

$$\bar{X} = \frac{\Sigma X}{N}$$

$$\bar{X} = \text{Mean}$$

X = Individual score

N = Total number of scores

Ó = Indicates “sum of”

Example: The following marks 70, 30, 20, 90, 40 are secured by the 5 candidates in a term end examination conducted by a Municipality School. Calculate Mean.

Calculation of Mean

Candidates	Marks
A	70
B	30
C	20
D	90
E	40
N=5	Óx=250

$$\text{Mean} = \frac{\Sigma x}{N}$$

=

$$\text{Mean} = \bar{X} = 50$$

- ii) **Calculation of Mean from Grouped Data:** The mean from grouped data is calculated by applying following formula:

$$\text{Mean} = \bar{X} = \frac{\Sigma fx}{N}$$

Ó = stands for “sum of”

f = Stands for frequency

X = Stand for the mid point of class intervals

N = Total number of cases

Calculate mean value of the following group data:

Calculation Mean Value

Class Interval	Frequency
30-34	2
25-29	3
20-24	6
15-19	4
10-14	5
	N=20

At first you have to calculate the mid point of the class interval. The method of calculating mid point is

$$\text{Mid Point} = LL + \frac{UL - LL}{2}$$

LL = Lower Limit

UL = Upper Limit

$$\text{The Mid Point} = 30 + \frac{34 - 30}{2} = 32$$

For the first class interval 30-34

Class Interval	Frequency	X	fx
30-34	2	32	64
25-29	3	27	81
20-24	6	22	132
15-19	4	17	68
10-14	5	12	60
	N=20		Σfx= 405

$$\bar{X} = \frac{\sum fx}{N}$$

Now by using the formula you can calculate the mean of data given you.

Mean =

$$= \frac{405}{20} = 20.25$$

Let us know some of the important properties of mean.

Following are some of the important properties of mean:

- The mean is used when a reliable and accurate measure of central tendency is needed.
- The mean is used when scores are distributed symmetrically around the central point.

Merits

- i) It is easy to compute
- ii) It is the best representative of the group.
- iii) It is reliable.

Demerits

- i) The value of mean depends on value of each item in the series.
- ii) When scores are widely discrepant this measurement cannot be used.
- iii) When scores are skewed mean can not be used.

3.4.2 Median

The median is a value that divides a distribution into two equal halves. The median is useful when the data is in ordinal scale, i.e., some measurements are much bigger or much smaller than the other measurement value. The mean of such data will be biased toward these extreme values. Thus, the mean is not a good measure of distribution, in this case. The median is not influenced by extreme values. The median value, also called the central or halfway value, (50th percentile, i.e., 50% value below median value, and 50% above it) is obtained in the following way:

- List the observations in order of magnitude (from the lowest to the highest value, or vice versa).
- Count the number of observations = n .
- The median value is the middle value, if n is odd {i.e., $(n+1)/2$ } and the mean of two middle values, if n is even {i.e., $(n/2)$ and the next value}

i) Calculation of Median from ungrouped data

Below we have given a few examples of how to calculate Median.

Example :

Case 1: The weights of 7 women are given in Table below, then calculate median value.

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	72

The median value is the value belonging to observation number $(7 + 1)/2$, which is the fourth one value: 43 kg.

Case: If there are 8 observations as given in Table below then what will be median:

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	49
8	72

The median would be 43.5 kg {the average of '(n/2=8/2) 4th value i.e. 43' and 'next value, i.e., 44'}; the median in this case would be (43+44)/2 = 43.5 kg}.

ii) Calculation of Median from Grouped Data

Let us calculate median for a grouped data given in below table.

Number of patients	Number of clinics	Cumulative frequency
0 - 19	5	5
20 - 39	8	13
40 - 59	10	23
60 - 79	11	34
80 - 99	19	53
100 - 119	10	63
120 - 139	9	72
140 - 159	8	80
Total	80	

$$L + \left[\left(\frac{N}{2} - F \right) \times \frac{d}{f} \right]$$

The steps for calculation of median from grouped are as follows:

Step1: The total of frequency is first divided by 2, i.e., 80/2 (=40). The cumulative frequency 40 will correspond to the class interval (80-99). This is called the median interval.

Step2: The formula is Median =

Step3: Record all values of symbol variables from the table as given below:

L (=80) is the lower limit of the median interval,

F (=34) is the cumulative frequency of the class, preceding to median class,

d (=20) is the width of class interval,

f (=19) is the frequency of median class.

Step4: Replace the symbol values with numeric values as noted in step3 in the formula,

Therefore, Median = $80 + [(40-34) \times 20] / 19 = 80 + 6.32 = 86.32$ patients.

Merits and Demerits

Merits

- i) It is rigidly defined
- ii) It is easily understood and easy to calculate. In some cases it can be located merely by inspection.
- iii) It is not at all affected by extreme values.
- iv) It can be calculated for distribution with open end classes.

Demerits

- i) In case of even member of observation, median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms.
- ii) It is not based on all the observation for example the median of 10,25,50,60 and 65 is 50. We can replace the observations 10 and 25 by any two values which are smaller than 50 and the observation 60 and 65 by any two values greater than 50, without affecting the value of median. This property is sometimes described by saying that median is insensitive.
- iii) It is not amenable to algebraic treatment.
- iv) As compared with mean, it is affected much by fluctuations of sampling.

Uses

- i) Median is the only average to be used while dealing with qualitative data which cannot be measured quantitatively but still can be arranged in ascending or descending order of magnitude, e.g. to find the average intelligence and average honesty among a group of people.
- ii) It is to be used for determining the typical value in problems concerning wages, distribution of wealth, etc.

3.4.3 MODE

Let us consider the following statements.

- i) The average height of an Indian (male) is 5'6".
- ii) The average size of the shoes sold in a shop 7.
- iii) An average student in a hostel spends Rs. 150 p.m.

In all above cases, the average referred to its mode.

Mode is the value which occurs most frequently in a set of observations and around which the other items of the set, cluster densely. In other word, mode is the value of the variable which is predominant in the series. According to AM Tuttle "mode is the value which has the greatest frequency density in this immediate neighborhood". Thus in case of dissent frequency distribution mode is the value of X corresponding to maximum frequency. Let us calculate mode from the data given below.

X:	12	3	4	5	6	7	8
F:	49	16	25	22	15	7	3

The value corresponding to the maximum frequency, viz 25 is 4. Hence mode is 4

Let us calculate mode of a grouped data given in the table below:

Class Interval	Frequency
30-34	2
25-29	3
20-24	6
15-19	4
10-14	5
	N=20

Following step will be used in the calculation of mode:

Step-1 Formula for calculating mode

$$\text{Mode} = L +$$

L = (20) Lower limit of the modal class

F_1 = (6) frequency of the modal class

F_0 = (3) frequency of the class proceeding modal class

$\frac{(f_1 - f_0)}{2(f_1 - f_0 - f_2)} \times h$ F_2 = (4) frequency of the class succeeding modal class

h = (4) magnitude of class interval

Step-2 Replace the symbol values with the numeric values as noted in the step-1 in the formula

The calculated value of mode is:

$$\text{Mode} = 20 + \left[\frac{4 - 3}{2 \times 20 - 4 - 3} \right] \times 4$$

$$= 20 + 1.08 = 21.08$$

If the distribution is moderately asymmetrical, the mean, median and mode obey the following empirical relationship:

$$\text{Mode} = 3 \text{ Median} - 2 \text{ Mean}$$

Merits and Demerits of Mode

Merits:-

- Mode is relatively comprehensible and easy to calculate.
- Mode is not at all affected by extreme values.
- Mode can be conveniently located even if the frequency distribution has class intervals of unequal magnitude classes proceeding and succeeding it are of the same magnitude.

Demerits:

- i) Mode is ill defined. It is not always possible to find an early defined mode. In some cases, we may come across distribution with two modes, such distribution are called bimodal. If a distribution has more than two modes it is said to be multimodal.
- ii) It is not based upon all the observation.
- iii) It is not capable of further mathematical treatment.
- iv) As compared with mean, mode is affected to a greater extent by fluctuation of sampling.

3.5 MEASURES OF DISPERSION

The mean, median, and mode are measures of the central tendency of a variable, but they do not provide any information of how much the measurements vary or are spread. This module will describe some common measures of variation (or variability), which in statistical text books are often referred to as measures of dispersion. Measures of dispersion or variability of a data give an idea up to which extent the values are clustered or spread out. In other words, it gives an idea of the homogeneity and heterogeneity of data. Two sets of data can have similar measures of central tendency but different measures of dispersion. Therefore, measures of central tendency should be reported along with measures of dispersion. There are various measures of dispersion. Those are discussed below:

3.5.1 Range

It is the simplest measure of dispersion. This can be represented as the difference between maximum and minimum values, or simply, as the maximum and minimum values for all observations.

Example : If the weights of 7 women are as given in Table below, then what is the range?

S.No.	Weight of women (kg)
1	40
2	41
3	42
4	43
5	44
6	47
7	72

The range would be $72 - 40 = 32$ kg.

Although simple to calculate, the range does not tell us anything about the distribution of the values between the two extreme ones.

3.5.2 Percentiles

A second way of describing the variation or dispersion of a set of measurements is to divide the distribution into percentiles (100 parts). As a matter of fact, the concept of percentiles is just an extension of the concept of the median, which may also be called the 50th percentile. Percentiles are points that divide all the measurements into 100 equal parts. The 30th percentile (P30) is the value below which 30% of the measurements lie. The 50th percentile (P50), or the median, is the value below which 50% of the measurements lie. To determine percentiles, the observations should be first listed from the lowest to the highest just like when finding the median. However, in case of grouped data, percentile can be calculated on similar lines of calculating the median.

3.5.3 Mean Deviation

It is the average of deviation from arithmetic mean, where $|x|$ denotes

Mod, considering all differences 'as positive' or 'in absolute value'.

a) Calculation of Mean Deviation (A.D.)

i) **Ungrouped Data-** The formula used for calculation of mean deviation is:

$$\text{Average Deviation} = \frac{\text{Sum of all deviations}}{N}$$

$$\text{A.D.} = \frac{\sum |X_i - \bar{X}|}{n}$$

X = deviation of the raw score Mean

$|x|$ = absolute deviation (disregarding the positive and negative sign)

N = Number of scores

Σ = sum total

Example: Calculate Mean Deviation (A.D.) from the following scores

10, 20, 30, 40, 50

Table: Calculation of Mean Deviation from ungrouped data

Score	Deviation (Raw Score –Mean)	x	$ x $
10	10-30	-20	20
20	20-30	-10	10
30	30-30	0	0
40	40-30	10	10
50	50-30	20	20
$\Sigma x=150$			$\Sigma x =60$

$$\text{Mean} = \frac{\sum x}{N} = \frac{150}{5} = 30$$

$$\text{A.D.} = \frac{\sum |x - \text{Mean}|}{N} = \frac{60}{5} = 12$$

Thus A.D. = 12

3.5.4 Standard Deviation (S.D.)

Standard deviation is the only measure of dispersion which has algebraic treatment. It is the most stable measure of variability. The concept of S.D. was first suggested by Karl Pearson in 1893. Here all the deviations of the scores from mean are taken into account. In short it is considered as 'Root-Mean-Square-Deviation from Mean'. When the deviation are squared positive and negative signs become positive. When we take positive square root of the deviations, it is known as S.D. It is usually known as σ (sigma).

The formula used to calculate standard deviation is

$$\begin{aligned} \text{S.D.} &= \sqrt{\frac{\sum (X - M)^2}{N}} \\ &= \sqrt{\frac{\sum d^2}{N}} \end{aligned}$$

$\sigma =$

Where,

$\sum$ = sum total

d = deviation (score-mean)

N = total number of cases

a) Calculation of S.D. from Ungrouped Data

The formula to calculate S.D. from ungrouped data is

$\sigma =$

Example. Find out the S.D. of the following scores:

8, 9, 10, 11, 12, 13, 14, 15

Procedure for the calculation of S.D are as follows.

- Calculate Mean
- Calculate deviation against each score
- square the deviations
- find the total or sum of squared deviations
- divide sum of squared deviation by N
- Find the square root of the division.

Table: Calculation of S.D from Ungrouped Data

Score (X)	Deviation d(X-M)	d	d ²
8	8-12	-4	16
9	9-12	-3	9
10	10-12	-2	4
11	11-12	-1	1
12	12-12	0	0
13	13-12	1	1
14	14-12	2	4
19	19-12	7	49
Σ x=96			Σ d ² =84

$$\text{Mean} = \frac{\Sigma x}{N} = \frac{96}{12} = 8$$

$$\text{S.D.} = \sqrt{\frac{\Sigma d^2}{N}} = \sqrt{\frac{84}{12}} = \sqrt{7} = 2.64$$

b) Calculation of S.D. from Grouped Data

$$\frac{\Sigma x}{N} = \frac{96}{12}$$

In grouped data, deviations are taken from the mid points of the class intervals. The deviations are squared and multiplied by frequency of the said class interval. Then the root, mean of square deviations is to be calculated.

The formula to calculate S.D. is

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N}}$$

Where,

Σ = sum total

f = frequency

d² = Square of deviation

N = total number of frequencies

Table: Calculation of Standard Deviation from Grouped Data

Class Interval (C.I)	Frequency (f)
10-14	2
15-19	3
20-24	4
25-29	5
30-34	6
	N=20

Computation of S.D. is given below.

C.I.	f	X	fx	X-M	d	d ²	fd ²
10-14	2	12	24	12-20.5	-8.5	72.25	144.50
15-19	8	17	136	17-20.5	-3.5	12.25	98.00
20-24	6	22	132	22-20.5	1.5	2.25	13.50
25-29	2	27	54	27-20.5	6.5	42.25	84.50
30-34	2	32	64	32-20.5	11.5	132.25	264.50
	N=20		Óf x=410				Ófd ² =605.00

Mean =

$$= \frac{410}{20} = 20.5$$

S.D. =

$$= \sqrt{\frac{605.00}{20}} = \sqrt{30.25}$$

$$= 5.5$$

The standard deviation is 5.5

3.5.5 Coefficient of Variation

100 times the coefficient of dispersion based upon standard deviation is called coefficient variation (c.v), i.e.,

$$C.V.= 100 \times \frac{\sigma}{\bar{X}}$$

According to profession Karl Pearson who suggested this measure, C.V.is the percentage variation in the mean, standard deviation being considered as then total variation in the mean.

Fun comparing the variability of two series, we calculate the co-efficient of variations for each series. The series having greater c.v. is said to be more variable than the other and the series having lesser c.v. is said to be more consistent than the other.

In this session you read about measures of central tendency and measures of dispersion, now answer the questions given in Check Your Progress-2

Note: a) Write your answer in about 50 words.

b) check your progress with possible answers given at the end of the unit.

1) What are the different measures of central tendency?

.....

.....

.....

.....

.....

2) What are the different measures of dispersion

.....

.....

.....

.....

.....

3.6 CORRELATION AND REGRESSION

3.6.1 Correlation: Concept and Meaning

Correlation is relationship between the two sets of continuous data; for example the relationship between height and body weight. Correlation statistics are used to determine the extent to which two independent variables are related and can be expressed by a measure called 'coefficient of correlation'. The correlation coefficient may be positive or negative and therefore it may vary from '-1' to '+1'. Positive correlation means that values of two different variables increase and decrease together. For example, height and weight correlate positively. Negative correlation means that if the value of one variable decreases then the value of the other variable increases (inverse relationship). For example, literacy and number of children in family may correlate negatively.

The strength of a correlation is determined by the absolute value of the correlation coefficient; the closer the value to 1, the stronger the correlation. For example, a correlation of -0.9 indicates an inverse relationship between two variables and shows a stronger relationship than that associated with a correlation of +0.2 or -0.5. Correlation between two variables is shown by scatter plot (Figure 1) below.

Correlation analysis is important because it can be used to predict values of one variable on the basis of value of other variable. A correlation does not mean causation but it also does not mean absence of causation, that is, if two variables exhibit strong correlation, then, one of the variables may cause the other. Correlation data is, therefore, not sufficient evidence for causation.

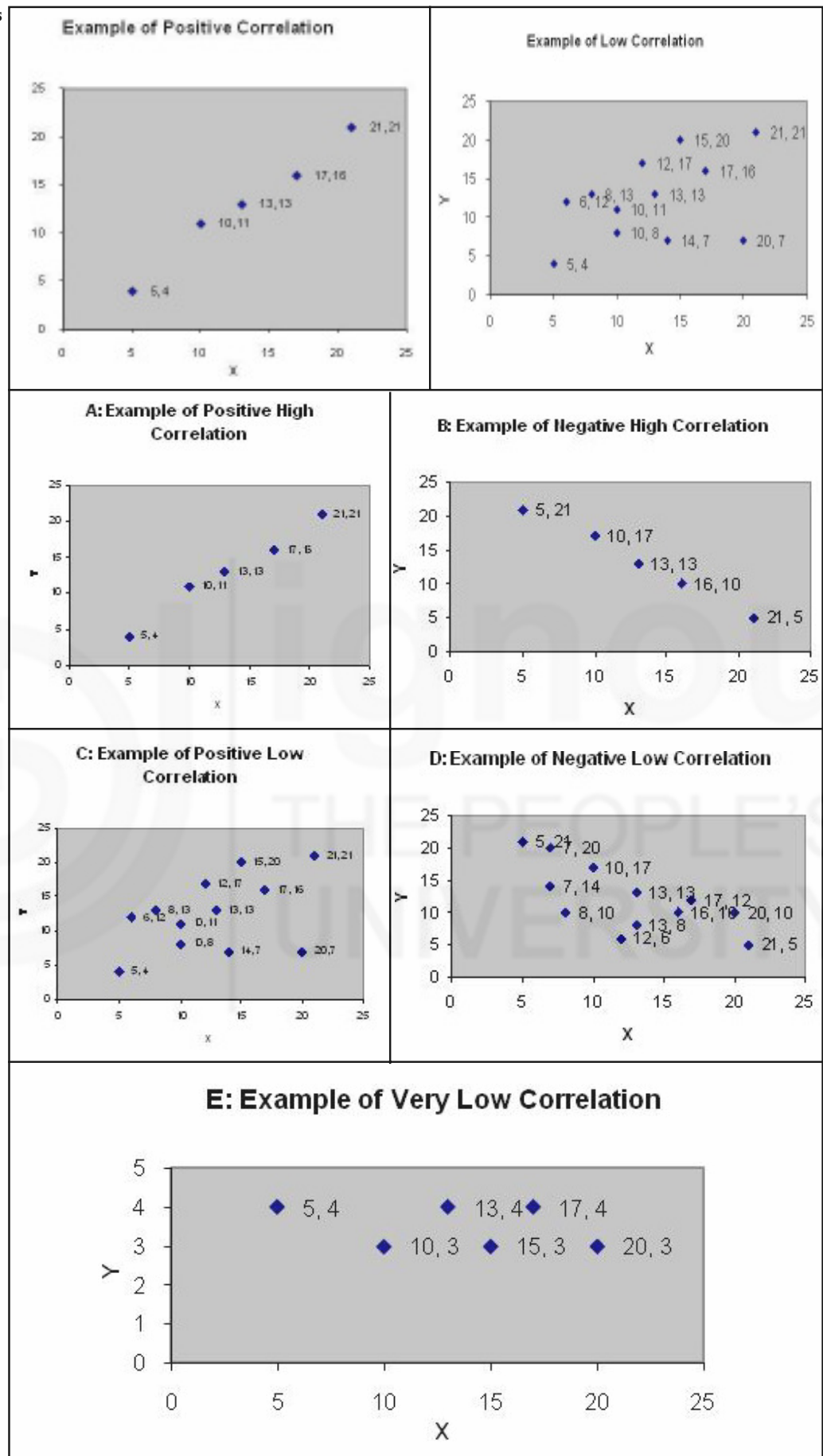


Fig. 3.1 : Scatter Diagram showing relation between two variables

The slopes of both the lines are identical in these two examples, but the scatter around the line is much greater in the second. Clearly the relationship between variables y and x is much closer in the first diagram.

If we are interested only in **measuring** the association between the two variables, then **Pearson's Correlation Coefficient (r)** gives us an estimate of the strength of the linear association between two numerical variables. Pearson's Correlation Coefficient can either be calculated by hand or the value of r can be obtained using either a calculator with built in capability to do the calculation or a variety of computer software programs. Note that in case there is curvilinear relationship, the value of r will be shown to be zero. The correlation coefficient has the following **properties**:

- 1) For any data set, r lies between '-1' and '+1'.
- 2) If $r = +1$, or -1 , the linear relationship is perfect, that is, all the points lie exactly on a straight line. If most of the points lie on the line, then it is very strong relationship and r is near to 1. If $r = +1$, variable y increases as x increases (i.e., the line slopes upwards). (See Diagram A.) If $r = -1$, variable y decreases as x increases (i.e., the line slopes downward). (See Diagram B.)
- 3) If r lies between 0 and +1, the regression line slopes upwards, but the points are scattered about the line. (See Diagram C.) The same is true of negative values of r, between 0 and -1, but in this case the regression line slopes downward. (See Diagram D.)
- 4) If $r = 0$, there is very low linear relationship between y and x. This may mean that there is no relationship at all between the two variables (i.e., knowing x tells us nothing about the value of y). (See Diagram E.).

$$\frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}}$$

Calculation of the Pearson's Correlation Coefficient

Formula for calculation of Karl Pearson's correlation co-efficient is:

$$r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}}$$

r = correlation coefficient

x = deviation from $\bar{x}$ (Arithmetic mean) of the first set of variables

y = deviation from $\bar{y}$ (Arithmetic mean) of the second set of variable

∑ = sign of summation

S_1 = standard deviation of the first set of variables

S_2 = standard deviation of the second set of variables

N = number of items in each set of variables

Example: Calculation the correlation coefficient between the following scores of history and mathematics.

Calculation the coefficient of correlation between the following scores of history and mathematics

Students	A	B	C	D	E
History (X)	65	56	69	60	75
Mathematics (Y)	60	76	74	80	85

Computation of coefficient of correlation

Student	History	Deviation		Mathematics From A.M.=65	Deviation From A.M.=65	y ²	xy
	X	x	x ²	Y	y		
A	65	0	0	60	-15	225	0
B	56	-9	81	76	+1	1	-9
C	69	+4	16	74	-1	1	-4
D	60	-5	25	80	+5	25	-25
E	75	+10	100	85	+10	100	+100
325		Σx ² =222		375	Σy ² =352		Σxy=62
=65		$\bar{Y} = \frac{375}{5} = 75$					

$$\text{Coefficient of Correlation (r)} = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \Sigma y^2}} = \frac{62}{\sqrt{222 \times 352}}$$

$$\frac{62}{\sqrt{78144}} = \frac{62}{\sqrt{280}} = +0.22$$

3.6.2 Regression: Concept and Meaning

In common language ‘regression’ means to return or to go back. In statistics, the term ‘regression’ is used to denote backward tendency which means going back to average or normal. The term ‘regression’ was first used by Sir Francis Galton.

Regression shows a relationship between the average values of two variables. So regression is average value of one variable for a given value of the other variable. It is useful for calculation of cause and effect relationship. The best average value of one variable associated with the given value of the other variable may be estimated or predicted by mean of an equation known as “Regression Equation” and also by the help of a line called as “Regression Line” which shows for a given value of other variable.

In order to estimate the best average values of the two variables, two regression equations are required and they are used separately. One equation is used for estimating the value of the first variable (X), this is called “Regression Coefficient of X on Y” or “Regression Equation of X on Y” and the second equation is used

for estimating the value of the second variable (Y) for a given value of the first variable called “Regression Coefficient of Y on X” and “Regression Equation of X on Y”.

The formula for calculation of regression coefficient are:-

1) Regression Coefficient of X on Y is $b_{xy} = r \frac{S_x}{S_y}$

2) Regression Coefficient of Y on X is $b_{yx} = r$

S_x = Standard Deviation of X series

S_y = Standard Deviation of Y series

r = Correlation coefficient between X and Y

1) Regression Equation of X and Y is

$$X - \bar{X} = r (Y - \bar{Y})$$

2) Regression Equation of Y and X is

$$Y - \bar{Y} = r (X - \bar{X})$$

X = Value of X

Y = Value of Y

$\bar{X}$ = Arithmetic Mean of X series

$\bar{Y}$ = Arithmetic Mean of Y series

S_x = Standard Deviation of X series

S_y = Standard Deviation of Y series

r = Correlation coefficient between X and Y

Example: obtain lines of regression for the following data:

Computation of Regression Equation

$\bar{X}$	$(\bar{X} - 5)$ x	x^2	Y	$(Y - 12)$ y	y^2	xy
1	-4	16	9	-3	9	12
2	-3	9	8	-4	16	12
3	-2	4	10	-2	4	4
4	-1	1	12	0	0	0
5	0	0	11	-1	1	0
6	+1	1	13	+1	1	1
7	+2	4	14	+2	4	4

8	+3	9	16	+4	16	12
9	+4	16	15	+3	9	12
$\bar{O}X = 45$ $\bar{O}x=0$ $\bar{O}x^2=60$ $\bar{O}Y=108$ $\bar{O}y=0$ $\bar{O}y^2=60$ $\bar{O}xy=57$						

Regression Coefficient (b_{xy}) = r = r =

$$= \frac{9 \times 57 - 0 \times 0}{9 \times 60 - 0^2} = \frac{9 \times 57}{9 \times 60} = \frac{19}{20} = 0.95$$

Regression Coefficient (b_{yx}) = r = =

$$= \frac{19}{20} = 0.95$$

i) The regression equation of X on Y is

$$X - \bar{X} = r (Y - \bar{Y}) \text{ is}$$

$$\bar{X} - 5 = 0.95 (Y - 12) = 0.95Y - 11.4$$

$$\bar{X} = 0.95Y - 11.4 + 5$$

$$\bar{X} = 0.95Y - 6.4$$

ii) The regression equation of Y on X is

$$Y - \bar{Y} = r (\bar{X} - \bar{X}) \text{ is}$$

$$Y - 12 = 0.95 (\bar{X} - 5) = 0.95\bar{X} - 4.75$$

$$Y = 0.95\bar{X} - 4.75 + 12$$

$$Y = 0.95X + 7.25$$

Differences between Regression and Correlation

Sl.	Correlation	Regression
1	Correlation quantifies the degree to which two variables are related. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does.	Regression finds out the best fit line for a given set of variables.
2	With correlation you don't have to think about cause and effect. You simply quantify how well two variables relate to each other.	With regression, you do have to think about cause and effect as the regression line is determined as the best way to predict Y from X.

3	With correlation, it doesn't matter which of the two variables you call "X" and which you call "Y". You'll get the same correlation coefficient if you swap the two.	With linear regression, the decision of which variable you call "X" and which you call "Y" matters a lot, as you'll get a different best-fit line if you swap the two. The line that best predicts Y from X is not the same as the line that predicts X from Y.
4	Correlation is almost always used when you measure both variables. It rarely is appropriate when one variable is something you experimentally manipulate.	With linear regression, the X variable is often something you experimentally manipulate (time, concentration...) and the Y variable is something you measure.
5	In correlation, on the other hand, our focus is on the measurement of the strength of such a relationship.	In regression analysis, we examine the nature of the relationship between the dependent and the independent variables. In regression we try to estimate the average value of one variable h m the given
6	In correlation, all the variables are implicitly taken to be random in nature.	In regression, at our level, we take the dependent variable as random, or stochastic, and the independent variables as non-random or fixed.

In this session you read about correlation and regression, now answer the questions given in Check Your Progress-3

Check Your Progress 3

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) Differentiate between correlation and regression.

.....

.....

.....

.....

.....

.....

3.7 HYPOTHESIS TESTING AND INFERENCIAL STATISTICS

3.7.1 Understanding True Difference

The analysis and interpretation of the results of our study must be related to the objectives of study. It is important to tabulate the data in univariate and/ or bi-variate or multivariate tables appropriate to the research objectives. We may find some interesting results. **For example**, in a study on nutrition, we find that 30% of the women included in the sample are anaemic as compared to only 20% of the men. How should we interpret this result?

- The observed difference of 10% might be a **true difference**, which also exists in the total population from which the sample was drawn.
- The difference might also be **due to the chance**; in reality there is no difference between men and women, but the sample of men just happened to differ from the sample of women. One can also say that the observed difference is due to sampling variation.
- A third possibility is that the observed difference of 10% is due to defects in the study design (also referred to as **Bias**). For example, we only used male interviewers, or omitted a pre-test, so we did not discover that anemia is a very important topic for women which require a female investigator.

If we feel confident that an observed difference between two groups cannot be explained by bias, we would like to find out whether this difference can be considered as a true difference. We can only conclude that this is the case if we can **rule out chance** (sampling variation) as an explanation. We accomplish this by applying a test of significance. A **test of significance** estimates the likelihood that the observed result (e.g., a difference between two groups) is due to chance or real. In other words, a significance test is used to find out whether a study result, which is observed in a sample, can be considered as a result which indeed exists in the study population from which the sample was drawn.

3.7.2 Tests of Significance

Different sets of data require different tests of significance. Throughout this module, two major sets of data will be distinguished.

- Two (or more) **groups**, which will be compared to detect **differences**. (e.g., men and women, compared to detect differences in anemia.)
- Two (or more) **variables**, which will be measured in order to detect if there is an **association** between them. (e.g., between anemia and income.)

In order to help you choose the right test, a flowchart and matrices will be presented for different sets of data. We will discuss how significance tests work. Please keep in mind that independent groups are treated as independent populations.

i) How to state Null (H_0) and Alternative (H_1) Hypothesis:

In statistical terms the assumption that **no real difference exists between groups** in the total study (target) population (or, that **no real association exists** between variables) is called the **Null Hypothesis (H_0)**. The **Alternative Hypothesis (H_1)** is that there **exists a difference between groups** or that a **real association exists** between variables. Examples of null hypotheses are

- There is no difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males do not drink more alcohol than females.
- There is no association between families' income and malnutrition in their children.

If the result is statistically significant, we reject the **Null Hypothesis (H_0)** and accept the **Alternative Hypothesis (H_1)** that there is real difference

between two groups, or a real association between two variables. Examples of alternative hypotheses (H_1) are:

- There is a difference in the incidence of measles between vaccinated and non-vaccinated children.
- Males drink more alcohol than females.
- There is an association between families' income and malnutrition in their children.

Be aware that 'statistically significant' does not mean that a difference or an association is *of practical importance*. The tiniest and most irrelevant difference will turn out to be statistically significant if a large enough sample is taken. On the other hand, a large and important difference may fail to reach statistical significance if too small a sample is used.

ii) The Concept of Type I and Type II Error

There are four ways in which conclusion of the test might relate to in our study (i) true positive (ii) true negative and (iii) false positive and (iv) false negative. These may be expressed in terms of error in statistical test of significance in following terms:

Type I error (α): We reject the null hypothesis when it is true, or false positive error, or type I error ' α ' (called alpha). It is the error in detecting true effect.

In the above example, type I error would mean that the effects of two drugs were found to be different by statistical analysis, when, in fact, there was no difference between them.

Type II error (β): We accept the null hypothesis when it is false or false negative error; or simply, type II error ' β ' (called beta) can be stated as failure to detect true effect. In the above example, type II error would mean that the effects of two drugs were not found different by statistical analysis, when in fact there was difference.

The definition can be summarized as given below.

Actual Situation			
Investigator's Decision		True H_0	False H_0
	Accept Null hypothesis	Correct Acceptance	Error (Type II)
	Reject Null hypothesis	Error (Type I)	Correct Rejection

Note: Alpha (α) and beta (β) are the Greek letters and are used to denote probabilities for **type I error** and **type II error** respectively.

We would like to carry our test, i.e., choose our critical region so as to minimize both types of errors simultaneously, but this not possible in a given fixed sample size. In fact decreasing one type of error may very likely increase the other type. In practice, we keep type I error (α) fixed at a specified value (i.e., at 1% or 5%).

3.8 STATISTICAL TESTS

Depending on the aim of your study and the type of data collected, you have to choose appropriate tests of significance. Before applying any statistical test, state the null hypothesis in relation to the data to which the test is being applied. This will enable you to interpret the results of the test. The following sections will explain how you will choose an appropriate statistical test to determine differences between groups or associations between variables. Although there are many statistical tests used in drawing inferences, here we will confine our discussion to four main types of tests:

- i) χ^2 test
- ii) T-test
- iii) Z- test
- iv) F-test

3.8.1 Chi-Square Test (χ^2)

Chi-square test is termed as a non parametric test. Karl Pearson first introduced the concept of chi-square and its application in testing statistical hypothesis. The value of chi-square is determined by (i) taking the difference between each observed frequency (f_o) and the corresponding expected theoretical frequency (f_e) (ii) squaring each difference (iii) dividing each squared difference by the corresponding expected theoretical frequency and then (iv) adding all the quotient. The value of chi-square is represented by the symbol χ^2

Thus $\chi^2 =$

Uses of chi-square Test

The chi-square test is very powerful tool in the hands of statisticians for testing hypothesis of a variety of statistical problems. The most important purposes served by the application of test of chi-square are follows:

- 1) test of goodness fit – the chi-square test is used for the comparison of observed frequencies with the expected theoretical frequencies in a sample.
- 2) test of Independence- the chi-square test is widely used to test the independence of attributes.
- 3) Test of homogeneity- the chi-square test is also used to test the homogeneity of attributes in respect to of a particular characteristic.

Formula used for computation of

$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

χ^2 = Chi-Square

F_o = Observed frequency

F_e = Expected frequency

Σ = Sum total

Example: Compute the chi-square of data given in table below:

Statistical Tools

Computation of Chi-square test

		f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
1.	Favorably	20	27	-7	49	1.81
2.	Unfavourably	40	27	13	169	6.25
3.	Undecided	21	27	- 6	36	1.33
		81	81			9.39

Follows steps will be used in assessing the level of significance:

Step-1 Determining the Degree of freedom- The Chi-Square test depends on degree of freedom. The degree of freedom deals with rows and columns of a table. The formula to calculate degree of freedom is

$$df = (c-1)(r-1)$$

df = degree of freedom

C = columns of a table

r = rows of the table

the above table (question) has 3 rows and 2 columns.

$$df = (C-1)(r-1)$$

$$= (3-1)(2-1)$$

$$= 2 \times 1$$

$$= 2$$

Step-2 Determining the Critical Value- χ^2 has pre-determined value. It requires significance level (5% or 1%) for the computed degree of freedom.

The df is 2. The critical value at 5% level is 5.991 and at 1% level is 9.210 by referring to χ^2 table.

Step-3 Comparing the critical value of Chi-Square with Computed Value- the computed χ^2 value is 9.39. It is higher than 5% and 1% level table value. So it is significant. Consequently null hypothesis is rejected in favour of alternative hypothesis.

3.8.2 T -Test

A t-Test is a statistical hypothesis test. The T-Statistic was introduced by W.S. Gossett under the pen name “student”. Therefore, the T-test is also known as the “student T-test”. The T-test is a commonly used statistical analysis for testing hypothesis, since it is straight forward and easy to use. Additionally, it is flexible and adoptable to a broad range of circumstances. The T-test is applied, if you have a limited sample, usually sample size is less than 30.

$$t = \frac{\bar{d}}{S(\bar{X}_1 - \bar{X}_2)}$$

Where,

t = t-test

$\bar{d}$ = mean difference

S = standard deviation

$\bar{X}_1$ = Mean of first set of variables

$\bar{X}_2$ = Mean of second set of variables

Calculation of $\bar{d} = \frac{\sum d}{N}$

$$S = \sqrt{\frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}}$$

Example: An IQ test was administered to 5 person before and after they were trained. The result are given below.

Candidates	I	II	III	IV	V
IQ before Training	110	120	123	132	125
IQ after Training	120	118	125	136	121

Test whether there is any change in IQ after training programme

Candidates	I Q before training x_1	I Q before training x_2	Difference $(x_2 - x_1) d$	d^2
I	110	120	10	100
II	120	118	-2	4
III	123	125	2	4
IV	132	136	4	16
V	125	121	-4	12
			$\sum d = 10$	$\sum d^2 = 140$

Estimated standard deviation of population = 6

=

=

=

=

=

$$= 2$$

$$t = \frac{\bar{d}}{s(\bar{X}_1 - \bar{X}_2)} = \frac{2}{2.45} = 0.816$$

3) Level of significance: $\alpha=0.01$

4) Decision At 0.01 level of significance for $5-1=4$ degrees of freedom, the critical value of $t = 4.6$ (using t-table) but the computed value of $t = 0.816$ is less than the critical value of $t = 4.6$ [$t = 0.816 < t = 4.6$]. Hence the computed value of $t = 0.816$ falls in the acceptance region. Thus the null hypothesis ($r_1 = r$) is accepted. So it may be concluded that there is no change in IQ after

$$\frac{\sqrt{100/5-1} \cdot 10}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \sigma / \sqrt{n} = \sqrt{\frac{30}{5}} = 2.45$$

3.8.3 Z -Test

Z-Test is another type of test like T-test applied to compare sample and population means to know if there is a significant difference between them. Z- Test is usually applied in large sample size, having more than 30 sample.

The formula for calculation of Z-Test is

=

Where

x_1 = mean of the first variable

x_2 = Mean of the Second variable

S_1 = Standard deviation first equation

S_2 = Standard deviation second equation

n_1 = Standard size of first

n_2 = Standard size of Second

Example: The score in mathematics for boys and girls is given in table below, calculate whether there is significant difference in score between them.

Scores of Boys		Scores of Girls	
40	30	22	42
35	20	33	19
25	11	26	26
26	36	33	29
24	39	44	39
20	44	20	49
45	19	41	23
43	28	33	15
28	36	37	40
33	27	27	26
29	34	18	27
31	18	19	28
41	47	44	11
49	16	32	31
21	47	22	29
34	22	36	25

$$\bar{X}_1 = 31.19$$

$$\bar{X}_2 = 29.56$$

$$S_1 = 10.13$$

$$S_2 = 9.56$$

$$S_1^2 = 102.802$$

$$= 85.67$$

$$Z =$$

$$= \frac{1.62}{2.42} = 0.67$$

Interpretation: The tabled value of z is 1.96. Since $Z < -1.96$ ($0.67 < 1.96$), we reject H_0 . It means that there is no significant difference between scores of boys and girls.

3.8.4 F -Test

The F-test was first developed by R.A. Fisher. Hence it is known as Fisher's test or more commonly as F test. The f-test is used either for testing the hypothesis

about the equality of two population variances of the equality of two or more population means. The ratio of two sample variances.

The formula for calculation of f is:

$$F =$$

Where,

= variance of first set of data

= variance of second set of data

Example

The time taken by workers in performing a job by method I and method II is given below.

Method I	20	16	23	27	23	22	
Method II	27	33	42	35	32	34	38

Do the data show that the variances of time distribution in a population from which these samples are drawn do not differ significantly

Solution

Computation of Variances

$$S^2 = \frac{\sum d_i^2}{n} - \left\{ \frac{\sum d_i}{n} \right\}^2$$

Method-I		
X_1	$X_1 - 22 = d_1$	d_1^2
20	-2	4
16	-6	36
26	4	16
27	5	25
23	1	1
22	0	0
	$\sum d_1 = +2$	$\sum d_1^2 = 82$

Method-II		
X_2	$X_2 - 34 = d_2$	d_2^2
27	-7	49
33	-1	1
42	8	64
35	1	1
32	-2	4
34	0	0
38	4	16
	$\sum d_2 = 3$	$\sum d_2^2 = 135$

Method-I

$$= 13.55$$

$$\text{Variance} = 13.55$$

Variance

=

$$= 19.28 - 0.18 = 19.10$$

$$\text{Variance} = S_2^2 = 19.10$$

Computation of F-test statistic

$$\text{Test Statistic } F = \quad = .709$$

$$3) \text{ degrees of freedom } v_1 = n - 1 = 6 - 1 = 5$$

$$\text{And } V_2 = n_2 - 1 = 7 - 1 = 6$$

- 4) Decision- at 5% level of significance the critical value of $F=4.95$ for $v_2=6$ and $v_1=5$ degrees of freedom. But the computed value of $F=.709$ is less than the critical value of $F=4.95$. Hence the null hypothesis $\sigma_1^2 = \sigma_2^2$ is accepted. So it may be concluded that the variance of time distribution in a population from which the samples are drawn do not differ significantly.

In this session you read about different deferential statistics, now answer the questions given in Check Your Progress-4

Check Your Progress 4

Note: a) Write your answer in about 50 words.

b) Check your progress with possible answers given at the end of the unit.

- 1) What is t-test and where it is applied?

.....

.....

.....

.....

.....

- 2) What is Chi-square?

.....

.....

.....

.....

.....

3.9 LET US SUM UP

Statistics is a science that deals with the collection, organization, analysis, interpretation, and presentation of information that can be presented numerically and/or graphically to help us answer a question of interest. The information or data collected may be classified as qualitative and quantitative. It may also be classified as discrete or continuous. Frequency distribution is an improved way of presenting a data. For better and more concise presentation of the information contained in a data set, the data is subjected to various calculations. If one wants to further summarize a set of observations, it is often helpful to use a measure which can be expressed in a single number like the measures of location or measures of central tendency of the distribution. The three measures used for this purpose are the mean, median, and mode. Measures of dispersion, on the other hand, give an idea about the extent to which the values are clustered or spread out. In other words, it gives an idea of homogeneity and heterogeneity of data. Two sets of data can have similar measures of central tendency but different measures of dispersion. Therefore, measures of central tendency should be reported along with measures of dispersion. The measures of dispersion include range, percentiles, mean deviation and standard deviation.

The results we obtain by subjecting our data to analysis may actually be true or may be due to chance or sampling variation. In order to rule out chance as an explanation, we use the test of significance. In this unit we have confined our discussion to four tests i.e. χ^2 test, Z- test, t-test and f-test.

Correlation is relationship between the two sets of continuous data; for example relationship between height and body weight. Correlation statistics is used to determine the extent to which two independent variables are related and can be expressed by a measure called the coefficient of correlation. Regression, on the other hand, deals with the cause and effect relation between two sets of data. Simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared *residuals* of the model (that is, vertical distances between the points of the data set and the fitted line) as small as possible. The regression line, thus, obtained helps us to predict the value of dependent variable for a given value of independent variable.

Annex I: Table of chi-square values

Degrees of freedom	χ^2 value if $\alpha = 0.05$	χ^2 value if $\alpha = 0.01$
1	3.84	6.63
2	5.99	9.21
3	7.81	11.34
4	9.49	13.28
5	11.07	15.09
6	12.59	16.81
7	14.07	18.48
8	15.51	20.09
9	16.92	21.67
10	18.31	23.21
11	19.68	24.72
12	21.03	26.22

Annexure-II

Degrees of freedom	t-value if chosen P $\alpha = 0.05$	t-value if Chosen P $\alpha = 0.01$
1	12.71	63.66
2	4.30	9.92
3	3.18	5.84
4	2.78	4.60
5	2.57	4.03
6	2.45	3.71
7	2.36	3.50
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.20	3.11
12	2.18	3.05
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.90

18	2.10	2.88
19	2.09	2.86
20	2.09	2.85
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.80
25	2.06	2.79
30	2.04	2.76
40	2.02	2.70
60	2.00	2.66
120	1.98	2.62
infintie	1.96	2.58.

3.10 KEYWORDS

Independent variable : The characteristic being observed or measured which is hypothesized to influence an event or outcome (dependent variable), and is not influenced by the event or outcome, but may cause it, or contribute to its variation.

Dependent variable : A variable whose value is dependent on the effect of other variables (independent variables) in the relationship being studied.

Mean : The mean (or, arithmetic mean) is also known as the average. It is calculated by totalling the results of all the observations and dividing by the total number of observations.

Median : The median is the value that divides a distribution into two equal halves. The median is useful when some measurements are in ordinal scale, i.e., much bigger or much smaller than the rest.

Mode : The mode is the most frequently occurring value in a set of observations. The mode is not very useful for numerical data that are continuous. It is most useful for numerical data that have been grouped. The mode is usually used to find the norm among populations.

Range : This can be represented as the difference between maximum and minimum value or, simply, as maximum and minimum values.

Percentiles	: Percentiles are points that divide all the measurements into 100 equal parts. The 30 th percentile (P ₃) is the value below which 30% of the measurements lie. The 50 th percentile (P ₅₀), or the median, is the value below which 50% of the measurements lie.
Mean Deviation	: This is the average of deviation from arithmetic mean
Standard Deviation	: This denotes (approximately) the extent of variation of values from the mean.
Parametric statistical test:	Is a test whose model specifies certain conditions about the parameters of the parent population from which the sample was drawn.
Non-parametric statistical test	: Is a test whose model does not specify conditions about the parameters of the parent population from which sample was drawn.
Normal Distribution	: The normal distribution is symmetrical around the mean. The mean, median, and mode assume the same value if observations (data) follows a normal distribution.
Sampling Variation	: Any value of a variable obtained from the randomly selected sample (e.g., a sample mean) cannot assume the true value in the population. The variation is called a sampling variation.
Test of Significance	: A test of significance estimates the likelihood that an observed study result (e.g., a difference between two groups) is due to chance or real.

3.11 REFERENCES AND SELECTED READINGS

Altman, D.G. (1991), *Practical Statistics for Medical Research*, Chapman and Hall, London.

Barker, D.J.P. (1982), *Practical Epidemiology*. (3rd ed.), Churchill Livingstone Edinburgh, UK.

Bradford, H. A. (1984), *A Short Textbook of Medical Statistics* (11th ed.), Hodder and Stoughton London, UK.

Castle, W.M. and North P.M. (1995), *Statistics in Small Doses*. Churchill Livingstone Edinburgh, UK.

Bose, A (1988), *Statistics*, Calcutta Book House, Calcutta

Fletcher, R. H., S. W. Fletcher and E. H. Wagner (1996), *Clinical Epidemiology: The Essentials*, Lippincott Williams and Wilkins, 351 West Canadian Street Baltimore, Maryland, USA.

Glaser, A.N. (2000), *High-yield Biostatistics*, Lippincott Williams and Wilkins, 227 East Washington Square, Philadelphia, USA.

Greenhalgh, T.(1998), *How to Read a Paper: The Basics of Evidence Based Medicine*, BMJ publishing group, BMA House, Tavistock Square, London,UK.

Hicks, C.M. (1999), *Research Methods for Clinical Therapists. 3rd Edition*, Churchill Livingstone, Robert Stevenson House, 1-3 Baxter's Place,Leith Walk, Edinburgh, UK.

Kelsey, J.L., W.D. Thompson and A.S. Evans (1986), *Methods in Observational Epidemiology*, Oxford University Press, Oxford, UK.

Kidder, L.H. and C.M. Judd (1986), *Research Methods In Social Relations*, CBS College Publishing, New York, USA.

Kleinbaum, D.G., L.L. Kupper and H. Morgenstern (1982), *Epidemiologic Research - Principles and Quantitative Methods*, Van Nostrald Reinhold, New York, USA.

Riegelman, R.F. (1981), *Studying a Study and Testing a Test*, Little Brown and Company, Boston, MA, USA.

Schlesselman, J.J. (1982), *Case-Control Studies - Design, Conduct, Analysis*, Oxford University Press, Oxford, UK.

Siegel, S. (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill Book Company.

Swinscow, T.D.V. and M.J. Campbell (1998), *Statistics at Square One* (11th ed.), British Medical Association, London, UK.

3.12 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress 1

- 1) What are the important types of data?

There are two types of data: (i) qualitative data, viz., occupation, sex, marital status, religion, and; (ii) quantitative data viz., age, weight, height, income, etc. These may be further be categorized in two types viz., discrete and continuous data.

- 2) What do you understand by non-parametric test?

A non-parametric statistical test is a test whose model does not specify conditions about the parameters of the parent population from which sample was drawn.

Check Your Progress 2

- 1) What are the different measures of central tendency?

The three measures of central tendency are the mean, median, and mode.

- 2) What are the different measures of dispersion?

The measures of dispersion are range, percentiles, mean deviation and standard deviation.

Check Your Progress 3

1) Differences between Correlation and Regression

The main difference between correlation and regression is that the correlation quantifies the degree to which two variables are related. You simply are computing a correlation coefficient (r) that tells you how much one variable tends to change when the other one does. While regression finds out the best fit line for a given set of variables.

Check Your Progress 4

1) What is t-test and where it is applied?

A t-Test is a statistical hypothesis test. The T-Statistic was introduced by W.S. Gossett under the pen name “student”. Therefore, The T-test is also known as the “student T-test”. The T-test is a commonly used statistical analysis for testing hypothesis, since it is straight forward and easy to use. Additionally, it is flexible and adoptable to a broad range of circumstances. The T-test is applied, if you have a limited sample, usually sample size is less than 30.

2) What is Chi-square?

Chi-square test is termed as a non parametric test. Karl Pearson first introduced the concept of chi-square and its application in testing statistical hypothesis. The value of chi-square is determined by (i) taking the difference between each observed frequency (f_o) and the corresponding expected theoretical frequency (f_e) (ii) squaring each difference (iii) dividing each squared difference by the corresponding expected theoretical frequency and then (iv) adding all the quotient. The value of chi-square is represented by the symbol χ^2 .

UNIT 4 DATA PROCESSING AND ANALYSIS

Structure

- 4.1 Introduction
- 4.2 Data Measurement and its Types
- 4.3 Tabulation and Interpretation of Data
- 4.4 Let Us Sum Up
- 4.5 Keywords
- 4.6 References and Selected Readings
- 4.7 Check Your Progress – Possible Answers

4.1 INTRODUCTION

The purpose of data analysis is to identify whether research assumptions were correct or not, and to highlight possible new views on the problem under study. The ultimate purpose of analysis is to answer the research questions outlined in the objectives with the collected data. However, before we look at how variables may be affecting one another, we need to summarize the information obtained on each variable in simple, tabular form, or, in a figure.

Some of the variables may produce numerical (continuous) data, while other variables produce categorical data. In analyzing our data, it is important, first, to determine the type of data that we are dealing with. This is crucial because the type of data used largely determines the type of statistical techniques that should be used to analyze the data. Once the data is processed, tables and graphs are prepared, and the report writing work may be initiated.

After studying this unit, you should be able to:

- define data and describe various types and nature of data.
- describe techniques of data processing, tabulation, and presentation.
- describe and interpret data from tables that have been generated.

4.2 DATA MEASUREMENT AND ITS TYPES

4.2.1 Data Measurement

Measurement is the process of observing and recording the observations that are collected as part of a research effort and the process of assigning numbers to objects or observations. But do we measure abstract concepts like happiness, quality of life, personality, opinion, etc? You have to understand that the qualities of good measurement are: (i) Precise; (ii) Unambiguous; (iii) Free from errors; (iv) Valid; (v) Reliable, and; (vi) Practical.

The tools, which are developed for measurement/ collection of data, should be

- valid to measure the characteristic which it is intended to measure
- reliable to the extent to which an experiment/ test/ or any measuring procedure yields the same result on repeated trials

- sensitive enough to detect differences in a characteristics
- specific enough to represent only the characteristic of interest
- appropriate to the objectives of the study
- provides adequate distribution of response in the study population
- meets the objective of the study

These are key points to keep in mind while developing effective tools for data collection. However, the following factors may affect the validity of the data, so remedial steps should be taken accordingly:

Respondents: reluctance, modesty, having little knowledge about details of the research problem to answer, may not admit ignorance, guessing about response, feel boredom due to long questioning, fatigue, and anxiety, etc.

Situation: lack of support to field investigators, lack of assurance on anonymity and confidentiality, etc.

Investigator/ Interviewer: may reword or reorder questions, style, and look, carelessness in filling the reply, incorrect coding, or calculation of scores.

Instrument: complex words, ambiguous meaning, poor printing, inadequate space for answer, response choice omission.

4.2.2 Types of Data

Data can be classified as either

- i) Primary or secondary data
- ii) Time series, cross section and panel data
- iii) Categorical and numerical data

i) Primary or Secondary Data

Primary Data

Primary data means original data that have been collected specially for the purpose in mind. Research where one gathers this kind of data is referred to as field research. Tools used for gathering primary data, for example, a questionnaire, an interview schedule etc.

Secondary Data

Secondary data are data that have been collected for another purpose and where we will use a statistical method with the primary data. This means that after performing statistical operations on primary data the results become known as secondary data. Research where one gathers this kind of data is referred to as desk research. The source for gathering secondary data, for example data from a book.

ii) Time series, Cross-Section and Panel Data

Cross-sectional data refers to data collected by observing many subjects (such as individuals, firms or countries/regions) at the same point of time, or, without regard to differences in time. Analysis of cross-sectional data usually consists of comparing the differences among the subjects. For

example, we want to measure current obesity levels in a population. The cross-sectional data provides us with a snapshot of that population, at that one point in time. Note that we do not know, based on one cross-sectional sample, if obesity is increasing or decreasing; we can only describe the current proportion.

Time series data is also known as longitudinal data, which follows one subject's changes over the course of time. For example, the average production of wheat from 1990 to 2009.

Panel data combines both cross sectional and time series data and looks at multiple subjects and how they change over the course of time. Panel analysis uses panel data to examine changes in variables over time, and differences in variables between subjects.

iii) Categorical and Numerical Data

Categorical or Nominal: Data that can be divided into categories or groups such as male and female, and can take only discrete and not decimal values. They are called categorical or nominal data. There are two types of categorical data: they are nominal or ordinal. In nominal data, the variables are divided into a number of named categories. These categories, however, cannot be ordered one above another (as they are not greater or lesser than each other).

Continuous or Numerical: Data that can take any value, including decimal values, are called continuous data. Data that can be measured on a scale is said to be scalar. We speak of numerical data if they are expressed in numbers. There are two types of numerical data: they are discrete or continuous. Discrete data are a distinct series of numbers.

In this section, you studied about the data measurement and its types. Now answer the following questions.

Check Your Progress 1

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) What is panel data?

.....

.....

.....

.....

.....

2) What do you understand by primary data?

.....

.....

.....

.....

4.3 TABULATION AND INTERPRETATION OF DATA

The following operations need to be done to bring data into a presentable form:

- Data coding, editing and feeding
- Data tabulation
- Figures and graphs

We will now discuss these operations, one after the other, for both categorical and numerical data.

4.3.1 Data Coding, Editing and Feeding

Once data is collected the researcher thinks about the processing and analysis of the data. This is a crucial stage of research work. Here, the researcher has to consult the guide and other academicians who have done research in the related field. Therefore, it is advised that data processing must be planned in advance, and, necessarily, during the time of questionnaire formulation; and, during the time of data collection. Nowadays, with the advent of the computer, researchers might think it is the job of computer assistant, but it is not so. Before either giving it to the computer assistant, or doing it manually, a lot of activities have to be completed: (i) editing of the data which you received from the field; (ii) coding of the data; (iii) preparation of the master chart if the data is to be computed manually; and (iv) presentation of data.

i) Editing of the data

The editing of data is the first step of data processing and analysis. After collecting data, either through the questionnaire or schedule, you have to edit it. You have to carefully check for missing and wrongly entered data. Therefore, in large scale surveys and research, the company undertaking the research project appoints supervisors, or, editors, for proper checking of data. An example of editing of data is given below.

Example – Suppose that you conduct a survey of the occupation status of the people in a village, for which you have to collect data from heads of households. The questionnaire for the purpose is as follows

Name of the Respondent	: Rama Singh
Age	: 56
Sex	: Male
Cast	: General
Occupation	: Private Service
Income of the Respondent	: 10,000
Wife's name	: Rita Singh
Age	: 51
Occupation	: House wife
Monthly Income	: 800/-

In this questionnaire, the wife of Rama Singh is a housewife, but she earns Rs. 800. She may be a housewife, but by knitting she may be earning Rs. 800 per month. Therefore, her occupation may be categorized as self-employment. Like this, you have to cross check the data and try to get accurate information which will minimize error and strengthen your research findings. Therefore editing of the data is very essential before sending it for processing.

ii) Coding of the data

The questionnaire must be properly coded. Before either, sending it for feeding into the computer, or entering it in the master chart, the coding of data is necessary. The coding of data will make data entry easy. Coding of data means assigning numerical symbol to each response of the question. The purpose of giving numerical symbols is to translate raw data into numerical data, which may be counted and tabulated. An example of coding of the marital status and education is given below.

Marital Status

Married	:	01
Separated	:	02
Widow	:	03
Divorced	:	05
Never Married	:	06
Any Other	:	07

The coding of data and checking that codes are properly inserted in the questionnaire must be done during the editing of the questionnaire. It is always better to prepare a code book for your questionnaire. The coding can also be done during the time of data collection, if the code book is available with you.

iii) Entering data in the Master Chart

If you are doing tabulation manually, it is always wise to enter the data into a master chart. The master chart is a large sheet which will enable you to enter all the codes of different variables into it. It will help you to generate tables easily. An example of master chart is given below.

Master Chart

Sl. No	Name of the respondent	Variables				
		Sex	Age	Marital Status	Education	Occupation
1						
2						
3						
4						
5						
Total						

In the master chart you can enter the data of 14 sample respondents. Likewise you can expand the number of respondents in the column and variables in the rows. It is always better to enter code (numerical number) in the master chart.

iv) Entering Data into the Computer

Computers are widely used for the analysis of data. It makes the calculation much faster. The excel sheet, and the SPSS package can be used in social science research. The following steps are used in the analysis of data by using the SPSS package.

- i) Entering of data in the SPSS statistical package
- ii) Selection of procedure from the menu
- iii) Selection of variables for analysis
- iv) Examination of the output.

v) Presentation of Statistical Data

Statistical data are collected to serve a purpose. Therefore, data may be presented in such a way that it may be easily grasped and the conclusion may be drawn promptly. Generally, the following three methods are used for the presentation of statistical information.

- a) **Textual Presentation:** In this method, statistical information is presented in text form. Generally, this type of presentation is made in a descriptive way. It requires careful reading of the text in order to grasp the meaning and significance of the facts and figures given, therein. But, for most people, it is not a suitable and effective method of presentation of statistical information, because it is not easy for the reader to single out the individual information and figures. The advantage of this method is that an ordinary person can prepare and present the text and a layman can read it and grasp it.
- b) **Tabular Presentation:** In this method, statistical information is presented in the form of a table. Facts and figures are gathered, and then, incorporated in tables. Generally, this type of presentation is made in a tabular form with rows and columns. Tables summarize statistical data in a logical and orderly manner. The main advantage of this method is that tables are brief, concise, and contain only relevant figures. Tables also facilitate the comparison of figures. The only disadvantage of this method is that the presentation of tables and the interpretation of data require some skills and techniques.
- c) **Graphical Presentation:** In this case, statistical information is presented in the form of graphs and charts. Facts and figures are gathered first, and then, they are depicted in graphs and charts for presentation. Generally, this type of presentation is made through figures, diagrams, charts, or graphs. The main advantage of this method is that the facts and figures become more attractive and appealing to eye. A disadvantage of this method is that facts cannot be shown in detail and accurately.

4.3.2 Data Tabulation

Tabulation is an orderly and systematic arrangement of numerical data presented in rows and columns for the purpose of information, comparison, and interpretation. So, a statistical table is a systematic arrangement of statistical data into rows and columns. It summarizes the data in a logical and orderly manner for the purpose of presentation, comparison, and interpretation. Tabulation is, thus, a scientific process and means of recording, statistical data in a systematic and orderly manner.

A statistical table has the following five parts.

- i) **Title:** each table must have a title which convey the contents of the table. It should be clear, concise and self-explanatory. It should be written on the top of the table.
- ii) **Stub:** this is a column used for mentioning the items and their heading. It is the left most column of the table. A stub is generally marked with rows and in each row, an item is mentioned. The stub should be clear and self-explanatory.
- iii) **Caption:** this is the heading for columns, other than the stub. It is the upper part of the Table. Caption should be properly columned and worded. Sometimes, below the caption, the units of measurement and column-numbers, is called a box head.
- iv) **Body:** the main part of the table. It contains the data which are exhibited in the table. The figures inserted therein should be distinct.
- v) **Source & Footnote:** the last part of a table. If the researcher is procuring data from a secondary source, then the source of the data needs to be mentioned. For example, if you are citing data from the Census of India then the year and the department and the state need to be mentioned. After mentioning the sources, the researcher has to provide a footnote, e.g., in the same cell, if you are giving the figure, and in parenthesis, the percentage to the total, and then, it must be mentioned in the footnote. An example of a table is given below:

Title		
		Caption
STUB	BODY	BOX HEAD

Source:

Footnote:

4.3.3 Types of Tabulation

Tabulation is done based on the data. The following types of tables are generally constructed.

- i) Construction of frequency distribution table
- ii) Construction of cross- tabulation
- iii) Construction of figures and graph

i) Construction of Frequency Distribution Table

A frequency distribution table can be of two types:

- Simple frequency distribution
- Grouped frequency distribution

In constructing a simple frequency distribution, the observations are not divided into groups or classes. Only individual values are shown, whereas, in the grouped frequency distribution the observations are divided into groups or classes.

Here is a simple frequency distribution table.

Table 4.1: Mark obtained by class fifth students in mathematics

Marks	Tally of marks	Frequency (No. of students)
20	HHH	5
21	HHH-HH II	12
25	HHH-HH III	13
30	HHH-HI	10
Total		40

Table 4.2: Grouped frequency distribution

Marks obtained	Tally of marks	Frequency
1-10	HHH-HH I	11
11-20	HHH-HH-HH IIII	20
21-30	HHH-HH-HH IIII I	21
31-40	HHH-HH III	13
41-50	HHH IIII	9

While preparing the group frequency distributions, the following points have to be taken into consideration.

- The groups must not overlap, otherwise there will be confusion about which group a measurement belongs to.
- There must be continuity from one group to the next, which means that there must be no gaps. Otherwise, some measurements may not fit in a group.

- The groups must range from the lowest measurement to the highest measurement so that all of the measurements have a group to which they can be assigned.
- The groups should normally be of an equal width, so that the counts in different groups can easily be compared.

ii) Construction of Cross Tabulation

So far, we have made tables containing frequency distributions for one variable at a time, in order to partially describe our data. Depending on the objectives of our study, and the study type, we may have to examine the relationship between several of our variables at the same time. For this purpose it is appropriate to construct cross tabulation of data. Depending on the objectives and the type of study, different kinds of cross tabulations may be required. The examples of cross tabulation are given below. Here, three different types of cross tabulation of data have been given.

Example 1: A study was carried out on the degree of job satisfaction among doctors and nurses in rural and urban areas. To describe the sample, a cross tabulation was constructed which included the sex and the residence (rural, or urban) of the doctors and nurses interviewed. This was useful, because in the analysis, the opinions of male and female staff had to be compared separately for rural and urban areas.

Table 4.3: Type of teachers by residence

Residence	Type of teachers		Total
	Principles	Teachers	
Rural	10 (16%)	69 (38%)	79 (33%)
Urban	51 (84%)	113(62%)	164 (67%)
Total	61(100%)	182 (100%)	243(100%)

Interpretation: Table 4.3 shows that a higher percentage of teachers than principals work in rural areas, but, that, overall, a greater proportion of teaching staff works in urban areas (67%).

Table 4.4: Sex of teachers by residence

Residence	Sex of teachers		Total
	Male	Female	
Rural	54 (43%)	25 (21%)	79 (33%)
Urban	71 (57%)	93 (79%)	164 (67%)
Total	125(100%)	118 (100%)	243(100%)

Interpretation: It can be concluded from Table 4.4 that there are more males serving in rural areas than females. These males in rural areas are apparently teachers.

To obtain an overview of the distribution of principals and teachers by gender in rural and urban areas, we can construct the following two-by-four cross-table.

Table 4.5: Residence and sex of principals and teachers

Teaching staff		Residence		Total
		Rural	Urban	
Principals	Males	8 (10%)	35 (21%)	43 (18%)
	Females	2 (3%)	16 (10%)	18 (7%)
Teachers	Males	46 (58%)	36 (22%)	82 (34%)
	Females	23 (29%)	77 (47%)	100 (41%)
Total		79(100%)	164 (100%)	243(100%)

Interpretation: This table shows, in a glance, that male nurses dominate the rural health services. It also indicates that males dominate in the medical profession, (18% M > 7% F doctors) but, that, overall, there are more female than male nurses, and that the females are mainly clustered in towns.

The data in the tables is usually listed in absolute figures, as well as in relative frequencies (percentages or proportions). As already seen in Unit 3, for numerical data (such as age) the mean, median, and/or mode, with standard deviation may be calculated as well, to describe the sample.

General hints while constructing tables

- Make sure that all the categories of the variables presented in the tables have been specified, and that they are mutually exclusive (i.e., no overlaps and no gaps) and are exhaustive.
- When making cross-tabulations, check that the column and row counts correspond to the frequency counts for each variable.
- Also, check that the grand total in the table corresponds to the number of subjects in the sample. If not, an explanation is required. This could be presented as a footnote. (missing data, for example.)
- Think of a clear title for each table. Also, be sure that the headings of rows and columns leave no room for misinterpretation.
- Number your tables and graphs and keep them together with the objectives to which they are related. [Numbering of tables and graphs in a chapter (e.g., 4) may be like table 4.1, table 4.2 and Figure 4.1 Figure 4.2, etc.] This will assist in organizing your report and ensure that work is not duplicated.

iii) Construction of figures and graph

If your report contains many descriptive tables, it may be more readable, if you present the most important ones in figures. The most frequently used figures for presenting data include

• Bar charts	for categorical data
• Pie charts	
• Histograms	for continuous data diagrams
• Line graphs	
• Scatter	
• Maps	

We will now look at example of the abovementioned figures that can be used for presenting data.

1) *Bar Chart*

The data from Example 2 can be presented in a bar chart, using either absolute frequencies or relative frequencies/percentages and an example is given in Figure 4.1 below.

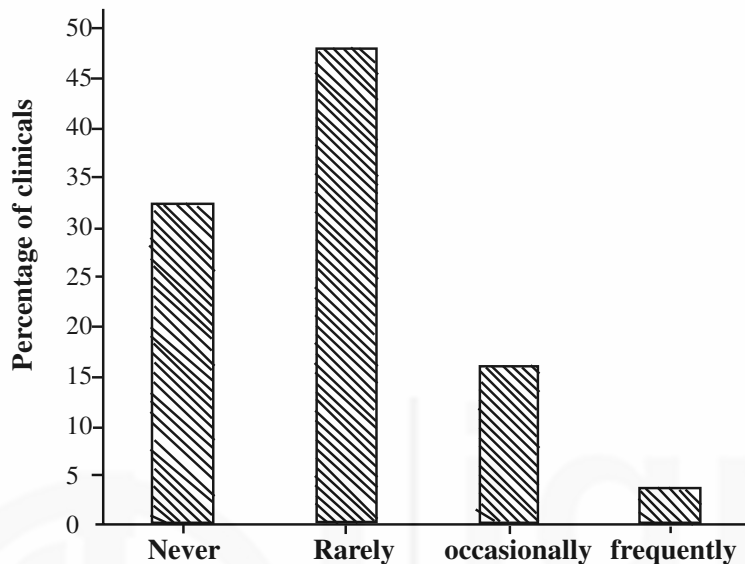


Figure 4.1: Relative frequency of shortage of anti-malaria drugs in rural health institutions (n=148)

Note that the sample size must be indicated if you present the data in percentages.

2) *Pie Charts*

A pie chart can be used for the same set of data, providing the reader with a quick overview of the data presented in a different form. A pie chart illustrates the relative frequency of a number of items. All the segments of the pie chart should add up to 100%.

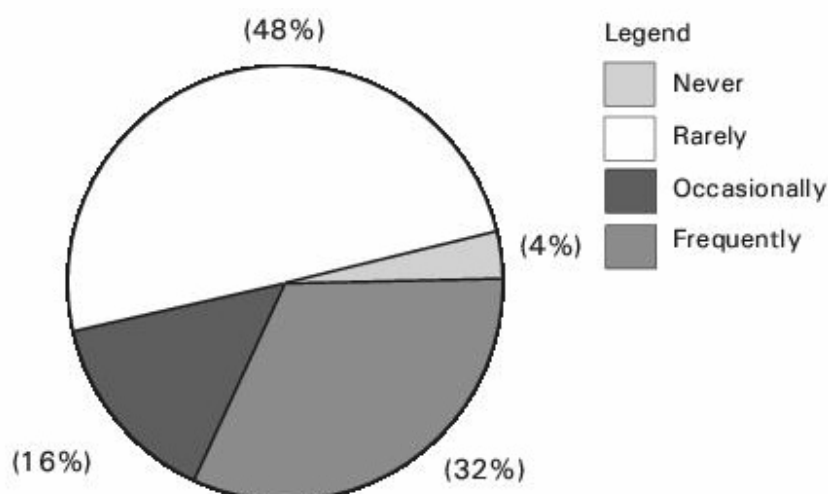


Figure 4.2: Relative frequency of shortage of anti-malaria drugs in rural health institutions (n=148)

3) Histograms

Numerical data are often presented in histograms, which are very similar to the bar charts which are used for categorical data. An important difference, however, is that in a histogram the bars are connected (as long as there is no gap between the data), whereas in a bar chart the bars are not connected, as the different categories are distinct entities. An example of histogram is given in Figure 4.3.

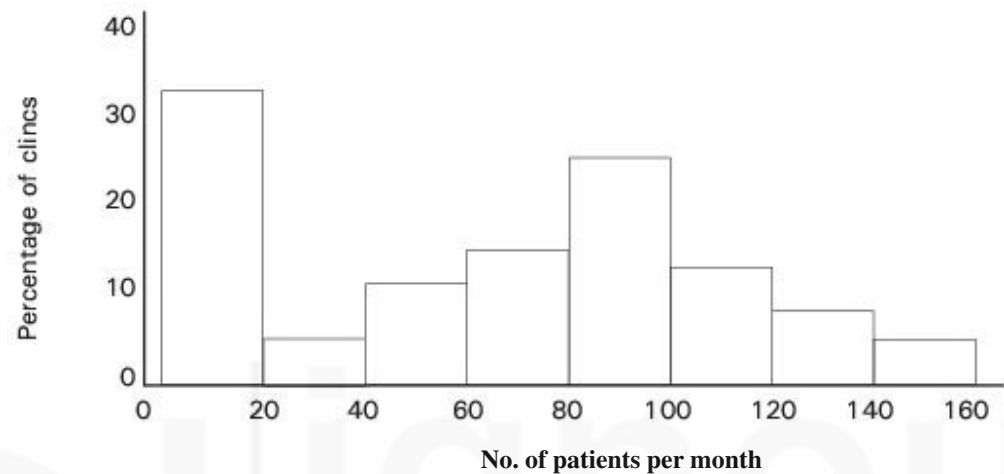


Figure 4.3: Percentage of clinics treating different numbers of malaria patients in one month (n=80).

4) Line Graphs

A line graph is particularly useful for numerical data if you wish to show a trend over time. An example of a line graph is given in Figure 4.4 .

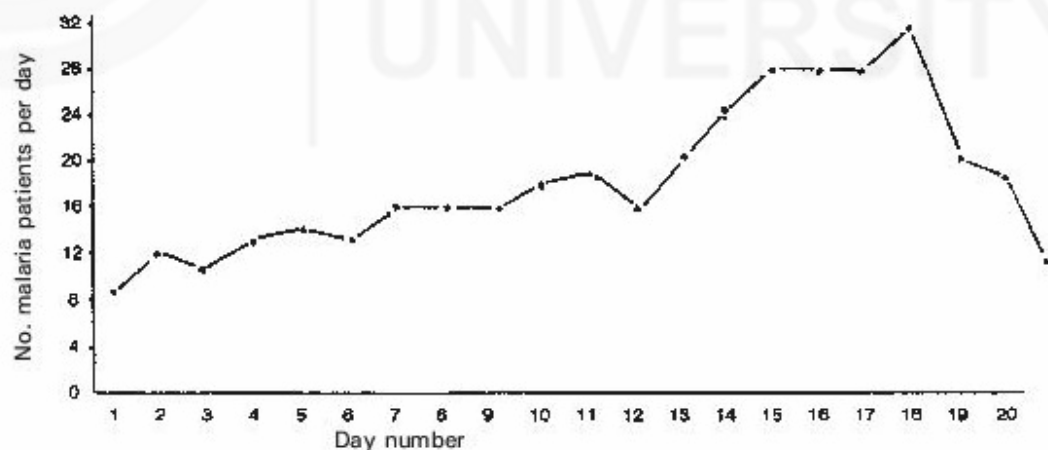


Figure 4.4: Daily number of malaria patients at the health centres in District X

It is easy to show two or more distributions in one graph, as long as the difference between the lines is easy to distinguish. Thus, it is possible to compare frequency distributions of different groups, i.e., the age distribution between males and females, or cases and controls.

5) Scatter diagrams

Scatter diagrams are useful for showing information on two variables which are possibly related. The example of a scatter diagram, given below, is used where we are dealing with the concepts of association and correlation.

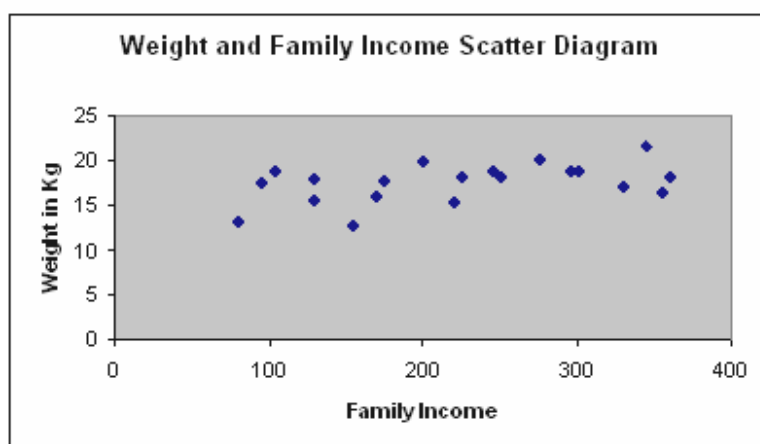


Figure 4.5: Weight of five-year-olds according to annual family income

Note: It is important that all figures presented in your research report have numbers, clear titles, and are clearly labelled (or keyed).

6) Maps

In addition to the figures above, the use of maps may be considered to present information. For instance, the area, where a study was carried out, can be shown in a map. If the study explored the epidemiology of cholera, a map could be produced showing the geographical distribution of cholera cases, together with the distribution of protected water sources, thus illustrating that there is an association. If the study related to vaccination coverage, a map could be developed to indicate the clinic sites and the vaccination coverage among under-fives in each village, perhaps showing that home-clinic distance is an important factor associated with vaccination status.

In this section, we discussed about the tabulation and interpretation of data. Now answer the following questions.

Check Your Progress 2

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) What is meant by coding of data?

.....

.....

.....

.....

.....

.....

2) What is a pie chart and where is it used?

.....

.....

.....

.....

.....

4.4 LET US SUM UP

Understanding data quality and its measurement are of utmost importance for any researcher. Poor quality data cannot be analysed properly, and, may also give results which may not be valid, or, sometimes give results which could have adverse consequences for society if those results are used as a base for policy making. The presentation of data needs to be made in simple, as well as cross table format. It is always advisable to prepare a cross tabulation plan as per study requirements, and to present data in graphical way for clarity.

4.5 KEYWORDS

Data Measurement	: Measurement is the process of observing and recording the observations that are collected as part of a research effort.
Type of Data	: Broadly there are two types of data: (i) quantitative and; (ii) qualitative which can be further classified as categorical, nominal and continuous data.
Data Quality	: The quality data can be characterized as: (i) precise, (ii) unambiguous, (iii) free from errors, (iv) valid, (v) reliable, and (vi) practical.
Data Processing	: Means the generation of frequency distribution and cross tabulation and calculation of other statistical measures.
Frequency Distribution	: Preparation of tables which distribute respondents according to a particular characteristic of sample, or research outcome.
Cross Tabulation	: This is a process of generating tables giving the outcome of interest in columns, and various characteristics of respondents, or factors affecting outcomes in rows.
Data Interpretation	: Is drawing valid and meaningful conclusions from the tables generated with the help of collected data.
Report Preparation	: Is the process of documenting the whole process of research conducted to identify the problem, or to prove some relationships, or for proving the success of some programme related activities.

4.6 REFERENCES AND SELECTED READINGS

Gibaldi, J. (1995), *MLA Handbook for Writers of Research Papers*, Modern Language Association of America, New York.

Yang, J. T. et al. (1996), *An outline of Scientific Writing: For Researchers with English as a Foreign Language*, World Scientific Publishing, Singapore.

Trochim, W. M. (1999), *The Research Methods Knowledge Base*, 2nd Edition, Online textbook, URL: <<http://www.socialresearchmethods.net/kb/>>.

Training modules of International Development Research Council, Canada.

Health System Research Modules published by WHO.

Modules on Primary Health Care, Agha Khan Foundation, Geneva.

4.7 CHECK YOUR PROGRESS - POSSIBLE ANSWERS

Check Your Progress 1

- 1) What is panel data.

Panel data combines both time series and cross sectional data and looks at multiple subjects and how they change over the course of time. Panel analysis uses panel data to examine changes in variables over time, and differences in variables between subjects.

- 2) What do you understand by primary data?

Primary data is original data that have been collected specially with a purpose in mind. Research where one gathers this kind of data is referred to as field research, for example: a questionnaire.

Check Your Progress 2

- 1) What is meant by coding of data?

A questionnaire must be properly coded. Before feeding it into a computer, or entering the data into the master chart, coding of data is necessary. The coding of data will make data entry easy. Coding of data means assigning a numerical symbol to each response of the question. The purpose of giving numerical symbols is to translate raw data into numerical data, which may be counted and tabulated.

- 2) What is a pie chart and where is it used?

A pie chart can be used for providing the reader with a quick overview of the data presented in a different form. A pie chart illustrates the relative frequency of a number of items. All the segments of the pie chart should add up to 100%.

UNIT 5 REPORT WRITING

Structure

- 5.1 Introduction
- 5.2 Types of Report
- 5.3 Writing the Research Report
- 5.4 The Preliminary Pages of Research Report
- 5.5 Main Components or Chaptering of Research Report
- 5.6 Style and Layout of the Report
- 5.7 Common Weaknesses in Report Writing and Finalizing the Text
- 5.8 Let Us Sum Up
- 5.9 References and Selected Readings
- 5.10 Check Your Progress – Possible Answers

5.1 INTRODUCTION

A research report is considered a major component of any research study as the research remains incomplete till the report has been presented or written. No matter how good a research study, and how meticulously the research study has been conducted, the findings of the research are of little value unless they are effectively documented and communicated to others. The research results must invariably enter the general store of knowledge. Writing a report is the last step in a research study and requires a set of skills somewhat different from those called for in actually conducting a research.

After reading this unit you will be able to:

- follow the various steps involved in writing a research report.
- explain the various components of a research report
- identify common mistakes committed while writing a research report.

5.2 TYPES OF REPORT

Research reports vary greatly in length and type depending on the subject. For example banks and other financial institutions prefer short balance sheet type of tabulations for their annual report. In mathematics, the report may consist of many algebraic notations, whereas a chemists report may be in the form of symbols and formulae. Students of literature usually write a long report critically analysing a writer or book.

The news items found in newspapers are also a form of report writing. Other examples of reports include book reviews, reports prepared by government bureaus, PhD theses, etc. Any research investigation may be presented in like a technical report, a popular report, an article, a monograph, or, at times, even in the form of an oral presentation. The technical report is prepared for specialists who have interest in understanding the technical procedure and terminology used in the research project. The report will be in technical language. In the technical

report, the main emphasis is on: (i) the methods employed; (ii) assumptions made in the course of study and; (iii) the detailed presentation of the findings, including their limitations and supporting data.

Popular data is intended for persons who have limited interest in the technical aspects of the research methodology and research findings. The audience will include laymen and even top executives who want summary reports. The popular report is one which gives emphasis on simplicity and attractiveness. The simplification should be sought through clear writing, minimizing of technical, particularly mathematical details, and liberal use of charts and diagrams. Attractive layouts along with large print and many subheadings is another feature of a popular report. In such a report, emphasis is given on practical aspects and policy implications.

5.3 WRITING THE RESEARCH REPORT

Once the data collection and analysis work is over, the researcher will start writing the research report. Social and development research reports need to

- have a logical, clear structure
- be to the point
- use simple language, and have a pleasant layout

Just as an architect has to draw a layout plan for a house that is being designed, you first have to make an **outline** for your report. This outline will contain a head, a body, and a tail. **The head** consists of a description of your problem within its context (the country and research area), the objectives of the study and the methodology followed. This part should not comprise more than one quarter of the report, otherwise it becomes top-heavy. **The body** will form the bigger part of your report: it will contain the research findings. **The tail**, finally, consists of the discussion of your data, conclusions, and recommendations.

Before you start writing, it is essential to group and review the data you have analysed by objective. Check whether all data has indeed been processed and analysed as you planned in the research protocol/proposal which is duly approved. Draw major conclusions and relate these to the research literature. Again, you may be inspired to go back to your raw data and refine your analysis, or to search for additional literature to answer questions that the analysis of your data may evoke. Compile the major conclusions and tables or quotes from qualitative data related to each specific objective. You are now ready to draft the report.

The research report will have, broadly, three parts.

Part I : The Preliminary Pages

Part II : The Main Text of the Research Report

Part III : The End Matter

5.4 THE PRELIMINARY PAGES OF RESEARCH REPORT

The preliminary pages of the research report should have the following main constituents.

- Title and cover page
- A foreword
- Preface
- Acknowledgements
- Table of contents
- List of tables
- List of figures
- List of appendices
- List of abbreviations
- Executive Summary

i) Title and Cover page

The cover page should contain the title, the names of the authors with their designations, the institution that is publishing the report with its logo, (e.g., Health Systems Research Unit, Ministry of Health), the month, and the year of publication. The title could consist of a challenging statement or question, followed by an informative subtitle covering the content of the study and indicating the area where the study was implemented. However, this is suggestive in nature and should not be considered standard. It would be appropriate if the cover page is designed by an expert in computer graphics who may be suggested to include some important photograph related to identity of organization or problem under study or from the field within the background. Design software may be used. An example of a title of a research report is given in the box below.

Title of the research report

Labour Migration and its Implication on Rural Economy of Indo-Gangetic Plains of India

ii) Foreword

A **foreword** is usually a short piece of writing found at the beginning of a book or other piece of literature, before the introduction. This may or may not be written by the primary author of the work. Often, a foreword will tell of some interaction between the writer of the foreword and the story, or, the writer of the story. A foreword to later editions of a work often explains how the new edition differs from previous ones. Unlike a preface, a foreword is always signed. An example of a foreword is given in the box below.

Foreword

Migration of all kinds, particularly income seeking migration across state boundaries, has attracted much attention in recent scholarly and policy literature. This study provides sufficient evidences of the effect of labour migration, more specifically, male outmigration on the rural economy of the Indo-Gangetic region. The number of districts of high and moderately high male outmigration has increased. The findings reveal the holistic scenario of migration led changes in agricultural and household domains. I am sure that this volume would be of great interest to researchers, policy makers, and development agencies while framing strategies for agricultural and rural development.

iii) Preface

A preface, by contrast, is written by the author of the book. A preface generally covers the story of how the book came into being, or how the idea for the book was developed; this is often followed by thanks and acknowledgments to people who were helpful to the author during the time of writing. A preface is an introduction to a book or other literary work written by the work's author. An example of preface is given in the box below.

Preface

The present study was conducted in three states of Bihar, Uttar Pradesh, and Punjab to study various aspects of labour migration, and its impact on rural economy in the Indo-Gangetic plains in India. The study focused on labour outmigration across two states of the Indo-Gangetic Region and in-migration in Punjab. The results of this study would help researchers, policy makers and planners as well as development agencies in addressing various issues of labour migration and its implication in India.

iv) Acknowledgements

It is good practice to thank those who supported you technically or financially in the design and implementation of your study. You should not forget to thank your research guide and your employer, too, who has allowed you to invest time in the study; and, the respondents may be acknowledged. You should not forget to acknowledge the contribution of computer professionals, library staff, local officials, and the community at large that provided the information. Acknowledgements are usually placed right after the title page or at the end of the report, before the references. An example of acknowledgement is given in the box below.

Acknowledgements

I take this opportunity to thank the Indian Council of Agricultural Research for providing funds and facilities for the project. I offer my sincere thanks to the Director, Indian Agricultural Research Institute for his encouragement and support for pursuing this study. I am also grateful to the head, Division of Agricultural Economics, IARI for providing all needed support, encouragement, and technical guidance. All the Research Associates, Senior Research Fellows and technical assistants working under the project deserve special appreciation for their hard work and sincere efforts in completing this project.

v) Table of Contents

A table of contents is essential. It provides the reader a quick overview of the chapters with major sections and sub sections of your report, and page references, so that the reader can go through the report in a different order, or skip certain sections. The sections and sub sections within each chapter may be given numbers that are specific to the chapter. For example, a section in chapter III may be given no as 3.1; and, a sub section as 3.1.1. An example of a table of contents is given below.

S. No.	Contents	Pages
1	Introduction	
2	Review of Literature	
3	Methodology	
3.1	Data	
3.2	Analytical Tools	
3.3	Profile of Area Under Study	
4	Research Findings	
4.1	Macro Level Evidences	
4.2	Evidences from filed Survey	
5	Discussion	
6	Conclusions and Policy Implications	
7	References	
	Appendix	

vi) List of Tables

If you have many tables or figures, it is essential to list these also in a table of contents with formatted with page numbers. The initial letters of the key words in the title are capitalized and no terminal punctuation is used. An example is given below.

List of Tables

S. No.	Name of the Table	Pages
2.1	Sampling Pattern of Households in the Study Area	
3.1	Migrants by Last Residence in India	
3.2	Total Inter-State Migrants by Place of Birth in Major States	
3.3	Social Characteristics of Households in the Study Area	
.	.	
.	.	
.	.	

vii) List of Figures

The list of figures appears in the same format as the list of tables, titled List of Figures.

viii) List of Appendices

The appendices will contain any additional information that the researcher have collected while carrying out the study. It may be a questionnaire, a letter of appreciation, a government notification, etc. The list of appendices appears in the same format as the list of tables.

ix) List of Abbreviations (optional)

If abbreviations or acronyms are used in the report, these should be stated in full in the text the first time that they are mentioned. If there are many, they should be listed in alphabetical order as well. The list can be placed before the first chapter of the report.

The table of contents and lists of tables, figures, abbreviations should be prepared last, as only then can you include the page numbers of all chapters and sections, sub-sections in the table of contents. Then, you can also finalise the numbering of figures and tables and include all abbreviations. An example of a List of Abbreviations follows.

x) List of Abbreviations

List of Abbreviations	
AI	: Agreement Index
CMIE	: Centre for Monitoring of Indian Economy
CV	: Coefficient of Variation
DEA	: Data Envelopment Analysis

xi) Executive Summary

The summary should be written only *after* the first or even the second draft of the report has been completed. It should contain

- a very brief description of the problem (**Why** this study was needed)—the main
- objectives (**What** has been studied)
- the place of study (**Where**)
- the type of study and methods used (**How**)
- the major findings and conclusions
- the major (or all) recommendations.

The summary will be the first (and for busy programme manager/decision makers most likely the only) part of your study that will be read. Therefore, it demands thorough reflection and is time consuming. Several drafts may have to be made, each discussed by the research team as a whole.

As you may have collaborated with various groups during the drafting and implementation of your research proposal, you may consider writing **different summaries** for each of these groups. For example, you may prepare different summaries for policymakers and programme managers, for implementing staff of lower levels, for community members, or for the public at large (newspaper, TV). In a later stage, you may write articles in scientific journals.

In this section, we discussed about the types of report and the contents to be included in the preliminary pages of research report. Now answer the following questions.

Check Your Progress 1

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

1) What should be included on the cover page of a research report?

.....

.....

.....

.....

.....

2) What is the importance of writing an acknowledgement in a research report and where should it be placed?

.....

.....

.....

.....

.....

5.5 MAIN COMPONENTS OR CHAPTERING OF RESEARCH REPORT

The Main Text includes the following chapters

- Introduction
- Review of Literature
- Methodology
- Research Findings
- Discussion
- Conclusion and Recommendations
- Summary

5.5.1 Chapter 1: Introduction

The introduction is a relatively easy part of the report that can best be written after a first draft of the findings has been made. It should certainly contain some relevant (environmental/ administrative/ economic/ social) background data and information about the topic on which you are carrying out research for example if you are doing research on primary education, then a brief about the status of primary education, such as their number, state-wise break up, expenditure on primary education, etc., need to be described. You may make additions to the corresponding section in your research proposal, including additional literature, and use it for your report.

Then, the statement of the problem should follow, again, revised from your research proposal with additional comments and relevant literature collected during the implementation of the study. It should contain a paragraph on what you hope/ hoped to achieve from the results of the study. Enough background should be given to make clear to the reader why the problem was considered worth investigating.

The general and specific objectives should also be included in this chapter. If necessary, you can adjust them slightly for style and sequence. However, you should not change their basic nature. If you have not been able to meet some of the objectives of the project, this should be stated in the methodology section, and in the discussion of the findings. The objectives form the heart of your study. They determined the methodology you chose and will determine how you structure the reporting of your findings.

5.5.2 Chapter 2: Review of Literature

Global literature can be reviewed in the introduction to the statement of the problem if you have selected a problem of global interest. Otherwise, relevant literature from individual countries may follow as a separate literature review after the statement of the problem. A literature review is a body of text that aims to review the critical points of current knowledge and or methodological approaches on a particular topic. Literature reviews are secondary sources, and, as such, do not report any new or original experimental work. Its ultimate goal is to bring the reader up to date with current literature on a topic, and forms the basis for another goal, such as future research that may be needed in the area.

A well-structured literature review is characterized by a logical flow of ideas; current and relevant references with consistent, appropriate referencing style; proper use of terminology; and an unbiased and comprehensive view of the previous research on the topic. One research study should be presented in one paragraph and it should mention the name of the researcher, year of study, topic and area of study, sample size, main objectives, and findings of the study. An example of a review is given in the box below.

Review of Literature

Singh (2008) conducted a study on labour out-migration from the Indo-Gangetic plains of India. The study provides sufficient evidence of the effect of male out-migration on the rural economy of the Indo-Gangetic plains of India. Male out-migration has resulted in gender role reversal in terms of decision making on important household and farm issues. Besides, the women of the migrant households had to take up many male specific activities, like land preparation, seed selection, broadcasting, irrigation, and herbicide application. The study also proved that the crop returns of non-migrant households were significantly higher than that of migrant households in case of both rice and wheat cultivation. The technical, allocative and economic efficiencies of non-migrant households was much higher than the migrant households in both rice and wheat cultivation.

5.5.3 Chapter 3: Methodology

The methodology adopted in conducting the study must be fully explained. The scientific reader would like to know about the basic design of the study, the

methods of data collection, information regarding the sample used in the study, the statistical analysis adopted and the factors limiting the study. The methodology section should include a description of

- the study type
- major study themes or variables (a detailed list of variables on which data was collected may be annexed)
- the study/ target population(s), sampling method(s) and the size of the sample(s)
- data collection techniques used for the different study populations
- duration of data collection
- how the data was collected and by whom
- procedures used for data analysis, including statistical tests (if applicable)
- any constraints and its management
- limitations of the study.

If you have deviated from the original study design presented in your research proposal, you should explain to what extent you did so, and why. The consequences of this deviation for meeting certain objectives of your study should be indicated. If the quality of some of the data is weak, resulting in possible biases, this should be described as well under the heading 'limitations of the study'. An example of methodology is given in the box below.

Methodology

Data Collection/Sample

A micro level study based on primary cross section data was designed to attain the objectives of this project. The survey was conducted in three states; Bihar, Uttar Pradesh and Punjab. A systematic interview schedule was used to collect information on various aspects of labour migration and its impact on rural economy of Indo-Gangetic Plains of India. The data was collected for 200 families with migration and 200 families without migrating members.

Analytical tools

Various statistical tools were used in the analysis of data. Those are mean, standard deviation, correlation, t-test, and regression.

5.5.4 Chapter 4: Research Findings

A detailed presentation of the findings of the study with supporting data in the form of tables and charts, together with a validation of the results is the next step in writing the main text of the report. The result section of the study should contain the statistical summaries and reductions of data, rather than raw data. All the results should be presented in a logical sequence and split into readily identifiable sections.

The systematic presentation of your findings in relation to the research objectives is the crucial part of your report.

The list of data by objectives will help you to decide how to organise the presentation of data. The decision concerning where to put what can best be

made after all data have been fully processed and analysed, and before the writing starts.

When all data have been analysed, a detailed **outline** has to be made for the presentation of the findings. This will help the decision-making on how to organise the data, and is *an absolute precondition for optimal division of tasks among group members in the writing process*.

At this stage you might as well prepare an outline for the whole report, taking the main components of a research report as a point of departure.

An **outline** should contain

- the headings of the main sections of the report
- the headings of subsections
- the points to be made in each section
- the list of tables, figures and/or quotes to illustrate each section.

The outline for the chapter on findings will predictably be the most elaborate.

The first section under findings is usually a description of the study/ target population. When different study populations have been studied, you should provide a short description of each group before you present the data pertaining to these informants.

Then, depending on the study design, you may provide more information on the problem you studied (size, distribution, characteristics). Thereafter, in an analytic study, the degree to which different independent variables influence the problem will be discussed.

For better understanding, an example of how the research findings are tabulated and presented in the form of findings is given in the following table. An analysis of table 5.1 is given in the box below.

Table 5.1: Social Characteristics of Migrants (Percentage)

Particulars	Bihar	UP	Overall
Number	245	308	553
i) Age Profile			
Up to 30 Years	69.80	56.49	62.39
31 to 45 Years	26.53	35.39	31.46
Above 46 Years	3.68	8.11	6.15
ii) Literacy Status			
Illiterate	33.88	19.16	25.50
Primary	50.20	29.87	38.00
Matriculation and above	15.92	50.97	36.48
iii) Social Status			
Upper Caste	22.86	9.42	15.37
SC/ST/BC	77.15	90.58	84.63

Analysis of Table 5.1

The socio-economic characteristics of the migrants are depicted in Table 3. The table clearly shows that in UP there had been 308 migrants from 200 households while Bihar had only 245 migrants from 200 households. On an average, 62 percent of the migrants were below 30 years of age with a higher percentage of younger migrants from Bihar than from UP. Most of the migrants from both UP and Bihar were literate, and only 25 per cent of the total migrants from both UP and Bihar were illiterate. Most of the migrants belonged to a schedule caste or backward class, the percentage being higher in UP (91%) compared with Bihar (77%).

Tables and Figures in the text should be numbered and have clear titles. It is advisable to first use the number of the section to which the table belongs. In the final draft you may decide to number tables and figures in sequence. It is appreciated in case some pictures from the field are also appropriately presented to give visual presentation of the field information.

Include only those tables and figures that present main findings and need more elaborate discussion in the text. Others may be put in annexes, or, if they don't reveal interesting points, be omitted.

It is advisable to involve statistician/data analyst from the very beginning and in each process of the research so that he may provide meaningful tables and himself judge irrelevant findings.

Note: It is unnecessary to describe in detail a table that you include in the report. Only present the main conclusions.

The first draft of your findings is never final. Therefore, you might concentrate primarily on content rather than on style. Nevertheless, it is advisable to structure the text from the beginning in paragraphs and to attempt to phrase each sentence clearly and precisely.

5.5.5 Chapter 5: Discussion

The findings can now be discussed by objective or by cluster of related variables or themes, which should lead to conclusions and possible recommendations. The discussion may also include findings from other related studies that support or contradict your own. For easy understanding, the discussion of the table given in findings is given in the box below.

The socio-economic characteristics of the migrants are depicted in Table 3. The number of migrants gives an idea about frequency of migrants in a household. In UP, the percentage of households having more than one migrant was relatively higher when compared with that of Bihar. Most of the migrants in both the states were up to 30 years of age. This clearly indicates that young men in their productive age were more involved in migration. Similar results were reported by Sidhu et.al. (1997) and Kumar et.al. (1998) in their studies, that most of the migrants of both the states were literate and belong to the backward sections of the society. The underlying fact is that the backward classes belonging to the lower social hierarchy were more capable of doing menial jobs and tasks, which required lot of energy.

5.5.6 Chapter 6: Conclusions and Recommendations

The conclusions and recommendations should follow logically from the discussion of the findings. Conclusions can be short, as they have already been elaborately discussed in Chapter 5. As the discussion will follow the sequence in which the findings have been presented (which in turn depends on your objectives) the conclusions should logically follow the same order. Sometimes, it is advisable to present conclusion and recommendations in specific sections related to issues of importance/under investigation/objectives of the study for better clarity to different stake holders. The conclusions should be given in bullets so that it can easily catch the attention of the reader.

Remember that action-oriented groups are most interested in this section.

The conclusions should be followed by suggestions or recommendations. While making recommendations, use not only the findings of your study, but also supportive information from other sources. The recommendations should be generated from the findings and conclusions. It should not be generalized; rather it should be specific to particular stake holders in pure, actionable term which is feasible in relation to social context, policy and constitution of country, political acceptability, budget, time, etc. One should **not give general recommendations** such as, “Government should provide free treatment to everyone for all health problems”.

If your recommendations are short (roughly one page), you might include them all in your summary and omit them as a separate section in Chapter 6 in order to avoid repetition.

5.5.7 Chapter 7:References

This is the list of books/articles in some way pertinent to the research which was followed while conducting research. It should contain all those works which the researcher has consulted. The references in your text can be numbered in the sequence in which they appear in the report and then listed in this order in the list of references (Vancouver system). Another possibility is the Harvard system of listing in brackets the author’s name(s) in the text, followed by the date of the publication and page number, for example: (Sharma et. Al., 2000: 84). In the list of references, the publications are then arranged in alphabetical order by the principal author’s last name. You can choose either system as long as you use it consistently throughout the report unless some guidelines specifically ask for it (in case of research publications).

The references should be given in the following order.

- 1) Name of the author, last name first.
- 2) Title, underlined to indicate italics.
- 3) Place, publisher and date of publication.
- 4) Number of volumes.

Example

Kothari, C.R., *Quantitative Techniques*, New Delhi, Vikas Publishing House Pvt. Ltd., 1978.

For magazines and newspapers the order may be as follows.

- 1) Name of the author, last name first.
- 2) Title of the article, in quotation marks.
- 3) Name of the periodical, underlined to indicate italics.
- 4) The volume and number.
- 5) The date of the issue.
- 6) The pagination.

Example

Robert V. Roosa, "Coping with Short-term International Money Flows", *The Banker*, London, September, 1971, p.995.

5.5.8 Annexure

The annexes should contain any additional information needed to enable professionals to follow your research procedures and data analysis.

Information that would be useful to special categories of readers but is not of interest to the average reader can be included in annexes as well.

Examples of information that can be presented in annexes are

- tables, figures (graphs) and pictures referred to in the text but not included in order to keep the report short
- lists of hospitals, districts, villages, etc., that participated in the study
- questionnaires or check lists used for data collection
- A list of research team members.

Note: Never start writing without an outline. Make sure that all sections carry the headings and numbers consistent with the outline before they are word-processed. Have the outline visible on the wall so that everyone will be aware immediately of any additions or changes, and of progress made.

5.6 STYLE AND LAYOUT OF THE REPORT

The style of writing and layout of writing are two important components of report writing. However, revising and finalizing the text may also be considered as another important aspect in same line.

5.6.1 Style of Writing

Remember that your reader

- is short of time
- has many other urgent matters demanding his or her interest and attention
- is probably not knowledgeable concerning 'research jargon'.

Therefore, the rules are

- simplify- Keep to the essentials
- justify- Make no statement that is not based on facts and data

- do not quote the name of anyone who has provided the information
- in case of sensitive findings, one should think not to clearly mention name of village/ location, etc.
- quantify when you have the data to do so; avoid 'large', 'small' - instead, say '50%', 'one in three'
- the percentage 45.8 in table may be presented in the text as about 46% and 45.3% may be presented as approximately 45%
- be precise and specific in your phrasing of findings
- inform, not impress - avoid exaggeration
- use short sentences
- use adverbs and adjectives infrequently
- be consistent in the use of past and present tenses
- avoid the passive voice, if possible, as it creates vagueness (e.g., 'patients were interviewed' leaves uncertainty as to who interviewed them) and repeated use makes dull reading
- aim to be logical and systematic in your presentation.

5.6.2 Layout of the Report

A good *physical layout* is important, as it will help your report

- make a good initial impression
- encourage the readers
- give them an idea of how the material has been organised so the reader can make a quick determination of what he will read first.

Particular *attention* should be paid to make sure there is

- an attractive layout for the title page and a clear table of contents
- consistency in margins and spacing
- consistency in headings and subheadings, e.g.: **Font size 16 or 18 bold**, for headings of chapters; **size 14 bold** for headings of major sections; **size 12 bold**, for headings of sub-sections, etc.
- good quality printing and photocopying
- correct drafts carefully with spell check as well as critical reading for clarity by other team-members, your facilitator and, if possible, outsiders
- numbering of figures and tables, provision of clear titles for tables, and clear headings for columns and rows, etc.
- accuracy and consistency in quotations and references.

5.6.3 Revising and Finalising the Text

Prepare a double-spaced first draft of your report with wide margins so that you can easily make comments and corrections in the text. Have several copies made of the first draft, so you will have one or more copies to work on, and one copy on which to insert the final changes for revision. When a first draft of the findings, discussion, and conclusions has been completed, all working group members and facilitators should read it critically and make comments.

The following questions should be kept in mind when reading the draft.

- Have all important findings been included?
- Do the conclusions follow logically from the findings? If some of the findings contradict each other, has this been discussed and explained, if possible? Have weaknesses in the methodology, if any, been revealed?
- Are there any overlaps in the draft that have to be removed?
- Is it possible to condense the content? In general a text improves by shortening. Some parts less relevant for action may be included in annexes. Check if descriptive paragraphs may be shortened and *introduced or finished by a concluding sentence*.
- Do data in the text agree with data in the tables? Are all tables consistent (with the same number of informants per variable), are they numbered in sequence, and do they have clear titles and headings?
- Is the sequence of paragraphs and subsections logical and coherent? Is there a smooth connection between successive paragraphs and sections? Is the phrasing of findings and conclusions precise and clear?

The original authors of each section may prepare a second draft, taking into consideration all comments that have been made. However, you might consider the appointment of two editors amongst yourselves, to draft the complete version. The help from proof readers may also be taken to remove minor mistakes from the draft.

It is advisable to have one of the other groups and facilitators read the second draft and judge it on the points mentioned in the previous section. Then a final version of the report should be prepared. This time you should give extra care to the presentation and layout: structure, style and consistency of spelling (use spell check!).

Use verb tenses consistently. Descriptions of the field situation may be stated in the past tense (e.g., 'Five households owned less than one acre of land.'). Conclusions drawn from the data are usually in the present tense (e.g., 'Food taboos hardly have any impact on the nutritional status of young children.')

Note: For a final check on readability you might skim through the pages and read the first sentences of each paragraph. If this gives you a clear impression of the organisation and results of your study, you may conclude that you did the best you could.

5.7 COMMON WEAKNESSES IN REPORT WRITING AND FINALIZING THE TEXT

It is important to know the general mistakes committed in report writing and also the points to consider while finalising the text.

- i) **Endless Description** without interpretation is a pitfall. Tables need conclusions, not detailed presentation of all numbers or percentages in the cells which readers can see for themselves. The chapter discussion, in

particular, needs comparison of data, highlighting of unexpected results, your own or others' opinions on problems discovered, weighing of pro's and con's of possible solutions. Yet, too often the discussion is merely a dry summary of findings.

- ii) **Neglect of Qualitative Data** is also quite common. Quotes of informants as illustration of your findings and conclusions make your report lively. They also have scientific value in allowing the reader to draw his/her own conclusions from the data you present. (Assuming you are not biased in your presentation!). Presentation of important photographs also makes report attractive and explains facts better.
- iii) Sometimes qualitative data (e.g., open opinion questions) are just coded and counted like quantitative data, **without interpretation**, whereas they may provide interesting illustrations of reasons for the behaviour of informants or of their attitudes. This is serious maltreatment of data that needs correction.

In these sections you have read about the main text and end matter of the research report. You have also read about the style and layout of the research report. The general mistakes committed while writing a research report and the method of finalizing the text have also been given. Now, answer the questions that follow in Check Your Progress 2.

Check Your Progress 2

Note: a) Write your answer in about 50 words.

b) Check your answer with possible answers given at the end of the unit

- 1) What is the order of chaptering in a research report?

.....

.....

.....

.....

.....

.....

- 2) What information do annexes and appendices contain?

.....

.....

.....

.....

.....

.....

5.8 LET US SUM UP

The last part of any research is writing the research report. The report writing is an art as well as science. You have to identify who will be reading your report and the report should be prepared accordingly. A summary of report in the beginning is important. The report layout plan should be comprehensive and all aspects of report including realistic recommendations and future directions of research should be described.

5.9 REFERENCES AND SELECTED READINGS

Ackoff, R. L. (1961), *The Design of Social Science Research*, Chicago: University of Chicago Press.

Bailey, K. D. (1978), *Methods of Social Research*, New York.

Berelson, B. (1952), *Content Analysis in Communication Research*, Free Press, New York.

Berenson, B. and C. Raymond (1971), *Research and Report Writing for Business and Economics*, Random House, New York.

Kothari, C.R. (1978), *Quantitative Techniques*, New Delhi, Vikas Publishing House Pvt. Ltd.

Gatner, E.S.M and C. Francesco (1956), *Research and Report Writing*, Barnes & Noble Inc., New York.

Gaum, C.G., H.F. Graves and L. Hoffman (1950), *Report Writing*, 3rd ed., Prentice Hall, New York.

Gopal, M.H. (1965), *Research Reporting in Social Sciences*, Karnatak University, Dharwad.

5.10 CHECK YOUR PROGRESS – POSSIBLE ANSWERS

Check Your Progress 1

- 1) What should be included on the cover page of a research report?

The cover page should contain the title, the names of the authors with their titles and positions, the institution that is publishing the report with its logo, and the month and year of publication. The title could consist of a challenging statement or question, followed by an informative subtitle covering the content of the study and indicating the area where the study was implemented.

- 2) What is the importance of writing an acknowledgement in a research report and where should it be placed?

It is good practice to thank those who supported you technically or financially in the design and implementation of your study. Also your employer who has allowed you to invest time in the study and the respondents may be

acknowledged. You should not forget to acknowledge the contribution of computer professionals, library staff, local officials, and community at large that provided the information. Acknowledgements are usually placed right after the title page or at the end of the report, before the references.

Check Your Progress 2

- 1) What is the order of chaptering in a research report?

In any research report, the general trend of chaptering is as follows:

Chapter 1 Introduction

Chapter 2 Objectives

Chapter 3 Methodology

Chapter 4 Research Findings

Chapter 5 Discussion

Chapter 6 Conclusions and Recommendations

- 2) What information do annexes and appendices contain?

The annexes should contain any additional information needed to enable professionals to follow your research procedures and data analysis.

Information that would be useful to special categories of readers but is not of interest to the average reader can be included in annexes as well.