

BLOCK

4

DATA COLLECTION AND ANALYSIS

Unit 13

Tools and Techniques for Data Collection	241
---	------------

Unit 14

Sources of Data and its Analysis and Interpretation	260
--	------------

Unit 15

Organization and Graphical Representation of Data	279
--	------------

Unit 16

Analysis of Quantitative Data-I	302
--	------------

Unit 17

Analysis of Quantitative Data-II	334
---	------------

BLOCK 4: DATA COLLECTION AND ANALYSIS

Block Introduction

This Block consists of five units on collection, source, presentation, analysis and interpretation of data. The block begins with the Unit **“Tools and Techniques for Data Collection”** (unit-13) which discusses concept of research tools and techniques; characteristics of a good research tool, types of research tools and techniques like- test, questionnaire, interview, observation, scale, inventory, checklist etc.

The second unit of this Block, titled, **‘Sources of Data and its Analysis and Interpretation’** (Unit-14) deals with the sources and types of data (Primary and Secondary, quantitative and qualitative data). The Unit discusses some common techniques to be used for analysis and interpretation of quantitative and qualitative. The Unit also discusses various scales of measurement and its use in data analysis.

The third Unit of this Block, titled, **‘Organization and Graphical Representation of Data’** (Unit-15) talks about nature of data and its presentation. The Unit elaborates on organization of data in frequency distribution tables. Further, the Unit explains various methods of graphical representation of data like histogram, pie charts, frequency polygon, cumulative percentage curve/ogive.

Fourth Unit of this Block, titled, **‘Analysis of Quantitative Data-I’** (Unit-16) elaborately discusses the techniques of measuring the central tendency, its interpretation and use in interpretation of data. The major statistics used for calculating central tendency such as Mean, Median and Mode have been discussed in this Unit with suitable examples. The Unit also further discusses the concept, calculation, interpretation and uses of various statistical techniques of measuring dispersion such as Range, Quartile Deviation, Average Deviation and Standard Deviation. The Unit also further discusses the calculation and uses of Percentile and Percentile Ranks.

The last Unit of this Block, **‘Analysis of Quantitative Data-II’** (Unit-17) focuses on use of correlation and Normal probability curve in data analysis. So, Unit starts with the concept of correlation between two sets of scores, its degrees of relationships, coefficient index, techniques/methods of calculation, its interpretation and uses. The particular methods of correlation such as Rank Difference Method and Product Movement Method have been discussed. In this Unit, in the second part of the unit the concept of normal distribution and characteristics of normal probability curve have been discussed. Concept of Skewness and Kurtosis is also introduced in this Unit.

UNIT 13 TOOLS AND TECHNIQUES FOR DATA COLLECTION

Structure

- 13.1 Introduction
- 13.2 Objectives
- 13.3 Research Tools and Techniques: Meaning and Concept
- 13.4 Characteristics of a Good Research Tool
- 13.5 Tools for Data Collection
 - 13.5.1 Test
 - 13.5.2 Questionnaire
 - 13.5.3 Rating Scale
 - 13.5.4 Inventory
 - 13.5.5 Checklist
- 13.6 Techniques for Data Collection
 - 13.6.1 Interview
 - 13.6.2 Observation
- 13.7 Let Us Sum Up
- 13.8 References and Suggested Readings
- 15.9 Answers to Check Your Progress

13.1 INTRODUCTION

Research is an original contribution to the existing knowledge that helps in advancement in the field of education. For conducting research, relevant data can be gathered by using appropriate tools and techniques. Tools and techniques in research can be chosen based on the nature of research which a researcher undertakes. In order to get relevant and appropriate data for an educational research, a researcher can select appropriate tool(s) such as, psychological tests and inventories, questionnaires, checklists, rating scales etc. and techniques, namely, observation, interview, sociogram, content analysis etc. Each tool and technique differ in its nature, characteristics, design, administration and interpretation. Also, each tool is specifically designed for collection of information related to specific type or nature of research. As there are many tools available to assess a specific variable, a researcher has to select one tool to assess a variable under his/her study. If in case of non-availability of tool, a researcher may modify or develop his/her own tool. A good research tool has some basic characteristics and every researcher needs to be aware of them. This unit will familiarize you with some data collection devices.

13.2 OBJECTIVES

After studying this unit, you will be able to:

- explain the meaning of tool and technique of research;

- describe characteristics of good research tools;
- discuss different types of tools used for data collection and
- discuss different types of techniques used for data collection.

13.3 RESEARCH TOOLS AND TECHNIQUES: MEANING AND CONCEPT

Data collection is one of the most important steps of research and data is collected with the help of a research tool(s). Research tools are instruments used to support in the assessment of a particular variable which is under study. So, we can say that, anything that is used to collect information related to a variable under study is called a research tool or a research instrument. The research instruments or research tools are used to measure or collect data for a variety of variables extending from physical working to psycho-social health. Research tools are measurable and observable for the purpose of data analysis and interpretation. Different types of tools are used to gather data, depending upon the nature of the or data. Some of these tools only identify the absence or presence of certain aspects of a situation. Few other tools gather information related to qualitative descriptions which comprises of contrast and comparisons between or among the elements that exist in that situation. Some other tools are used to collect data related to quantitative measures which are in scores or in scale measure.

When a researcher plans to collect data, either he/she develops a research tool according to the objectives framed for the study, or selects an already developed standardized tool.

Normally data can be collected through testing and non-testing procedures. Under testing procedure, the data is obtained through a test or a questionnaire or an inventory or a check list or a rating scale, these types of procedures/instruments are called **research tools**. Under non testing, the data are collected through observation or interview and theses types of procedures are called **research techniques**. Different types of tools are used for different data collection techniques. For example, in observation technique one can use check list, rating scale, questionnaire etc. as research tools.

Research tools refers to the tools or means with the help of which a researcher assesses different variables under study in the data collection process. It is related not only to selection, designing, construction and assessment of tool but also to the conditions wherein the tool is administered. Research tools are very important and are needed to gather data from sample for further analysis and its interpretation of results to draw findings.

Check Your Progress 1

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

1. Discuss the meaning of research tools.

.....

.....

.....

.....

2. Do you think research tools are important? Justify your answer.

.....

13.4 CHARACTERISTICS OF A GOOD RESEARCH TOOL

The characteristics of a research tool are:

1. **Objectivity:** A test is called objective when examiner's personal judgement does not affect the scoring. Objective judgements are accurate and hence tend to be reliable. It can be seen from two aspects:
 - **Objectivity of items:** It implies that the items should be as simple as possible away from the vague words like perhaps, sometimes, too much etc.
 - **Objectivity of scoring:** It means that the personal biasness of researcher should not affect scores. Emotional conditions, prejudices and attitude of researcher should not affect scores.
2. **Reliability:** Reliability refers to consistency in measurement. It is basically concerned with the consistency of responses on the test given to same individuals at different occasions. This means that even if an individual takes the same test twice, the scores are the same. To determine the reliability of a test, different measure of correlations such as Test-retest, Split half, Parallel forms etc. are used.
3. **Validity:** A good research tool is expected to measure what it is supposed to measure. If a test achieves this objective, then the test is called valid. Validity is basic, we may consider validity in terms of categories such as high, moderate or low validity. Furthermore, validity is specific, a test may be valid for one purpose but not valid for another. A data collection tool must be a valid one.
4. **Originality:** The tool should be original and non- plagiarised, free of errors and should include all the required details.
5. **Usability:** It is also known as practicability. The usability depends on the ease with which a tool can be administered, interpreted and scored without excessive expenditure of time, money and efforts.
6. **Sensitivity:** A tool should be sensitive enough to record/ assess/ measure even a small change. This will make the tool more accurate and the findings will be more precise or exact.

Check Your Progress 2

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

3. How do we know that a research tool is good?

.....
.....
.....
.....

4. Discuss two different aspects of objectivity.

.....
.....
.....
.....

5. Differentiate between reliability and validity.

.....
.....
.....
.....

13.5 TOOLS FOR DATA COLLECTION

To collect appropriate and reliable data we can use various types of tools. The choice of a tool will depend on the nature of variable and type of data, which a researcher wants to collect in a particular situation. Data can be quantitative or qualitative in nature, about which you are going to study in the next unit. Let us now discuss about some of the tools one-by-one:

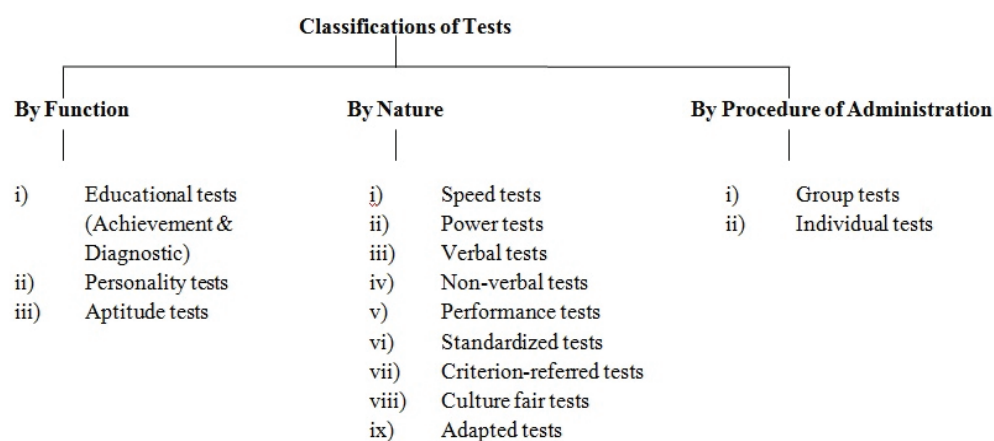
13.5.1 Test

A test is a systematic procedure for observing a person's behaviour and describing it with the help of a numerical scale. Test is helpful in getting a quantitative data.

It is essentially an objective and standardized measure of a sample of behaviour. Psychological tests which are available in research cells or departments are standardized tests. Teacher made test are non-standardized test.

Types of Tests

Tests are classified in different ways based on their function, nature and procedure of administration and are presented here under:



Let us now understand the meaning of this classification and will also get information about these tests in details.

A. Classification of the Tests as per their Function

(i) **Educational Tests** - Certain types of tests are developed for use in educational contexts. They include achievement tests and diagnostic tests. Achievement tests are designed to measure the effects of a specific programme of instruction or training. It measures the effects of relatively standardized sets of learning experiences. It generally represents a terminal evaluation of an individual's status on the completion of an instructional programme or training. An achievement test measures developed abilities/skills. Diagnostic tests on the other hand, are designed to analyze individual's specific strengths and weaknesses in a school subject and to suggest causes of his/her difficulties. In connection with the use of all diagnostic tests, one point deserves special emphasis. The diagnosis of learning disabilities is the prerequisite for a programme of remedial teaching.

(ii) **Personality Tests** - These tests most often refer to measures of affective aspects of behaviours such as emotional states, interpersonal relations, motivation, interests and attitudes. A personality test could be a self-report inventory, a situational test or another approach could be the application of projective technique.

(iii) **Aptitude Tests** - These measure the effects of learning under relatively (in comparison with achievement tests) uncontrolled and unknown conditions. They include general intelligence tests, multiple aptitude batteries, special aptitude tests (e.g. creativity tests) etc.

Intelligence tests are designed to measure the intelligence of an individual. Here, we must bear in mind that intelligence is not a single, unitary ability but a composite of several abilities. The term is commonly used to cover a combination of abilities required for survival and advancement within a peculiar culture. (e.g. Stanford Binet Test of Intelligence)

Multiple Aptitude Batteries yield a profile of test scores in set of relatively independent abilities identified through factor-analysis. One of the most widely used Multiple Aptitude Batteries is a Differential Aptitude Test (DAT). This DAT yields scores on eight abilities - Verbal Reasoning, Numerical Ability, Abstract Reasoning, Clerical Speed and Accuracy, Mechanical Reasoning, Space Relations, Spelling and Language Usage.

Creative Aptitude Batteries have been produced at the University of Southern California by J.P. Guilford & his colleagues, and 'Tests of Creative Thinking' by Paul Torrance.

B. Classification of the Tests as per their Nature

- (i) **Speed Tests** are those tests which contain such items which are of somewhat equal difficulty value. To solve the tests the time limits are fixed. e.g. our usual school/college tests.
- (ii) **Power tests** are those tests which contain items of different difficulty value arranged in increasing order of difficulty value. For such tests time limit is not fixed, it is flexible.
- (iii) **Verbal Tests** include question answer pattern in the form of 'words'. Such verbal tests can be either paper pencil (written) tests or oral tests.
- (iv) **Non-verbal Tests** necessarily contain items in the figural or symbolic form i.e. other than 'words'. Raven's Progressive Matrices Test is a non-verbal test.
- (v) **Performance Tests** require action on part of the respondent as his/her response. Bhatia's Intelligence test is a performance test.
- (vi) **Standardized Tests** imply uniformity of procedure in administering and scoring the tests. Another important aspect of standardized tests is the establishment of norms, reliability and the validity of the tests.
- (vii) **Criterion Referenced Tests** have been contrasted with the usual norm-referenced tests in which an individual's score is interpreted by comparing it with the scores obtained by others in the group on the same test. Unlike in a norm-referenced test, in a criterion referenced test an individual's score is interpreted by comparing it with an external criterion.

C. Classification of Tests as per the Procedure of Administration

- (i) **Group Tests** are designed as mass testing instruments. They permit the simultaneous testing/examination of large groups. The instruction and administration procedures are simplified so as to demand a minimum of training on the part of the test administrator/examiner
- (ii) **Individual Tests** are those which can be administered to one individual at a time. In view of the testing situation including response in the form of actual performance, the observer can record the entries in the performance sheet for one candidate at a time

Source : 13.5.1 is adopted from 'Tools and Techniques for Data Collection (Unit-13). MES-054- Methodology of Educational Research'. IGNOU (2006). New Delhi

13.5.2 Questionnaire

Questionnaires generally include a set of standardized items (question/statement form) that are based on a particular topic designed to collect data related to demographics, attitudes, behaviours or opinions. It is a kind of tool wherein factual data is collected from a very large sample. The main advantage of using a questionnaire is that it can be made with minimum cost and time and can

reach a widely scattered audience. With the use of technology (Google forms, Survey monkey) , it is now only a single click . A questionnaire consists of several types of questions. Let's discuss them:

Types of Questions in a Questionnaire

1. **Open ended questions:** There is no pre-determined set of answers because it calls for free response from the respondent in a space provided. No clues or preset answers are provided, hence the respondent has greater freedom to respond. The responses to these questions are difficult to analyze. It provides true and unexpected suggestions.

For example:

- Why did you choose a particular university/ college to pursue your graduation?
- What happened in mathematics class?

2. **Close ended questions:** There is a predetermined set of answers, because it provides the alternatives against each question have statement. The responses are limited in number. The responses are easy to analyze. Generally a questionnaire alternative response type questions (i.e. select Yes/No) and multiple choice type items (select right answer in the given choices)

Examples of Alternative Type Questions:

- (i) Are you happy with the faculty/teachers of this college? Yes () No
- (ii) Have you ever chosen a course in journalism? Yes () No
- (iii) Have you been honoured for your achievements? Yes () No ()

Example of Multiple-choice Questions:

1. What does ZPD stand for?
 - a) Zone of Psychosocial Distance
 - b) Zone of Proximal Development
 - c) Zone of Psychosocial Development
 - d) Zone of Proximal Distance

Advantages: The following are the advantages of Questionnaire:

1. It is easy to administer.
2. It collects data from every corner of the world.
3. It is economical in terms of time and money.
4. Respondents can answer questions at any time convenient to them and they can answer at their own pace.

Limitations: The following are the limitations of Questionnaire:

1. Questions have to be comparatively easy.

2. Generally, not possible to give assistance directly, if required by the respondents.
3. Some items may be misunderstood.
4. Only those who are literate can be the part of sample.
5. Wastage of time, money and efforts because there is no control over who all complete and send back the filled in questionnaire.
6. Deadline for returning the filled questionnaire is required.
7. Reminders to the respondents may be required.
8. Printing may be costly if it is lengthy.

When you develop the questionnaire, it is essential that you must ensure the following:

1. The covering letter of the questionnaire is conscripted in a friendly manner and shows its importance and relevance to the respondents.
2. The title of the questionnaire must be well stated.
3. It should consist of statement of purpose.
4. Instructions/ directions are complete and clearly mentioned.
5. The respondents should be given assurance about the confidentiality of their responses.
6. It is neatly arranged, looks attractive and is free from typographical errors.
7. Technical terms are clarified. Also, the clarity in language must meet the level of understanding of the respondents.
8. It avoids questions which reflects annoying or embarrassing tone which may arouse aggression in the respondents.
9. Items are arranged in classes/ groups which elicits easy and accurate responses.
10. The questions are pin pointed and objective without any clues or suggestions.
11. Statements/ questions are in order from easy to difficult.
12. It does not try to find the information which may be available from other sources such as records and registers maintained in school/ institution.
13. Restricted to single question which deals with single aspect.
14. Double negatives descriptive adjectives and adverbs are avoided
15. It is worded in such a manner that it can be easily tabulated, interpreted and summarized

Check Your Progress 3

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

6. What is the meaning of tests?

.....

7. On what basis are tests classified?

.....

8. What is the difference between speed test and power test.

.....

9. What points will you keep in mind while developing a questionnaire?

.....

13.5.3 Rating Scale

This is one of the oldest method/tool of personality assessment. Rating is generally applied to expression of judgement regarding some event, character or an object. The opinion or judgement is usually expressed on a scale. Rating is a technique with the help of which such opinions/judgements can be quantified. Rating can be done by parents, teachers, interviewers, judges etc. It resembles checklists but are used when finer discriminations are needed. These scales provide an indication of the amount of a specific attribute and use either number or description. For example:

How satisfied are you with the relaxation therapy used during your stress reduction programme?

Extremely dissatisfied

Extremely satisfied

0 1 2 3 4 5 6 7 8 9 10

Characteristics of a Rating Scale

1. These reflect valuable opinion about characteristics of one person by another person.

2. Make quantitative opinions about qualitative characteristics.
3. This is most commonly used tool to perform structured observations.
4. It provides flexibility to judge the presence of different characteristics among subjects.

Types of Rating Scale

1 Numerical Rating Scale: Numerical rating scale is one of the simplest type of rating scale. The rater simply marks a numeric figure which shows the degree to which a trait is present. The statement consists of traits and values from 1-5 are given to each trait that is rated.

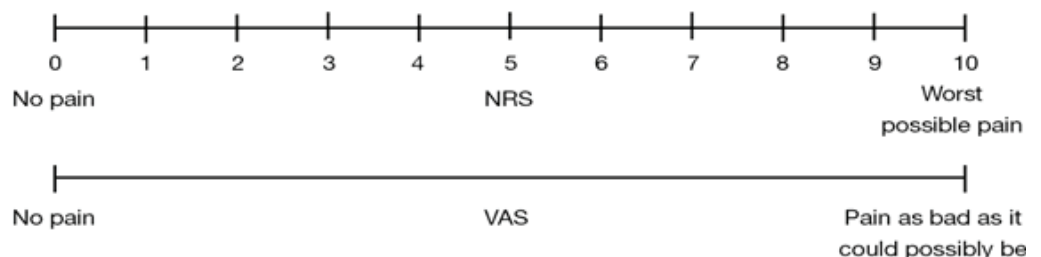
Example:

Tick the correct number showing the extent to which student-teacher exhibits skill of stimulus variation.

1. Unsatisfactory
2. Below average
3. Average
4. Above Average
5. Outstanding

2. Graphic Rating Scale /Visual Analog Scale

This scale is useful for assessing perception of physical stimuli such as shortness of breath, pain, sleep, etc. It is generally used in hospitals by doctors to conduct pain scale survey. Instead of using scale values, the ratings are made in a graphic form- a point anywhere along a band. Each attribute is followed by a horizontal line. The rating is made by placing a tick mark on the line. A set of classes identifies specific positions along the line, but the rater has a freedom to check in between these points.



Choose an option to describe your pain:

				
☐	☐	☐	☐	☐
Unbearable Pain	Intense Pain	Nagging Pain	Mild, Annoying Pain	No Pain

Source: <https://www.questionpro.com/visual-analogue-scale-vas.html>

Advantages

1. Rating scale is flexible and adaptable.
2. Can be used with a large sample size.
3. It assesses specific objectives of education which are important to teachers.
4. Evaluate processes such as demonstrating a skill, performing experiment in laboratory, teaching learning process etc.
5. It records qualitative and quantitative observations about observed performances.

Limitations

- Common sources of errors such as ambiguity, attitude of researcher, logical errors etc. affects the validity.
- Many aspects of an individual may not be rated.
- Scale may become unscientific due to chances of subjectivity.
- Generosity error -The raters would not like to run down their own people by giving them low ratings. The result is that high ratings are given in almost all cases.

13.5.4 Inventory

Inventories are self-reporting assessment tools that examine one or more characteristics of an individual's behaviour. It can be of many types such as interest inventory, personality inventory etc.

An inventory is like a questionnaire. In the inventory, the statements are put in first person, for example, "I think I am comparatively more sincere than others". In the questionnaire, there is a question in a second person, for example "Do you think you are more sincere than other persons?"

Advantages

1. Used in educational and vocational guidance.
2. Useful for teachers to find out the interest of learners/students.
3. Inventories are comparatively economic.
4. They are simple to construct.
5. Objectivity can be maintained.
6. Not much training is required to administer it.

Limitations

1. These instruments may be of limited value because of individuals' inability or unwillingness to report their own reactions accurately or objectively.
2. It is difficult to standardize inventories, which limits its usage.

13.5.5 Checklist

A checklist comparatively simple, inexpensive and reliable method of evaluating

sample (s). It is used to record the presence or absence of the phenomena under study. The main purpose of the checklist is to call attention to various aspects of an object or situation, to see that nothing of importance is overlooked. It includes list of words, statements/ items which describes a person, object or a situation and requires same format of response for all items. It can be a simple list of aspects that can be marked as present or absent “YES” or “NO”.

They are more easily constructed in comparison to other tools and can be useful for evaluating personality, behaviour, skills, conditions, manifestations etc. We can also use checklist to know about the availability of resources in a school/ college like computers, printers, books, sports material, smart classrooms etc. Let's see an example of checklist to know about students behaviour in a class.

Table 13.1 : Checklist for Students Behaviour

S.No.	Situation	Yes	No
1.	Took permission to talk		
2.	Attitude was good		
3.	Was Active		
4.	Was kind with others		
5.	Was attentive		
6.	Followed instructions		
7.	Cooperated peers in a group activity		
8.	Completed work		
9.	Participated in activities		

Advantages

1. Checklist is a simple method to record observation of any behaviour
2. Checklist is used in evaluation of activities.
3. It can be used in comparing individuals.
4. The chances of error in observation decreases.

Limitations

1. The degree of accuracy cannot be assessed.
2. It has limited use, as it does not indicate quality of performance.
3. Only the presence or absence of the ability, attribute, behaviour or performance parameter may be assessed.

Activity for Practice

- a) Try to make a checklist for yourself before going to market/ before examination/ before going to a vacation.
- b) Try to make a checklist to know about availability of facilities in a school.

Check Your Progress 4

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

10. List characteristics of rating scale.

.....

.....

.....

.....

11. Write advantages and limitations of checklist.

.....

.....

.....

.....

12. In what ways inventory is different from questionnaire?

.....

.....

.....

.....

13.6 TECHNIQUES FOR DATA COLLECTION

We can collect the requisite data for the research in hand by employing some techniques. Following techniques, if needed such as interview, schedules and observation schedules may be prepared for the purpose of data collection. Let us get information about these techniques of data collection in details.

13.6.1 Interview

Interview involves a verbal face-to-face or telephonic communication between the researcher and the respondent wherein the information is provided to the researcher. The conversation carried out during the interview has a definite objective. Interview is the most common tool used to collect data in qualitative and descriptive studies. The researcher tries to collect factual data related to perception, attitude, opinion etc.

Characteristics of Interview

1. Researcher and subject(s) will be strangers.
2. It can be modified as per the demand of the situation.
3. The information from the subject is recorded.
4. Every interview has a specific purpose.
5. Verbal answers for verbal questions are expected and recorded.

Types of Interviews

1. Structured Interviews

The structured interviews are formal in nature because there is a restriction of sequence and wordings to be followed by the researcher. It has fixed and

closed ended questions. It is known as standardized interview. It increases the reliability and credibility of data. Also, the responses from the sample are recorded on a standardized schedule.

Advantages

1. It avoids irrelevant, objectiveless interaction.
2. Recording, coding and analysis of data is easy.
3. Data gathered from two different samples can be compared.
4. Attention is not diverted.

Limitations

1. The scope of exploration is limited.
2. It may not cover all the views or responses of the sample.
3. In depth information may not be possible.

2. Unstructured Interviews

In unstructured interviews, subjects are motivated to speak freely and initially very few questions are asked for rapport building. The researcher also freely asks the questions as per researchers wish or as and when required. The unstructured interview is informal in nature wherein sets of questions can be used, but there is no restriction of following the sequence and wordings. Also, researcher is at times free to explain or substantiate the question during the interview. Therefore, it is characterized by flexibility of approach to questioning.

Advantages

1. Data can be used in explorative and qualitative studies.
2. Less prone to researcher's biases.
3. Information obtained in an informal manner enhances reliability and credibility of data.

Limitations

1. Analysis is relatively challenging.
2. Information of two candidates cannot be compared.
3. Time consuming
4. Researcher needs to be skilled enough to analyse data.
5. Interpretation is based on researcher's subjective feelings.

How to Conduct an Interview**

As a technique of data collection a good interview requires

- i) **Preparing for interview** - The following factors need to be determined in advance of the actual interview. Clear conception of the purpose and of what information is needed. i.e. kind of interview - individual or group. A clear outline or schedule of the best sequence of questions.
- ii) **Conducting an interview (execution)** - The initial task of securing

confidence and co-operation of the subjects of respondents what is called - rapport building. It requires expertise and sensitivity that is almost an art. For securing desired information, questions should be asked in planned sequence and not in a rigid manner. Moreover, a questions need to be asked in a stimulating and encouraging manner with comments and necessary explanations and recordings. The recording device selected should be used without distracting the interviewee. Before an interview-based study is undertaken, the interviewer should undergo rigorous training in conducting and recording of data in an interview.

- iii) **Recording and interpreting responses** - It is best to employ, if possible, a recording device which would retain the actual wording of the responses. Audio recording can be done through smart phones. If the responses to questions in the interview have to be noted down, it can be done either simultaneously with the interview or immediately after it.

****Source :** Some parts of 13.6.1 is adopted from 'Tools and Techniques for Data Collection (Unit-13)'. MES-054 : 'Methodology of Educational Research'. IGNOU (2006). New Delhi

13.6.2 Observation

The observation is a technique of collection of comprehensive information in qualitative research through events that can be observed through senses, with or without any mechanical devices. The subject can be observed with the help of following:

- **Log (field diary):** A log is a day-to-day record of interactions and events in the field. It is a historical listing of how researchers have spent their time and can be further used for planning purposes. This is for:
 - * reviewing what all work has been completed.
 - * keeping track of responses.
- **Anecdotes:** Anecdotes opts for definite types of behaviours for observation beforehand. They generally focus on behaviour of particular interest. The researcher records the information in an objective and accurate manner.
- **Video Recording:** Video recordings are done by mechanical devices commonly known as video tapes. Presently, it can be done through smart phone as well. These are used to record complex behaviours that might get unnoticed by the researcher.

Types of Observation

1. **Participant Observation:** In this type of observation the researcher participates as one of the members of the group, where subject of the study can be directly observed. This helps the researcher to take direct note of all the characteristics of the subjects who are being studied.
2. **Non-Participant Observation:** In this type of observation the researcher does not participates as one of the members of the group where subject of the study is present. The researcher observes the subject of the study with his/her understanding and experience, but without taking part in the event under study.

Stages in the Process of Observation: As a good data collection technique observation needs –

- a) **Planning:** The factors one must consider in detail while planning to use observation as a research technique includes-
- Definition of specific activities.
 - An appropriate group of subjects or individuals to observe.
 - Scope of observation - individual or group.
 - Determination of length of each observation period.
 - Deciding about the instruments and the form of recording and physical position of the observer.
 - Determining the special conditions required for subjects.
 - Preparation of proper tools for recording observation.
 - Getting oneself trained in terms of expertise of an observer.
- b) **Execution of Observation :** An expert execution demands -
- Proper arrangement of special conditions for the subjects.
 - Assuming the proper physical position for observation.
 - Focusing attention on the specific activities or units of behaviour under observation.
 - Observing directly, the length and number of periods and intervals decided upon.
 - Handling well the recording instruments to be used.
 - Utilizing the training received in terms of expertness.
- c) **Recording and Interpreting Observations :** The two common procedures for recording observations are –
- **Simultaneous** - when the observer goes on recording his/her observations simultaneously with the occurrence of the phenomena observed, as in time sampling.
 - **Soon after the observation** - when the observer undertakes to record his/her observations not simultaneously with his/her actual observation process, but immediately after he/she has observed for a unit of time, while the details are still fresh in his/her mind.

Advantages

1. Data can be collected where and when an event is taking place.
2. Natural responses are observed.
3. Validity of data is more.
4. Useful to sociological research.
5. Does not depend on sample's willingness to provide information.

Limitations

1. More prone to researcher's subjectivity or biases.
2. It is time consuming.
3. Analysis of data is difficult.
4. Researcher has to face practical difficulties.
5. The people being observed become conscious and begin to behave in an unnatural manner.

Check Your Progress 5

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

13. What are the different types of Interviews and Observation.

.....

.....

.....

14. Write the steps for conducting an interview.

.....

.....

.....

15. What are the limitations of observation technique?

.....

.....

.....

13.7 LET US SUM UP

In this unit you have learnt about various tools and techniques of data collection. You have also learnt that anything that is used to collect information related to a variable under study is called a research tool. The researcher selects the most appropriate tool for data collection which is further analysed to test the hypotheses. Some common and important tools and techniques of research which are regularly used in social science research are: Tests, Questionnaire, Interview, Observation, Rating scale, Inventory and Checklist. Some researchers use only one technique while some may use a number of them in combination. The choice of selecting a specific tool depends on the objectives, the variable under study, the type of information required, available resources, skill of using a particular tool and the socio and demographic attributes of the sample. All these research tools have some advantages or limitations as discussed in this unit. Questionnaires generally include a set of standardized items (question/statement form) that are based on a particular topic designed to collect data whereas interview

involves a verbal face to face or telephonic communication between the researcher and the respondent wherein the information is provided to the researcher. Observation is a useful technique of collecting authentic evidence in descriptive educational research.

Rating scale is a tool of research which is used to judge particular characteristics. A checklist is comparatively simple, inexpensive and a reliable method of evaluating sample (s). It consists of a list of response categories that is marked as checked if appropriate. Inventories are used in research while studying the inter-relationship between personality traits and variables like attitude, intelligence, achievement etc.

13.8 REFERENCES AND SUGGESTED READINGS

- Black, P. (2005). *Assessment for Learning: putting into practice (e-book)*, Mc Graw-Hill.
- Dweck, C. (2006). *Mind-set: The new psychology of success*. Random House LLC.
- Pathak, R. P. (2012). *Measurement and Evaluation in Education*. New Delhi: Pearson Publication.
- IGNOU. (2006). Tools and Techniques for Data Collection (Unit-13). MES-054: Methodology of Educational Research'. School of Education .New Delhi
- Suzanne (1995). *Assessment, Testing and Evaluation in Teacher Education*, USA: Ablex Publishing Corporation.
- Singh, A. K. (1997). *Tests, Measurements and Research Methods in Behavioural Sciences*. Patna: Bharti Bhawan.
- Shepard, L. A. (2000). *The role of assessment in a learning culture*. Educational Researcher, 4–14.
- Stiggins, R. (2005). *From formative assessment to assessment for learning: A path to success in standards-based schools*. Phi Delta Kappan, 324–328.
- Shaughnessy, J. J. et al. (2006). *Research Methods in Psychology*, Singapore: Mc.Graw-Hill.
- Sindhu, K. S. (2007). *New Approaches to Measurement and Evaluation*, Sterling Publication.
- Secolsky, C. (2011). *Handbook on Measurement and Evaluation in Higher Education*, Routledge.
- Thorndike, E. L. & Hagen, E.P. (1969). *Measurement and Evaluation in Psychology and Education*. New York: Johan Wiley and Sons Inc.
- William, D. (2011). *Assessment for Learning: why, what and how*, London: Institute of Education.
- Young, P. V. (1960), *Scientific Social Surveys and research*, prentice hall, New Delhi

13.9 ANSWERS TO CHECK YOUR PROGRESS

- 1) Anything that is used to collect information related to a variable under study is called a research tool or a research instrument.
- 2) Yes, research tools are important. Research tools are important because there are an important part of data collection. Without collection of data a researcher cannot proceed further.
- 3) A tool is said to be good if it consists of characteristics like objectivity, reliability, validity, originality, usability, sensitivity.
- 4) The two different aspects of objectivity are: Objectivity of items and Objectivity of scoring.
- 5) **Reliability:**
 - a. Reliability is consistency in measurement.
 - b. The reliability of a tool can be established with the help of: Test-retest reliability, Split half reliability, Parallel forms etc.

Validity:

- a. A tool is said to be valid when it measures what it is supposed to measure and perform as it is designed to perform.
 - b. Validity of a tool can be established with the help of content validity, concurrent validity, face validity etc.
- 6) A test is a systematic procedure for observing a person's behaviour and describing it with help of a numerical scale.
 - 7) By function , by nature and by the procedure of administration.
 - 8) Speed Tests are those tests which contain such items which are of somewhat equal difficulty value. To solve the tests the time limits are fixed. Power tests are those tests which contain items of different difficulty value arranged in increasing order of difficulty value. For such tests time limit is not fixed, it is flexible.
 - 9) Refer 13.5.2
 - 10) Refer 13.5.3
 - 11) Refer 13.5.5
 - 12) Refer 13.5.4
 - 13) Types of Observation: Participant and Non-Participant
Types of interviews: Structured and Unstructured
 - 14) As a technique of data collection good interview requires -
 - i) Preparing for Interview
 - ii) Conduct of Interview (execution)
 - iii) Recording and Interpreting Responses
 - 15) More prone to researcher's subjectivity. It is time consuming. Analysis of data is difficult. Researcher has to face practical difficulties.

UNIT 14 SOURCES OF DATA AND ITS ANALYSIS AND INTERPRETATION

Structure

- 14.1 Introduction
- 14.2 Objectives
- 14.3 Sources of Data
 - 14.3.1 Primary Data
 - 14.3.2 Secondary Data
- 14.4 Types of Data
 - 14.4.1 Quantitative Data
 - 14.4.2 Qualitative Data
- 14.5 Scales of Measurement
- 14.6 Techniques of Data Analysis and Interpretation
 - 14.6.1 Quantitative Data Analysis and Interpretation
 - 14.6.1.1 Statistical Measures for Quantitative Data Analysis
 - 14.6.2 Qualitative Data Analysis and Interpretation
- 14.7 Let Us Sum Up
- 14.8 Unit End Exercises
- 14.9 References and Suggested Readings
- 14.10 Answers to Check Your Progress

14.1 INTRODUCTION

In the previous units, you have been exposed to tools and techniques of data collection including test, questionnaire, interview, observation, scale, inventory and checklist, etc. Characteristics of a good research tool have also been discussed. In any research (fundamental or applied), after defining the research problem and finalizing the research design, data collection begins. For this, it is essential to know about data, its nature and sources. The present unit will deal about the sources and types of data. The difference between quantitative and qualitative data will be highlighted and common techniques to be used for its interpretation will also be explained.

14.2 OBJECTIVES

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

- identify various sources of data,
- define and differentiate primary and secondary data,
- differentiate between primary and secondary sources of data,
- compare and contrast the quantitative and qualitative data, and

- use appropriate techniques to interpret quantitative as well as qualitative data.

14.3 SOURCES OF DATA

The data are defined as numbers that represent measurements from the real world. Datum is a single measurement. We often read the news like 20 centimeters of continuous rain in Delhi or distance between one place to another in kilometers. This numerical information is called data. It may be easily realized that there is a large volume of data available around the world today. However, at times, it becomes difficult to derive logical conclusions from such data, if they are in raw form. Hence, it is important to ensure that the measured information is algorithmically derived and/or logically deduced and/or statistically calculated from multiple data.

When we undertake any research study, then the data may include facts, figures, information and other relevant materials that are considered important for that research study. Data are the important raw materials of research without which research study cannot proceed. In research, data provides a basis for the analysis which helps the researcher to reach conclusion or outcome of the research. The process of data collection begins after the research problem has been defined and the research design has been planned. Although there are many forms of data, every data has some common structures known as elements of analysis or units of analysis i.e., individuals, dimensions or variables which are studied.

There are two main types of data based on the sources of data. These are primary and secondary data.

14.3.1 Primary Data

Primary data are directly collected by the researcher, by visiting the field of study (population/sample) or interview or by experimentation. Thus, it is collected by the researcher directly from the sample/individual for the first time. Primary data are collected first hand and thus happen to be original. These data are also collected from the sample directly by the researcher using various tools and techniques.

Primary data are the data gathered at first hand, the responsibility of their compilation and promulgation, remaining under the same authority that originally gathered them. – Young (1973)

There are mainly two sources for collecting primary data:

- i) Individuals related to the problem/study and
- ii) Direct observations.

Types of Sources:

- a) **Direct primary sources-** Data collected directly by the researcher using, face-to-face interviews observations, schedule etc.
- b) **Indirect primary sources -** Data collected with the help of Questionnaires, telephonic interviews etc.

Some important tools and techniques used for collecting primary data are:

- Observation schedule

- Interview schedule
- Questionnaire
- Schedules
- Projective techniques
- Experiments
- Surveys etc.

Example

Mohit conducted a study on the impact of education on decision making ability of working women. For which, he collected the data through survey method using questionnaire. Here the data collected is a primary data as the data are collected directly from the sample.

Advantages of Primary data

- Primary data are original as they are collected directly by the researcher and therefore, liability, suitability and adequacy of primary data are ascertained.
- Primary data is at par with the time and place. The change in the information with respect to the time and place can easily be traced or noticed.
- The primary data can be used in different contexts and dimensions.
- Maximum research enquiry for the research study is based on primary data.
- The chances of subjectivity are minimum.
- The naturalness of the data is maintained.

On the other hand, concerning the time and cost-effectiveness of primary data is time-consuming and expensive. The collection of primary data requires a skilled person in research.

Precautions to be taken while Collecting Primary Data

- A researcher should avoid collecting useless or redundant data, for this, the researcher has to plan for data collection beforehand.
- The researcher should be objective while collecting data i.e., the personal biases of the researcher should not intervene with the process of data collection.
- Data should be collected systematically.
- Collected primary data and it should not be changed.
- While collecting data the sample or respondent should not feel as they are being inspected.

14.3.2 Secondary Data

Secondary data consists of both published and unpublished documents, letters, diaries, manuscripts, etc., which include government publications too. Such data are already available and has been collected and analyzed by someone else or used in some other research studies. It can also be in the form of findings of the researches that had already been conducted and to some extent is related to the present one.

Secondary data are the data that are compiled from the sources and of which the promulgating authority is different from that which controlled the collection of data first hand. - Young, (1973)

Major sources of secondary data are:

- Personal documents
- Public documents.

In secondary data, the nature of data collection is merely the compilation of already existing data. Secondary data may either be published or unpublished. Usually published data are available in various publications of Central, State and local governments, International Bodies and their subsidiary organizations, technical and trade journals, books, magazines and newspapers, reports and publications of various associations, reports or dissertations prepared by research scholars, universities, scientists/scholars in different fields, public records and statistics, historical documents and other sources of published information. Sources of unpublished data may be diaries, letters, unpublished biographies, autobiographies, reflective journals and also may be available with scholars and research workers, associations and bureaus and other public-private individuals and organizations.

Secondary data must be used very carefully by the researcher as it may be unsuitable and inadequate in the context of the research problem. The face value of the published statistics should not be the only criteria for the selection of secondary data as it has its limitations. The following characteristics of the secondary data should be considered while using it for the enquiry:

- **Reliability:** The researcher should know the secondary data, like- who has collected the data? What were the sources of data? At what time were the data collected? What was the method used to collect the data? etc.
- **Suitability:** In this context, the researcher should carefully scrutinize the definition of various terms and units of collection used at the time of collecting the data from the primary sources.
- **Adequacy:** If the data is found inadequate for the present inquiry, it should not be used by the researcher.

Check Your Progress 1

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

1. List various sources of primary data used in research?

.....

.....

.....

.....

2. Differentiate between primary and secondary data?

.....

.....

.....

.....

14.4 TYPES OF DATA

Data collected through various sources like primary and secondary can be in the form of quantitative and qualitative data. The nature of the data depends upon the type of tools and techniques used to collect the data. On the basis of the nature, the data are of two types: quantitative and qualitative data.

14.4.1 Quantitative Data

Quantitative data are measurable or quantifiable and can be expressed in numbers, figures or values or using some quantity. Quantitative data are expressed in numerical form and can represent size, length, duration, scores, percentage etc. For example, the total number of students who come under Economically Weaker Sections at the elementary level of a district in Delhi or 30% of students like online learning in schools etc.

Quantitative data rely on collecting numeric data from a large number of people using several tools (interview, tests, aptitude test etc.) with preset or preplanned questions and responses.

Different types of statistical measurement as well as calculation such as correlation, t-test, McNemar, chi-square, ANOVA, ANCOVA etc. can be used and results are interpreted. Since quantitative data can express relationships numerically thus can be useful if the researcher doesn't need to explore the reasons for behaviour. If you want to know how many students graduated or passed out from a particular university, quantitative data satisfies this purpose.

Quantitative data draws the criticism that it only captures a relatively narrow range of data about people's behavior i.e. what. According to Day (1998) other aspects of data like

- Who?
- What?
- When?
- Where?

But quantitative data tells us very little about "The Why?", i.e. the reasons for people's behaviour. This is partly a problem of lack of depth due to the more complex human behaviour which is difficult to measure and quantify.

14.4.2 Qualitative Data

Qualitative data tries to describe people's behaviour and such data says something about how people experience the social aspects of the world. It can be used to understand the meanings they give to behavior. Qualitative data are descriptive rather than numbers. While collecting the qualitative data, the researcher doesn't impose their interpretation on a situation (by asking direct, quantifiable questions). This type of data provides a detailed understanding of behaviour as it is concerned with discovering "the Why?" - the reasons for such behavior. Thus, this type of data provides a much realistic feel of the problem which cannot be achieved through using quantitative data. Hence, techniques that are employed to collect such type of data include observations and interviews.

Respondents may talk about things they see as significant and take the research into directions and places the researcher had not originally considered, thus it helps in pre-judging the responses- which makes the research finding more reliable.

It takes a lot of time in collecting such kinds of data and also, the researcher has to judiciously select which data to keep and which one to discard? For example trainee teacher's experience about the school internship programme, analysis of reflective diaries, interviews, observations etc.

Check Your Progress 2

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

3. What is the difference between quantitative and qualitative data?

.....
.....
.....
.....

4. Classify the following into quantitative or qualitative data:

Height of students; letters; case histories; autobiographies; census record; achievement test record; data from closed ended questionnaire; interview; weight of students.

.....
.....
.....
.....

14.5 SCALES OF MEASUREMENT

From the discussion so far, you must have developed an understanding that while measuring different variables through various tools and techniques, we get different types of data, which can be either quantitative or qualitative. The nature of measurement depends on the nature of variables and data collection devices. There are certain categories of measurement called scales. Let us discuss in brief:

Nominal Scale: This type of classification is made according to similarities and differences observed for some characteristics or properties. It gives only the nominal or minimum information about the individuals or objects.

For example:

a) What is your gender?

- Male
- Female

b) What is the colour of your hair?

- Black
- Grey
- Brown
- Any other

So in these situations, we only count the number of cases in each category and obtain the frequencies. Statistical technique 'Mode' is used in such cases.

Ordinal Scale: These scales are somehow more precise than nominal scales. e.g., ranks, order or merit position within the group or class for some quality, property or performance.

Example: Ranking of teams in a cricket match series

- Indian team -first
- Australian team-second
- New Zealand - third
- South Africa- fourth etc.

Students ranking in a class

- Student A - First
- Student B - Second
- Student C - Third
- Student D - Fourth
- Student E - Fifth etc.

Performance of a class in a school

- Class 8- first
- Class 6- second
- Class 7 – third
- Class 9- fourth likewise

In the above examples along with nominal information, it includes scores that reflect more i.e. quality or characteristics based on which the individuals/ objects are ordered. Statistics like median, percentile and rank-order correlation coefficient can be easily computed through such data.

Interval Scale: This scale not only involves the characteristics of both nominal and ordinal scales but also points out the relative qualitative and quantitative differences.

This scale has equal units of measurements but does not possess true zero as we cannot conclude that individual has zero intelligence, attitude, or self-concept. This scale is mostly used in social sciences i.e. education, psychology, sociological studies etc. Mostly parametric and non-parametric tests can be employed in data analysis where data is collected with an interval scale.

Ratio Scale: This is the highest type of scale in terms of measurement.

For example:

- height of the plant, height of the child
- Length of road
- Width of table

- The capacity of pot, bucket, auditorium etc.

These measures start with true zero. Measures are not only expressed in equal units and also possess absolute (true) zero. The majority of techniques of statistical analysis can be applied to these measures.

Let us discuss through an illustration

To study the impact of online teaching learning on students' achievement in a secondary level.

Independent Variable : Teaching style which has two categories, online and face -to -face. It is two distinct categories which comes in the nominal scale.

Dependent Variable : Scores on achievement test, these scores generally received by administering achievement tests whose range may be from 0 to 100 or 0 to 50. Scores comes in an Interval Scale.

14.6 TECHNIQUES OF DATA ANALYSIS AND INTERPRETATION

This step in the process of research follows after data collection. It is important to make sense of the information data provided by the individual respondents in the study. The analysis involves using the data to determine individual responses and then assembling the information to summarize it. Analysis and interpretation of the data involve arranging in appropriate forms and drawing conclusions, representing in tables, figures and pictures (graphics) to summarize it and then explain the conclusions to provide answers to research questions raised initially.

Data analysis and interpretation is the process of assigning meaning to the collected information/data and determining the conclusions, significance and implications of the findings. It is an important and challenging step in the process of research. In all research studies, analysis follows data collection.

According to **Kothari (1989)**, “**The term analysis refers to the computation of measures along with searching for patterns of relationship that exist among data-groups**”. The analysis involves estimating the values of unknown parameters of the population and testing hypotheses for drawing inferences.

14.6.1 Quantitative Data Analysis and Interpretation

In quantitative research, the researcher relies on a statistical analysis of the data, which is typically in numeric form. Statistical analysis involves breaking down the data into parts to answer the research questions. Statistical procedures such as comparing groups or relating scores of individuals provide information to address the research questions or hypotheses. Quantitative analysis involves analyzing trends, comparing groups or relating variables using statistical analysis and interpreting results by comparing them with prior predictions and past researches. This interpretation is the explanation as to why the result turned out the way it did and explains how the results either support or contradict the expected prediction of the study.

Quantitative analysis involves:

- The major focus is on quantifiable phenomena like- comparison, trends, relationships, differences between variables, etc.
- For this type of data statistical measures like correlation, t-test, ANOVA etc. are used.
- It involves deductive inferences.
- It involves testing of hypothesis i.e. hypothetico-deductive analysis.
- Data analysis leads to interpretation and generalization.

Now let us discuss how to analyze quantitative data in a research study. Following steps are followed in quantitative data analysis.

Step 1: Preparing and organizing data for analysis

Before analyzing and interpreting the data, the researcher should make necessary preparations so that analysis and interpretation could become systematic and logical. Some basic tasks, the researcher should perform are:

- **Editing of data:** The responses given by the sample respondents through questionnaires and schedules need to be edited by the researcher. Editing includes minute checking of the raw data for accuracy, usefulness and completeness.
- **Classification of data:** It involves dividing the data into various categories, classes, groups or heads.
- **Tabulation of data:** The process of transferring classified data from the data gathering tools to the tabular form to examine the data systematically is called tabulation of the data.

Step 2: Analysis and interpretation of quantitative data

Studying organized data to discover the inherent facts is known as the analysis of data. In this step, the data are studied through all dimensions. The plan for the analysis of data should be prepared before the actual collection of the data. Statistical techniques have contributed greatly in gathering, organizing, analyzing and interpreting quantitative data.

14.6.1.1 Statistical Measures for Quantitative Data Analysis

Quantitative data are obtained using tools which may be inventory, checklist, rating scale, questionnaire etc. Statistical measures are used to facilitate the analysis and interpretation of data. There are various statistical techniques for the analysis of data. Generally, statistical measures are classified into two types:

1. Descriptive
2. Inferential.

Descriptive statistics are computed to describe the characteristics of a sample or population in totality thus limiting generalization to a particular sample. Some of the widely used descriptive statistics are- mean, mode, median, variability, percentage, range etc.

Inferential statistical methods are used to draw generalizations beyond the sample. Some of the inferential statistical measures are 't' test, F-test, Median test, Mann-Whitney test etc.

The quantitative data are either parametric or non-parametric.

Parametric Data: Parametric data are collected mainly using interval or ratio scales. The conditions under which parametric tests are used are:

1. Assumption of normality (The data should be normally distributed)
2. The variables must have been measured on an interval or ratio scale.
3. Homoscedasticity (population must have the same variance)
4. Observations must be independent. Examples of parametric tests are 't' test and 'F' test.

Non-parametric Data: This type of data are obtained by applying nominal or ordinal scales of measurement. The data obtained is either counted or ranked. In the situations under which conditions for parametric tests are not fulfilled non-parametric tests are used. Non-parametric tests are less powerful or predictive than parametric tests. The conditions where non-parametric tests are used:

1. Where the size of the sample (N) is too small. For example $N=6$ or $N=5$
2. Assumption of normality is not fulfilled.
3. The data is measured on a nominal or ordinal scale. Examples of non-parametric tests are the Mann-Whitney test, sign test, median test etc.

Parametric and non-parametric tests are used for drawing inferences and estimations regarding the significance of sample statistics.

Following statistical techniques are generally used in the analysis of quantitative data:

Measures of Central Tendency: If we take scores of achievement motivation of students in a class and arrange the scores, we may find few students score very high and very low scores and the scores of most of the students lie between the highest and the lowest scores in the class. This tendency of the distribution of scores of a group lying between the extremes and shared by most students are called measures of central tendency. Mean or arithmetic mean, median and mode are the most common measures of central tendency.

Measures of Variability: Measure of central tendency is unable to provide a sufficient base for the comparison of two or more sets of scores and there is a need to consider the variability of scores to describe and compare the set of scores. Thus, variability is the tendency of the characteristics of a group to deviate from the average or central value. Thus, the measure of variability is to find out the expected range of variation above or below the central value or average. Range, quartile deviation, average deviation and standard deviation are the four measures of variability.

Measures of Relative Position: Median, quartile, decile and percentile are the most common measures of relative position. The relative position of an individual is measured with a reference group that can be answered using these

measures. It indicates the relative position of the given score among other scores. Percentile is a point on a score scale below which a given per cent of the cases lie. Example: P_{10} means a score point in a given distribution below which 10% cases lie and above which 90% of cases lie. Similar is the case with quartile and decile.

Correlation: Coefficient of correlation is a kind of ratio that expresses the extent to which changes in one variable are accompanied by changes in another variable. Suppose we want to study the relation of home environment on student's achievement. There are many types of correlation like linear, curvilinear, biserial, partial or multiple. The first and the foremost thing that can be interpreted using the coefficient of correlation is whether there is any relationship between variables, then the type of relationship exists and finally, it indicates the degree of closeness or significance of the relationship.

't' test: Quantitative data analysis also involves the analysis of two groups under one variable. Suppose we want to compare the mean scores of problem-solving abilities of male and female students of class XII. Here we need to apply t-test to study the significant difference between the mean scores of problem-solving abilities of the students based on their gender.

F test: If we have one variable having more than two levels. For example, A comparative study of the problem-solving ability of students of government, aided and private schools. We need to apply F test which is also called analysis of variance (ANOVA). When we compare one variable having more than two levels we call it one-way ANOVA.

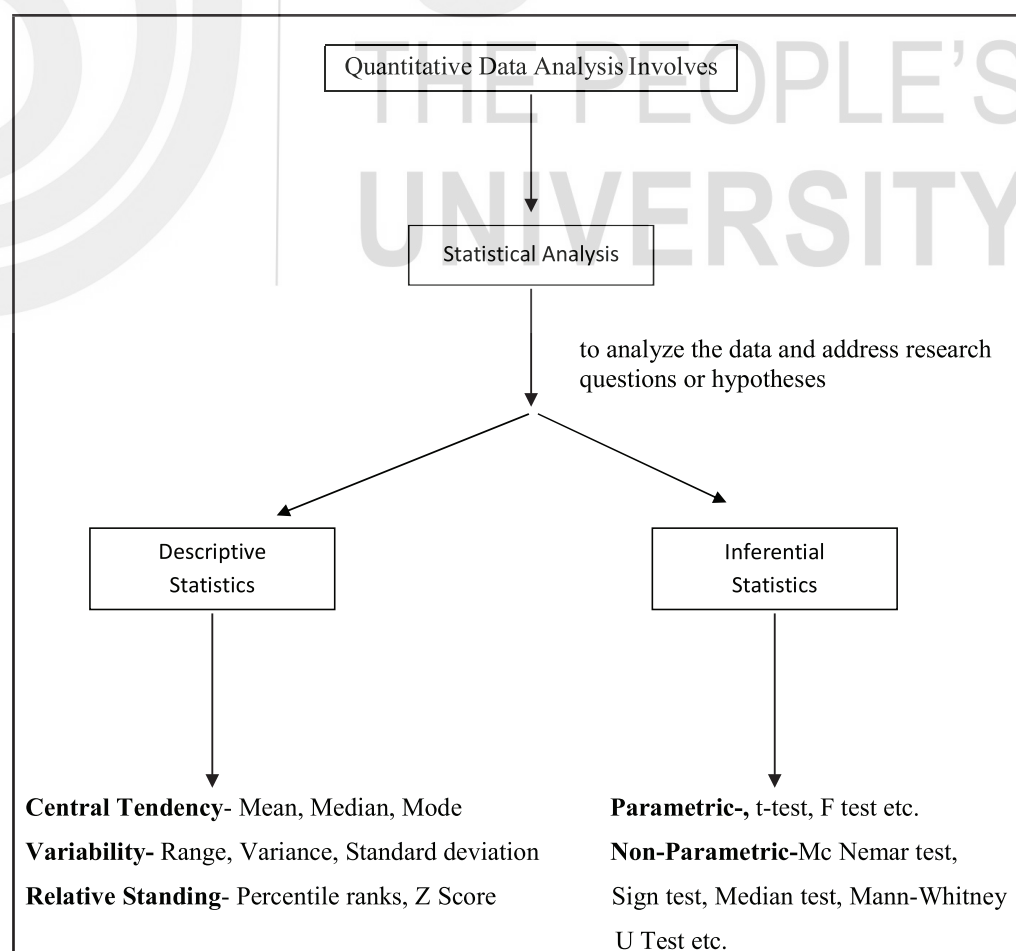


Figure 14.1: Schematic Representation of Quantitative Data Analysis

Check Your Progress 3

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

5. Differentiate between Descriptive and Inferential statistics?

.....

.....

.....

.....

14.6.2 Qualitative Data Analysis and Interpretation

In qualitative research, statistical techniques are not used to analyze the data, rather the researcher analyzes words (e.g., transcriptions from interviews) or images (e.g., photographs). Rather than relying on statistical procedures, the researcher analyzes the qualitative data i.e. words to group them into larger meanings of understanding, such as codes, categories or themes.

Qualitative data analysis involves making sense out of text and data images. It involves preparing data for analysis, conducting different analyses, going or eliciting a deeper and deeper understanding of the data, making an interpretation to get a larger meaning of the data. It is an ongoing process involving continual reflection about the data. It involves collecting open-ended data, based on asking general questions and developing an analysis from the information supplied by the participants. In this researcher collects qualitative data and analyze it for themes or perspectives and place the data in about 4-5 themes.

Some major types of qualitative researches are:

- Case study
- Ethnographic research
- Phenomenological research
- Narrative research

Some of the important aspects to be considered in different types of qualitative research have been presented in a tabular form below:

Types	Case Study Research	Ethnographic Research	Phenomenological Research	Narrative Research
Meaning	Inquiry in which the researcher explores in depth a program, an event, an activity, a process, an individual, a student, an institution, a community or a class.	The centrality of ethnographic study is culture. It involves describing, analyzing and interpreting culture sharing groups, shared patterns of culture, beliefs and language that develops over time.	Here the researcher identifies the essence of human experiences about a phenomenon that the participants describe. It is a theoretical point of view that advocates the study of direct experiences.	It involves re-storying or re-telling the participants' stories using structural devices such as plot, setting, climax etc. These are the transcribed experiences from the sample.

Characteristics	Study of a specific unit, present and past, intensive study, holistic study of a unit, cases are bounded by time and activity.	The cultural theme, a culture sharing group, context, fieldwork, researcher reflexivity.	The analysis of significant statements, generation of meaning and the development of essence description.	Individual experience, re-storying/ re-telling, collecting individual stories, coding for themes, context, the chronology of the experiences, participants collaboration.
Data collection	Involves both primary and secondary sources of data like an interview, observation, books, diaries, records, written documents etc.	Focus group discussions, life history, structured and semi-structured interviews, observations, spatial mapping, audio-visual material etc.	Open-ended questions, interviews, observations, documents, text, images etc.	Field text-discussions, conversations, interviews, journals, letters, family stories, photographs, memory boxes, etc.
Example	A case study of Eklavya School; A case study of educational opportunities of Kasturba Gandhi Balika Vidyalaya students.	The case analysis of an individual, events, activity or process of a particular community; A study of cultural practices in empowering women of a particular community.	Leadership strategies in high performing educational institutions, Understanding student's political participation in higher education institutes.	Biographies, experiences of a particular event, ethnographies, narrative interviews, oral histories.

After understanding major types of qualitative research, characteristics along data collection methods, let us try to discuss how the obtained data could be analyzed and interpreted? Qualitative data analysis could be done in the

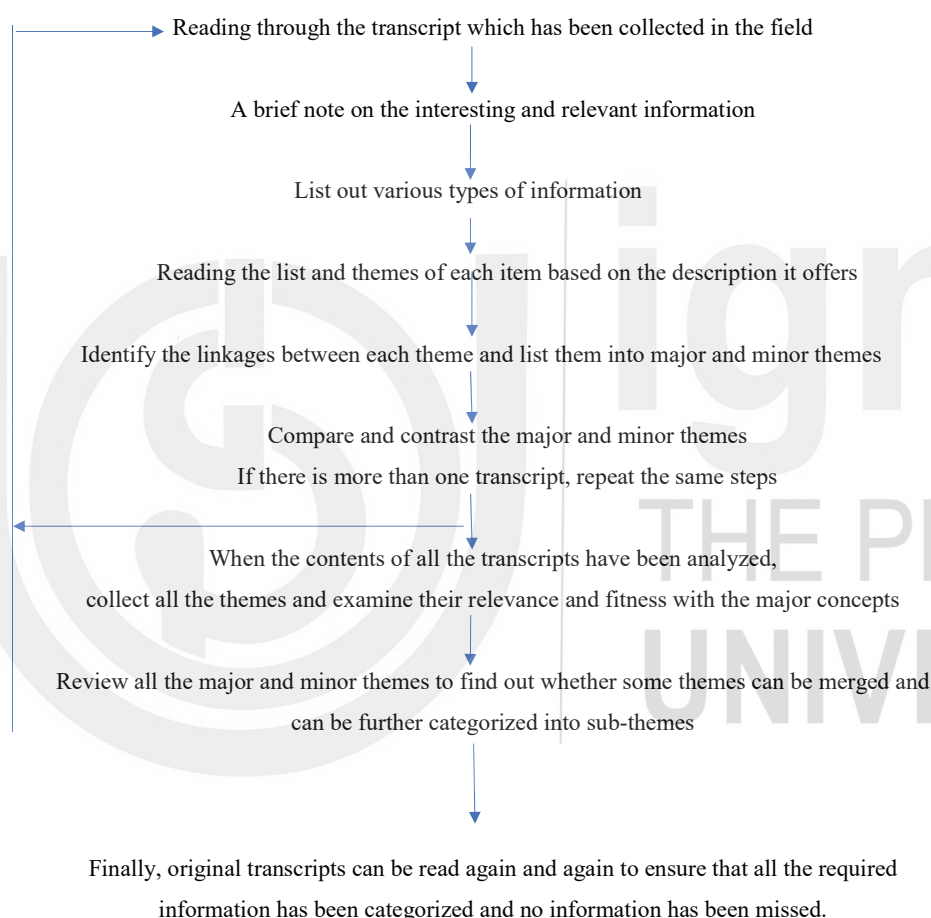
following steps:

Step 1: Analysis of Qualitative Data

Qualitative data analysis deals with the analysis of the organized data to discover the intrinsic or finer realities in terms of the specific research. The data are studied from diverse viewpoints to explore the new facts or to reinterpret already known existing facts.

Qualitative data analysis is concerned with the classification, organization and comparison of the data obtained from the field from interviews, focus group discussions, narratives, observations etc. Analysis of qualitative data generally involves following steps with some specificity concerning qualitative research.

For qualitative data analysis the following procedures need to be carried out:



Step 2: Interpretation of the Findings of Qualitative Data

After understanding the steps involved in the analysis of the qualitative data we need to interpret the findings without which the conclusion seems incomplete. Let us understand how we can interpret the findings of the qualitative data? Interpretation needs to be accompanied by:

- Reviewing the major findings of the analysis and answering the research questions.
- The interpretation of the findings should be supported by personal reflections of the researcher about the meaning of the data.
- Personal views and reflections of the researcher about the meaning of

the data should be compared and contrasted with the literature in the concerned area or context or setting.

- Limitations of the study should be explained reasonably.
- Suggestions for future research should be given by the researcher for the follow-up study.

Step 3: Validation of Qualitative Data Analysis and Interpretation

To establish the authenticity of the qualitative data analysis and its interpretation we need to demonstrate the accuracy of the findings and convince users about its accuracy. Let us find out what is validation in qualitative research?

The process of validation is the strength of qualitative research and it is based on determining whether the findings are accurate from the perspective of the researcher, and the participants of an account. Validation involves triangulation. Triangulation is the process used by the researcher to corroborate evidence from different individuals, types of data and methods of data collection. The researcher needs to triangulate different data sources of information by examining evidence from the sources and building an intelligible justification for the themes.

Triangulation plays an important role in the interpretation of the data. It can be done in different ways:

a) Triangulation: Integration of Qualitative and Quantitative Data

In this quantitative data (example- inventory, checklist, rating scale, questionnaire, etc.) as well as qualitative data (example- reflective diaries, interviews, observations, etc.) are integrated, matched and contrasted to arrive at common interpretations. It assumes that it is useful to bring a variety of data and methods of the same problem to reduce system bias in interpreting the results of the research study.

b) Triangulation: Comparing Multiple Qualitative Sources

This type of triangulation involves comparing and verifying the consistency of the data obtained by different means at different times using qualitative methods. Comparing the data collected from observations with interviews, data collected from observations with questionnaires, responses of the participants of a program in public with responses in private, checking for the consistency of the views of the participants about a program over some time.

c) Triangulation: Multiple Perspectives from Multiple Observer

This type of triangulation helps to reduce the bias or subjectivity of a single observer by involving multiple interviews and observations. For example, several interviewees or individuals could be involved in data collection, two or three individuals together collect the data of the same phenomenon, event

or individual's opinion could be interpreted through multiple perspectives with that potential biases could be reduced.

Check Your Progress 4

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

6. What are the major steps of qualitative data analysis?

.....

7. What is meant by triangulation?

.....

14.7 LET US SUM UP

In this unit, we have discussed various sources of data i.e., primary and secondary. We also described types of data - quantitative and qualitative. Apart from this, techniques of data analysis and interpretation. Quantitative and qualitative data differ largely in terms of analysis, where one depends upon quantitative measures or statistics and the other follows qualitative descriptions with minimal use of quantitative measures. Along with quantitative interpretations, the unit also dealt with the analysis of qualitative data, various types of qualitative researches such as phenomenological, case study, ethnography, narratives. Triangulation methods have also been discussed for validation of qualitative data analysis and its interpretation. In the next unit, we will be dealing with the statistical part of quantitative analysis.

14.8 UNIT END EXERCISES

1. Review some educational quantitative research studies and enlist the nature of data, types of statistics used and method of interpretation.
2. Explore at least one research study for each type of qualitative research. Review their data analysis method and make a comparison table.

14.9 REFERENCES AND SUGGESTED READINGS

- Ahuja, R. (2003). Samajik Sarvek shanvev Anusandhan, Jaipur, Rawat Publications.
- Baliya, S., Arora, R. & Sharma, O. P. (2013). Jaipur, Rajasthan Hindi Granth Academy.
- Best, J. W. (2012). Research in Education, New Delhi: Prentice Hall of India.

- Cohen, L., Manion, L., Morrison, K. (2007). Research Methods in Education, London and New York: Routledge Publication.
- Creswell, J. W. (2008). Research Design: Quantitative, Qualitative and Mixed Design, New Delhi: Sage Publications Inc.
- Creswell, J. W. (2011). Educational Research: Planning, Conducting and Evaluating Quantitative and Qualitative Research, Noida, Pearson Publication.
- Creswell, J. W. (2012). Qualitative Inquiry and Research Design: Choosing among Five Approaches, New Delhi: Sage Publications Inc.
- Garrett, H. E. (2005). Statistics in Psychology and Education. New Delhi: Paragon International Publishers.
- Gupta, S. P. (2015). Anusandhan Sandarshika, Allahabad, ShardaPustak Bhawan.
- Guilford, J. P. (1965). Fundamental Statistics in Psychology and Education, New York: Mc Graw Hill.
- Kothari, C. R. (1989). Research Methodology, New Delhi: New Age International Limited Publisher.
- Kaul, L. (2009). Methodology of Educational Research. New Delhi: Vikas Publications House Pvt. Ltd.
- Kerlinger, F. N. (20017). Foundations of Behavioural Research. New Delhi: Surjit Publication.
- Mangal, S. K. (2007). Statistics in Psychology and Education. New Delhi: Printice Hall Publication.
- Pandey, K. P. (2012). ShaikshikAnusandhan, Varanasi, Vishwavidyalaya Prakashan.
- Patton. M. Q. (2002). Qualitative Research and Evaluation Methods. Thousand Oaks: C.A: Sage.
- Scott, David & Usher, Robin (1996). Understanding Educational Research. New York: Rout ledge.
- Shank, G. D. (2002). Qualitative Research. Columbus, ott: Merrill, Prentice Hall.
- Sharma, B. (2004). Methodology of Educational Research. New Delhi: Vohra Publishers and Distributors.
- Sharma, S. R. (2003). Problems of Educational Research. New Delhi: Anmol Publications Pvt. Ltd.
- <https://www.statisticshowto.datasciencecentral.com/primary-data-secondary/https://www.slideshare.net/ChinchuC1/thematic-content-analysis-in-psychology>

14.10 ANSWERS TO CHECK YOUR PROGRESS

1. Observation, Interview, Questionnaire, Schedules, Projective techniques, Experiments, Surveys etc.

2. Primary data is directly collected by the researcher, by visiting the field of study or individual through interviews or schedules or questionnaires. It is called primary because the data are collected by the researcher directly from the sample/individual for the first time. Primary data are collected first hand and thus happen to be original. Whereas Secondary data are the data that are obtained by the researcher through published and unpublished documents, letters, diaries, manuscripts etc. Secondary data are the data that are already available and has been collected and analyzed by someone else.
3. Quantitative data are measurable and can be expressed in numbers, figures or values that can be expressed in quantity. Quantitative data are expressed in numerical form and can represent size, length, duration, scores etc. Quantitative data are obtained by applying various scales of measurement. Whereas the data obtained in the form of sentences or words and not in numbers are. Qualitative data are descriptive rather than in numbers. Observation and interviews are used to collect qualitative data.
4. Classify the following into quantitative and qualitative data.

Quantitative Data	Qualitative Data.
Height of students, census record, achievement test record, data from closed-ended questionnaires, the weight of students.	letters, case histories, autobiographies, interviews,

5. Descriptive statistics are computed to describe the characteristics of a sample or population in totality thus limiting generalization to a particular sample. Some of the widely used descriptive statistics are- mean, mode, median, variability, percentage, range etc.

Inferential statistical methods are used to draw generalizations beyond the sample. Some of the inferential statistics measures are 't' test, F-test, Median test, Mann-Whitney test etc.

6. Major Steps in qualitative data analysis are:
 - The researcher reads through the transcript which has been collected in the field and make a brief note on interesting and relevant information.
 - Listing of various types of information is made.
 - Read the list and each item based on the description it offers.
 - Identify the linkages between each theme and list them into major and minor themes.
 - Compare and contrast the major and minor themes.
 - If there is more than one transcript repeat the above process.
 - When the contents of all the transcripts have been analyzed, collect all the themes and examine each in detail in terms of its relevancy and fitness with the major concept.
 - Review all the major and minor themes to ascertain whether some

themes can be merged and some themes can be further categorized into sub-themes.

- Finally, original transcripts can be read again to ensure that all the required information has been categorized and no information has been missed.
7. Triangulation is the process used by the researcher to corroborate evidence from different individuals, types of data and methods of data collection. The researcher needs to triangulate different data sources of information by examining evidence from the sources and building an intelligible justification for the themes.



UNIT 15 ORGANIZATION AND GRAPHICAL REPRESENTATION OF DATA

Structure

- 15.1 Introduction
- 15.2 Objectives
- 15.3 Nature of Data
- 15.4 Organization of Data
 - 15.4.1 Frequency Distribution Table and Its Interpretation
- 15.5 Graphical Representation of Data
 - 15.5.1 Need of Graphical Representation of Data
 - 15.5.2 Types of Graphical Representations
 - 15.5.3 Graphical Representation of Discrete Variable/ Ungrouped Data
 - 15.5.4 Graphical Representation of Continuous Variable/Grouped Data
- 15.6 Let Us Sum Up
- 15.7 Unit End Activities
- 15.8 References and Bibliography
- 15.9 Answers to Check Your Progress

15.1 INTRODUCTION

While conducting educational research, statistic is used for analysis of information draw conclusions or inferences. This process consists of three steps: observations/collection of data or information, analysis of information/data analysis and drawing inferences/conclusions. With the help of observations, we obtain information either in the form of ideas or in numerical form. The information that is in the form of ideas is qualitative in nature and that is numerical form is quantitative in nature. Hence, we are likely to get either qualitative or quantitative data. Qualitative data may either be analysed using qualitative methods or may be converted into quantitative data through transforming the ideas into numbers. And the information collected in terms of numerical figures is known as the data. In this unit we shall discuss about nature methods of organization and graphical representation of data.

15.2 OBJECTIVES

After going through this unit, you will be able to

- explain nature of data;
- describe the procedure of classification and tabulation of data ;

Note: Few sections of the Unit has been partially taken from Unit-13 'Tabulation and Graphical Representation of Data', BES-127 'Assessment for Learning, B.Ed. IGNOU, 2017

- present data in frequency tables;
- explain the need of graphical representation of data; and
- explain various methods of graphical representation of data like histogram, pie charts, frequency polygon, cumulative percentage curve.

15.3 NATURE OF DATA

In the previous unit we have discussed about various types of data like primary or secondary data; quantitative or qualitative data. By now, you may have understood that quantitative data is collected by using various psychological tools or tests and the data is in the form of raw scores. For example, if you ask someone to rate the quality of food offered by a particular restaurant, the respondent can rate it on the scale of 1 to 5. Say, the respondent rated it as 5, then the obtained data is **raw score**. A raw score is **simply unaltered data from a test or observation**. You may get a set of raw data which is a collection of raw scores on the entire tool. Raw data does not have any meaning by itself or it may not give any proper insight unless it is not organized or transformed. For example, if a student has secured 45 marks in physics examination, then 45 is a raw score obtained by that particular student and the raw data set would be the raw scores obtained by the entire class. The raw score 45 does not indicate anything like what grade he/she has earned, what is the position secured in the class. For statistical treatment, we have to convert raw scores into derived scores. **Derived score** is the numerical description of an individual performance in terms of norms and thus helps to compare performance which is obtained after transforming a raw score. Common forms of derived scores are Z-score, T Score, Percentile etc. .

Apart from this on the basis of the nature of variable, quantitative data further could be classified into two types: Continuous Data and Discrete Data.

- **Continuous Data:** This type data is one which is capable of any degree of subdivision. Continuous data represents measurement and therefore their values cannot be counted but they can be measured. The scores of an intelligence quotient (I.Q.) test, achievement test and height or weight of a person fall into continuous series. Suppose if you are asked to measure your height with a meter rule, you might say you are 160 cm tall. However, if you had a more accurate rule, you might actually find you are 160.362 cm tall. This gap from 160 cm to 160.362 cm is in a truly continuous series and the precision of measurement fulfils the gaps.
- **Discrete Data:** Discrete data is a count that involves only integers. The discrete values can't be subdivided into parts. A discrete series exhibit the real gaps. For instance, if in a class there are 45 boys and 18 girls, then you know there is real gap between the two groups. So, this type of data cannot be measured but it can be counted. The national census is another example of discrete data.

On the basis of data organisation, data may be treated in two ways -ungrouped and grouped.

- **Ungrouped Data:** An **ungrouped** data is basically a list of numbers

which you first gather from an experiment or research study by using research tools. It is raw data, which is not sorted into categories or classified, or otherwise grouped. Let us take an example, A teacher administered achievement test on 20 students. Following are the scores:

25 33 36 28 35 35 25 45 23 20
21 36 45 33 34 35 33 33 45 40

This type of data is called **Ungrouped data**. This raw data is not giving any insight to the researcher like who got highest marks or who got lowest marks or who failed and so on. To get answers into such questions, we may sort the data into groups or into class intervals.

- **Grouped Data:** Grouped is the data that has been bundled together in categories or class intervals. When the sample is small, we normally analyse the data in ungrouped form and if obtained data is large then it has to be grouped. Let us understand through an example, Sanjay a science teacher conducted a unit test for 70 students. If the teacher arranges marks secured by each student in a serial order, it may be difficult to identify how many students secured average marks or how many failed and so on. But if, he converts the marks obtained into class intervals (for example, students who scored marks in between 0-10, 10-20, 20-30 and so on), and then represents them in a tabular form then the performance can be easily recognized. This type of **sequential arrangement of data in a table is called frequency distribution**. The term ‘**frequency**’ refers to the number of cases or objects in a category or class. Or we can say that **frequency** is the number of times a particular data point occurs in the set of data. For example, suppose 8 students scored marks below 35 in the term examination, then the ‘8’ represents frequency.

Check Your Progress 1

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

1. How is derived data different from raw data?

.....
.....
.....
.....

2. Differentiate between continuous and discrete data.

.....
.....
.....
.....

3. Compare between ungrouped and grouped data.

.....
.....
.....
.....

15.4 ORGANIZATION OF DATA

Now that you are aware about the importance of organization and classification of data, let us understand it further. The classification of scores is known as **frequency distribution** and a table representing them is known as frequency distribution table. Generally, the classification of data is done in two forms.

- Ungrouped frequency distribution
- Grouped frequency distribution

1) Ungrouped Frequency Distribution: This type of frequency distribution is prepared into when you have responses from a small sample, i.e. in which limited numbers are repeated. At first, the scores are arranged in ascending or descending order then tallies are marked for each score.

Let us understand through the same example discussed as in the earlier section 15.3 where a teacher administered achievement test to 20 students. Following are the scores:

25 33 36 28 35 35 25 45 23 20
21 36 45 33 35 35 33 33 45 40

Let us make the frequency distribution table for the above data. Table 15.1 shows frequency distribution table for ungrouped data.

Scores	Frequencies
20	1
21	1
23	1
25	2
28	1
33	4
35	4
36	2
40	1
45	3

Table 15.1: Frequency Distribution Table for Ungrouped Data

- 2) Grouped Frequency Distribution:** Grouped frequency distribution is a table in which total range of data is classified into class intervals of such a size. Generally, there are two ways to prepare class intervals :
- Exclusive Class Interval
 - Inclusive Class Interval

1. Exclusive Class Interval: In an exclusive class-interval, upper scores of the class-interval are not included in these class-intervals. For example, in a class interval of 15-20, 20 is an upper score and it be excluded. This upper score (i.e. 20) will be included in the next class-interval.

2. **Inclusive Class Interval:** In an inclusive class-interval, both the lower and upper scores are included in class interval. This type is considered to be the most convenient method.

Table 15.2 indicates two types of class intervals.

Exclusive Class Interval	Inclusive Class Interval
35-40	35-39
30-35	30-34
25-30	25-29
20-25	20-24
15-20	15-19
10-15	10-14
5-10	5-9
0-5	0-4

Table15.2: Types of Class Interval

15.4.1 Frequency Distribution Table and Its Interpretation

To develop the frequency distribution, we need to follow certain procedures and steps, which are explained below:

1. Calculation of range
2. Calculation of size of class interval
3. Indication of class intervals
4. Putting the tallies
5. Obtaining frequencies
6. Indication of frequency table

So now let us understand it with the help of an example. Following are the scores obtained by 35 students on an intelligence test and are shown below:

55 62 73 44 67 78 79 90 56 67 78 108
65 69 79 75 98 58 72 85 86 82 97 74
99 103 106 88 100 92 88 56 49 77 42

In order to make the frequency distribution, the following procedures need to be carried out:

1. **Calculation of Range:** Range is the difference between the highest score and the lowest score in the set of data. In the above case '108' is the highest score and '42' is the lowest score.

Therefore, Range= Highest Score-Lowest score

$$\text{Range} = 108 - 42 = 66$$

2. **Calculation of Size of Class Interval:** To calculate the size of the class interval, first finalise the number of classes in which you want to represent your data. The number of classes will depend on the size of data, larger the

data more the number of classes or vice versa. For calculating the size of a class interval, you may use the following formula;

Size of the Class Interval = $\text{Range} / \text{Numbers of classes desired}$

From the above example, it can be deduced:

(Note: here we decide to take seven classes by observing the size of data)

Size of class interval = $66/7 = 9.4 \approx 10$

Commonly used size of class intervals is 5 and 10 as they are easier to work in the calculation.

3. **Indication of Class Intervals:** Once the size of a class interval is decided, then the class intervals having the lowest to highest scores are arranged either in ascending or descending order. So in order to write the class interval, the lowest score, higher score and size of class interval is used.

If the lowest score is close to term i.e. 10, 15,.....30, 35, 40, 45....., the class interval can be started from these terms rather than actual scores.

In the above case, the lowest score is '42' (which is near to 40) and size of class interval is '10' and hence the first class interval would be 40-50. The next class interval would be 50-60, 60-70 and so on.

4. **Putting the Tallies:** After the class intervals are arranged, you may start putting tallies against each class intervals. For this the number of cases occurring in each class interval is noted and is denoted by using tallies. At this point, the style of putting tallies is to be paid attention. In order to put tallies, we may start from 1 and go up to 4, then after the fifth tally we mark it by drawing diagonal line as shown below. In the above example, the number of cases appearing in each class interval is represented using tallies as shown below:

Tally 1	Tally 2	Tally 3	Tally 4	Tally 5
I	II	III	IIII	IIII

5. **Obtaining Frequencies:** The marked tallies are counted separately for each class interval. In this example, the total number of tallies against each class interval is given below in table 15.3 :
6. **Indication of Frequency Table:** The final step is to check the total number of tallies to get the total number of cases, 'N'. This is found out by adding all the frequencies. In this case, the total frequency ($\sum f$) is 35, wherein ' $\sum$ ' stands for summation and 'f' for frequency. Final frequency table is shown below in table 15.3.

Class-Interval (x)	Tallies	Frequencies(f)
100-110	IIII	4
90-100	IIII	5
80-90	IIII	5
70-80	IIII III	9
60-70	IIII	5
50-60	IIII	4
40-50	III	3
		$N = \sum f = 35$

Table 15.3 : Frequency Distribution Table

On the basis of the frequency distribution table one can say that 19 students have an average I.Q level (i.e. 60-70, 70-80, 80-90) whereas only 4 students have very high I.Q. (i.e.100-110).

Activity 1

Analyze the frequency distribution table 15.3. Note down the interpretation that can be made out of it.

.....

.....

.....

Check Your Progress 2

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

4. Enumerate the ways for preparing class intervals.

.....

.....

.....

5. Discuss the steps involved in development of frequency distribution table.

.....

.....

.....

15.5 GRAPHICAL REPRESENTATION OF DATA

Many a times, numerical data is complex and difficult to interpret and understand. For example, in the above case, the marks scored by students have been converted into frequency distribution. But at times frequency distribution may not serve the purpose as it is not visually appealing and is complex. So, a more interesting and attractive kind of representation came into practice and that is the graphical form of data representation. In graphical representation the data is represented as geometric figures which can be easily interpreted and understood by anyone. But the geometric picture needs to be drawn keeping into account the proportion and measurements of data. Thus, it is possible to visualize and transform numerical data to picture or graphic format drawn considering a reasonable proportion. Graph represents the numerical data in geometric figures drawn on scales.

15.5.1 Need of Graphical Representation of Data

The graphical representation is important due to the following reasons:

- Graphical representations are attractive.
- It helps in easy visualisation.

- It facilitates trouble-free interpretation and judgements.
- It gives a bird's eye view of the entire data.
- It is easy to construct.

You may notice that, although we said graph is a form of pictorial/diagrammatic representation, these two are different. Now let us look at the major differences between graph and diagram in figure 15.1.

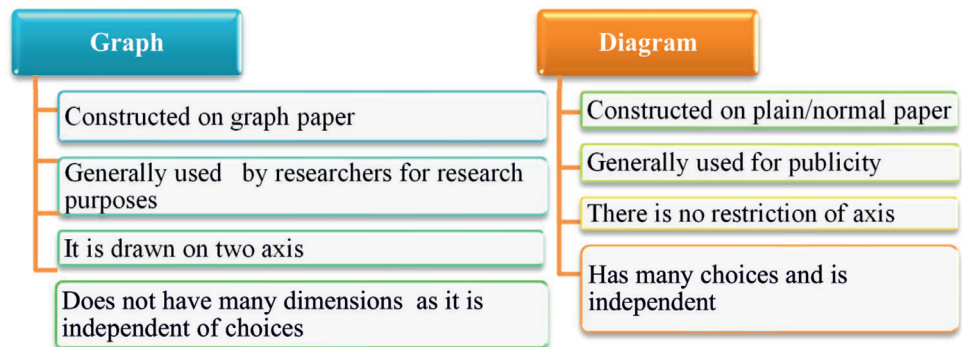


Figure 15.1: Major Differences between Graph and Diagram

When a researcher aspires to convert his/her data into graphical format, he/she needs to keep in mind the following points:

1. Draw two perpendicular lines. The point where two lines intersect is called 'origin' and is represented using '0' (zero).
2. The horizontal line is called 'X' axis. The 'x-axis is called abscissa(base).
3. The vertical line is called 'Y' axis. The y-axis is called ordinate (height).
4. The ordinate/height of the graph must be 75% of the abscissa/base. This is called 75% rule. But there is flexibility to dilate between 60% to 80%.
5. The graph generally has four quadrants as shown in figure 15.2. But educationists/psychologists usually use the (++) quadrants to utilize maximum space of the graph paper.

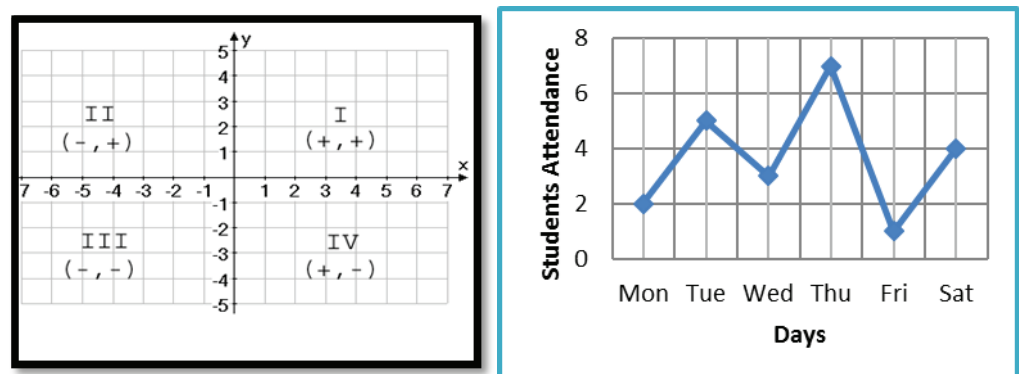


Figure 15.2: Drawing a Graph

15.5.2 Types of Graphical Representations

After having an understanding about the concept of graphical representation Let us now discuss various types of graphical representation. It is based on the type of data available for the presentation. You may have various types of graphical

representations for continuous and discrete variable.

- 1) Discrete variable in which nominal data are obtained. This type of ungrouped data is presented by the following graphical presentations:
 - i. Bar Graph or Bar Diagram
 - ii. Circle Or Pie Graphs/Diagram
 - iii. Line Graph
- 2) Continuous variable is one in which frequency distribution tables are prepared or data can be organised into class intervals. This type of grouped data is used in the following graphical presentation:
 - i. Histogram or Column diagram
 - ii. Frequency Polygon
 - iii. Cumulative Frequency Graph
 - iv. Cumulative Frequency Percentage Curve or Ogive

Check Your Progress 3

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

6. Write any four importance of graphical representation of data.

.....

.....

.....

.....

7. Write any three differences between graphs and diagram.

.....

.....

.....

.....

15.5.3 Graphical Representation of Discrete Variable/ Ungrouped Data

Usually, we use Line, Bar and Pie graphs are used for discrete data. Discrete data cannot be organized into class intervals, Since, it is available in the form of raw score it is called ungrouped data. Line, bar and pie graphs are used for analyzing pattern, trends and for comparison purpose. Now let us discuss about line, bar and pie graph graphs in details.

1. Bar Graph

By now, you know that the data for discrete variable are obtained on nominal scale or in terms of frequency rather than score. The frequencies are presented in the form of bar diagram, (rectangles with similar width) as used in bar graphs. You usually get the data in the form of frequencies, when you collect data by using a questionnaire, interview techniques or a rating scale. The following are

the steps used for constructing a bar graph:

- Select x-axis and y- axis on the graph paper. Generally, the x axis is the horizontal line and y-axis is the vertical line in the graph.
- The intersection of the x-axis and y-axis is the origin (marked as '0') of the graph.
- Choose a convenient scale for both the axis's.
- Mark the corresponding values against each variable on x-axis and y-axis and draw them as bars having equal widths.

Let us discuss here with an example and apply these steps to draw the bar graph. Consider a school, having the following number of girls in various sections of a particular grade:

S.No	Class	No. of Girls
1	8A	25
2	8B	20
3	8C	30
5	8D	15
6	8E	10

Here, the number of girls studying in various sections of grade eighth is given. In order to draw bar graph, number of girl students is taken on the y-axis and the corresponding grades are selected on the x axis. The resulting bar graph is given below in figure 15.3:

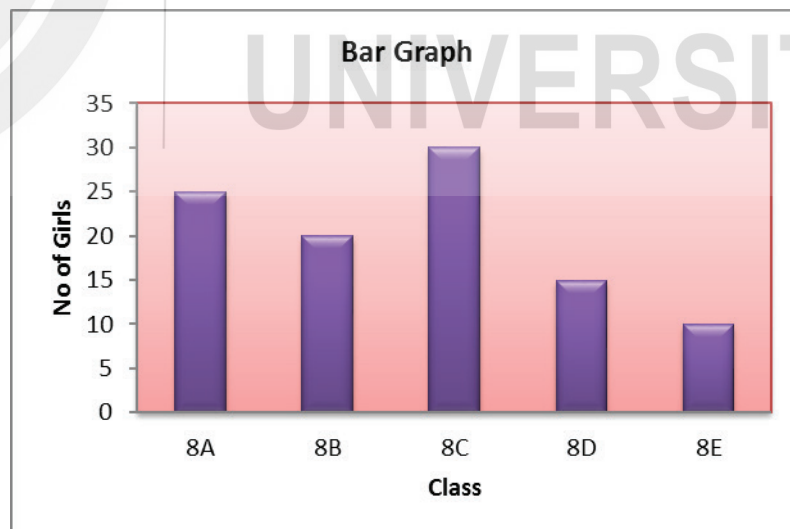


Figure 15.3: Bar Graph Showing Number of Girls Per Section

The next question is to how to interpret the bar graph. In this case, we can say that, the class 8C has greatest number of girls compared to rest of the classes and 8E has the least number of girls. What else can we infer? The difference in number of girls among classes 8E and 8C is 20. There are many more inferences that you can draw from the bar graph. Why don't you try it as an activity?

Activity 2

Draw other inferences from the bar-graph other than the ones already drawn above.

.....

.....

.....

Let's take another example. A researcher conducted a study on students' perception towards various modes of teaching -learning. He/she administered a questionnaire to 200 higher secondary school students. Following data received on students' perception on various mode of learning.

Modes of Teaching -Learning	F2F	Distance	Online	Blended
Rural Area Students	120	37	12	31
Urban Area Students	72	33	36	59

To represent the data given in the above example in a bar graph, we select two axes on the graph paper. Against the x-axis 'modes of learning' and y-axis, 'number of students (rural and urban areas)' are marked. The resulting bar graph is given below:

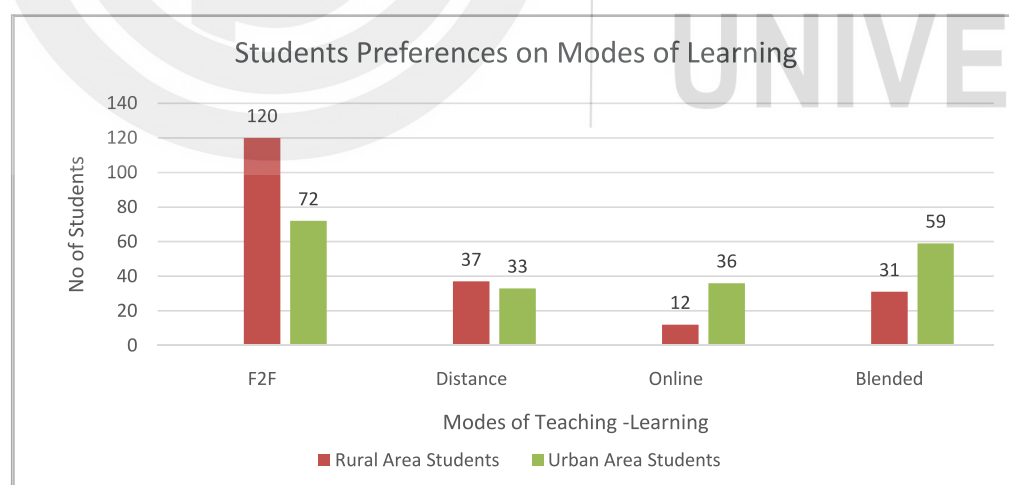


Figure 15. 4 : Bar Graph Showing Students Preferences on Teaching-Learning Modes

On the basis of this graph, we can say that face-to-face mode (f2f) is more preferred by rural area students as compared to urban area students. Whereas online mode is more preferred by urban area students as compared to rural area students.

Activity 3

In figure 15.4, bar graph shows students preference on various teaching-learning mode. Analyze the graph and draw inference other than the ones that is already discussed.

.....

.....

.....

2. Circle or Pie Graphs/Diagram

Another device for presenting a discrete data is the pie-diagram. The pie-diagram/pie graph is known as **circle graph** as we represent the statistical data as a circular figure considering weightage given to the proportions of data. The percentage is shown by circle in terms of angle. The angle of a circle at the center is 360° , which is converted into percentage i.e. 36° is equal to 10 percent and 1 percent is equal to 3.60° . The circle is shown for 100 frequency of a set of data. To construct a pie-diagram one should have the knowledge of angle measurements and percentages. Let us understand it with the help of an example.

Example: In a school, 180 students appeared in the higher secondary examination. 25 students failed, 35 passed in first division, 65 in second division and 55 in third division.

Let us try to put these details into a pie-diagram. To do so, you should have the knowledge that the value of a circle is 2π (2 pie). 2π is equal to $2 \times 180^\circ = 360^\circ$. Thus, the whole circle represents 360° . Thus, we will represent the total sample i.e. 180 through a circle having 360° . Let us see how it is done.

$$\text{Angle for I division} = (360 \times 35) / 180 = 70^\circ$$

$$\text{Angle for II division} = (360 \times 65) / 180 = 130^\circ$$

$$\text{Angle for III division} = (360 \times 55) / 180 = 110^\circ$$

$$\text{Angle for fail} = (360 \times 25) / 180 = 50^\circ$$

Type of Result	I Division	II Division	III Division	Fail	Total
In Frequency	35	65	55	25	180
In Degree of angles	70	130	110	50	360

In this way a scale for pie diagram is prepared. Now it is very easy to divide a circle into degrees to shown result in the visual presentation. The pie-diagram is shown in figure 15.5:

Pie Diagram Showing Higher Secondary Results

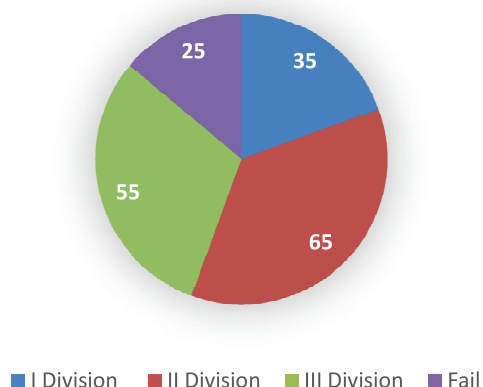


Figure 15.5 : Pie Graph

The pie diagram indicates that the second division and third division students are approximately equal. Whereas first division and failed students are also approximately equal. Pie diagram is easily understandable but is relatively difficult to prepare as it is a geometric figure.

Check Your Progress 4

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

8. What is bar- graph.

.....

.....

.....

9. What is pie-graph? Describe the procedure for preparing pie-graph.

.....

.....

.....

10. Develop pie-graph and bar graph for the following data. Which one you find most suitable for the following data and why?

S.No	Year	No of Students of Students Passed in Class 12th
1	2017	700
2	2018	820
3	2019	600
4	2020	500

.....

.....

.....

3. Line Graphs

Line graphs are one of the common mode of representation of statistical data. A line graph is commonly used to display change over time as a series of data points connected by straight line segments on two axes. The line graph therefore helps to determine the relationship between two sets of values, with one data set always being dependent on the other set. The data pertaining to the variables will be marked on two axes namely the x-axis and y-axis by choosing appropriate scales. Let us illustrate line graph through an example. Below, are given the details of absentees in a particular class.

Day	Mon	Tue	Wed	Thu	Fri	Sat
Absentees	2	5	3	7	1	4

Table 15.4 : Data for Line Graph

To represent the data given in the table 15.4 in a line graph, we select two axes on the graph paper. Against the x-axis 'day' and y axis, 'absentees' are marked. After that, appropriate scale is decided. In this case, as the number of absentees ranges from 1 to 7, we may choose 1 square of the graph as 1 along y-axis. Similarly against x-axis, each square can be chosen as a day. Then after, the data of absentees pertaining to each day are marked. The resulting line graph is given below:

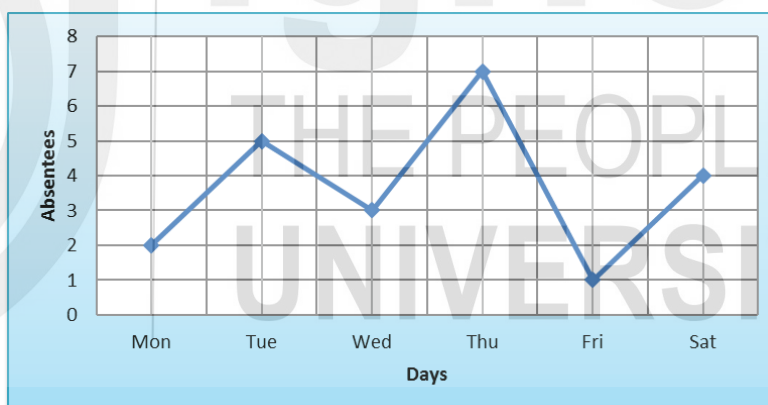


Figure 15.6 : Line Graph Showing Absentees in a School

Line graphs are useful in that they show data variables and trends very clearly and can help to make predictions about the results of data not yet recorded. They can also be used to display several dependent variables against one independent variable. So, more than one line will be drawn. To get better understanding let's take an example.

Example: Discipline wise pass percentage of result of high school are as follows (results is given in pass percentage):

Year	Humanities	Science	Commerce
2017	56	78	70
2018	58	80	90
2019	59	85	72

2020	60	89	98
2021	61	90	76

In the sample results are recorded year wise in different disciplines. There are three disciplines in the school, therefore three lines are drawn. The resulting line graph is given below in figures 15.7

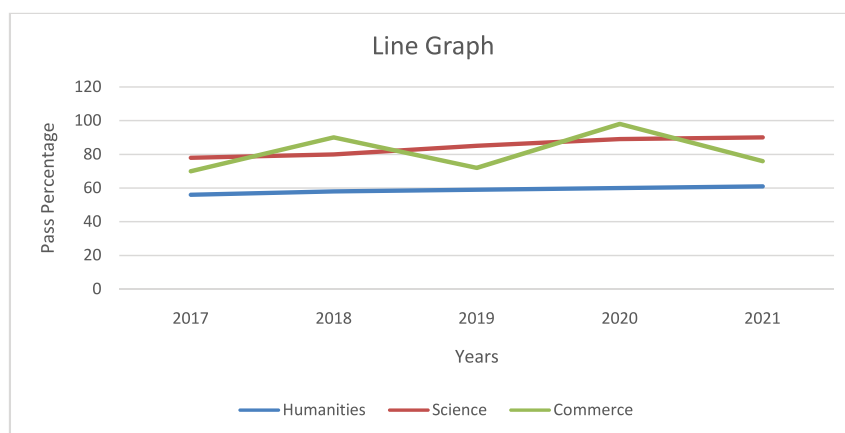


Figure 15.7: Line Graph for Discipline Wise Result of High School Passed Students

This presentation indicates that there is a progressive trend in science results while in humanities result have an upward trend. There is a great fluctuation in the results of commerce students as shown by the visual presentation of the data.

Check Your Progress 5

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

11. What is a line -graph? Higher secondary exams results of a school are given in the following table (results is given in pass percentage). Develop a line-graph.

S.No	Year	No of Students of Students Passed in Class 12th
1	2017	89
2	2018	92
3	2019	94
4	2020	98
5	2021	99

15.5.4 Graphical Representation of Continuous Variable/Grouped Data

Continuous variable is one that is organized in frequency table and class interval. It is called grouped data. The graphical presentation - histogram, frequency polygon and ogive are prepared for continues data/grouped data. Let us discuss the procedure for each of them.

1. Histogram or Column Diagram

We have studied the bar graph and the process followed in constructing them. Histogram is essentially a bar graph of a frequency distribution. Bar graph is used for discrete data but histogram is used for continuous data where the statistical data is arranged in class intervals. Here, the frequency is represented using vertical adjacent rectangles. Generally, the class interval is depicted on 'x' axis and frequency on 'y' axis. Thus, the base of the rectangle represents the class interval and height its frequency. You may conclude that histogram is the graphical representation of grouped data in the form of vertical bars (equal width) whose area is proportional to the frequency represented. It is to be noted that, histograms cannot be constructed with open end classes.

Now, let us discuss the process followed in the construction of a histogram. For that, the data is given in the Table 15.5.

Class Interval	Frequency(f)	Limits
30-34	8	29.5-34.5
25-29	5	24.5-29.5
20-24	3	19.5-24.5
15-19	6	14.5-19.5
10-14	2	9.5-14.5
5-9	3	4.5-9.5

Table 15.5 : Data for Histogram

To construct a histogram using the frequency distribution given above the following process is followed.

- First the limits of the class intervals are calculated. To compute limits, both lower limit and upper limit of each class interval is found out. For example, the lower and upper the limits of class interval 5-9 is 4.5 and 9.5 respectively and the class interval is written as 4.5-9.5.
- The lower limit and upper limits are plotted on the x-axis
- The frequencies are plotted on the y-axis.
- Thereafter, each class interval is depicted using adjacent rectangular bars of equal width.
- Keep in mind to select appropriate scales for both x-axis and y-axis.
- While constructing a histogram, 75% rule is followed i.e. the height of the figure should be approximately 75% of its width.

The histogram for the above frequency distribution is presented in Figure 15.8 :

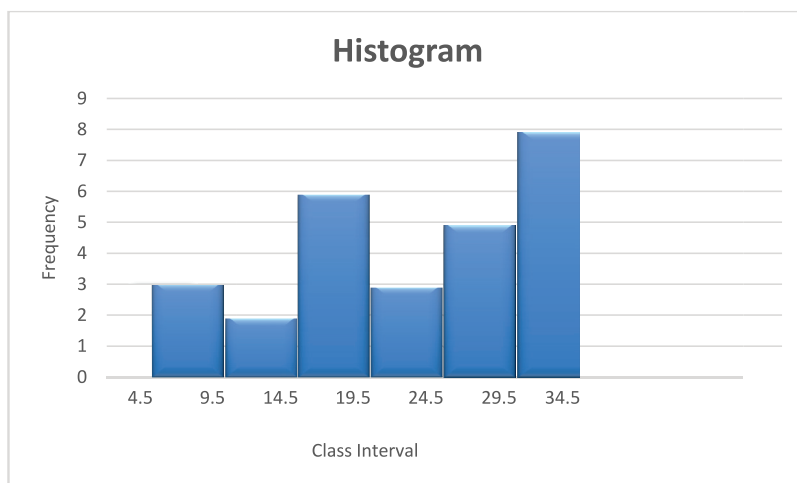


Figure 15.8 : Histogram

2. Frequency Polygon

Frequency polygon is a line graph representation of statistical data/frequency distribution. It is a closed figure with many sides. To construct a frequency polygon, the mid points of histogram are joined together and the two end sides are connected to the base line(x axis). As the end points are touching each other than they form a closed figure and hence the name **frequency polygon**. Now, let us look at steps followed in construction of a frequency polygon using the same data discussed in the previous section. The same data is reproduced in Table-15.6.

Class Interval	Frequency(f)	X(Mid -Point)
30-34	8	32
25-29	5	27
20-24	3	22
15-19	6	17
10-14	2	12
5-9	3	7

Table 15.6 : Data for Frequency Polygon

- To draw a frequency polygon, first of all, the mid points of class interval are found out and are represented using the letter 'X'.
- The mid points of class intervals are represented on x -axis.
- The frequency of class intervals are indicated on y- axis
- Then the corresponding frequency is plotted against each midpoint in the graph and is connected using straight lines.
- Finally, the start point and end point of the frequency polygon are connected to '0' on the x- axis. This can be achieved by adding a lower limit and higher limit (add an extra class interval at the lower/higher limit). This helps to construct a closed polygon.

The frequency polygon for the given frequency distribution(Ref:table15.6) is given in figure-15.9.

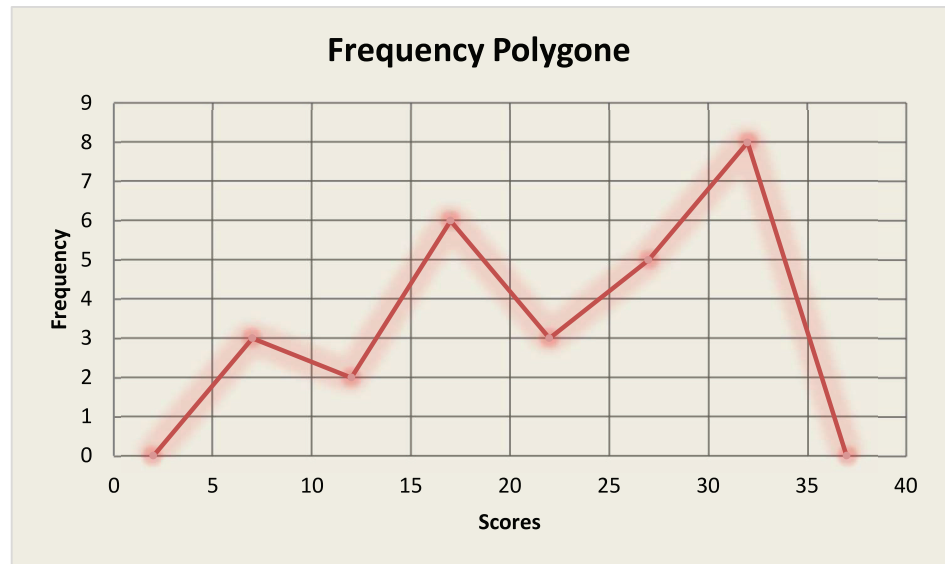


Figure 15.9 : Frequency Polygon



The histogram and frequency polygon are constructed on same graph on same scale. The area covered by histogram and polygon is equal. Mid points of bars of histogram is joined and its ends are also joined to the axis by adding one class interval at upper end and at lower end of the distribution.

3. Cumulative Frequency Graph

Another method of representing grouped and continuous data is through cumulative frequencies. In cumulative frequency graph, the frequencies are added and the resulting cumulative frequencies are plotted in the graph. Let us represent the data given in table 15.7 in a cumulative frequency graph.

Class Interval	Frequency(f)	Upper Limit	Cumulative Frequency
30-34	8	34.5	24
25-29	5	29.5	16
20-24	3	24.5	11
15-19	6	19.5	8
10-14	2	14.5	2
5-9 (Extra Class Interval)	0	9.5	0

Table 15.7: Data for Cumulative Frequency Graph

To draw the cumulative frequency graph, an extra class interval is added at the lowest limit whose frequency is 0. Thereafter, the frequencies are added and thus the cumulative frequencies are found out. Then these frequencies are represented on the graph. The final frequency polygon is given in Figure-15.10.

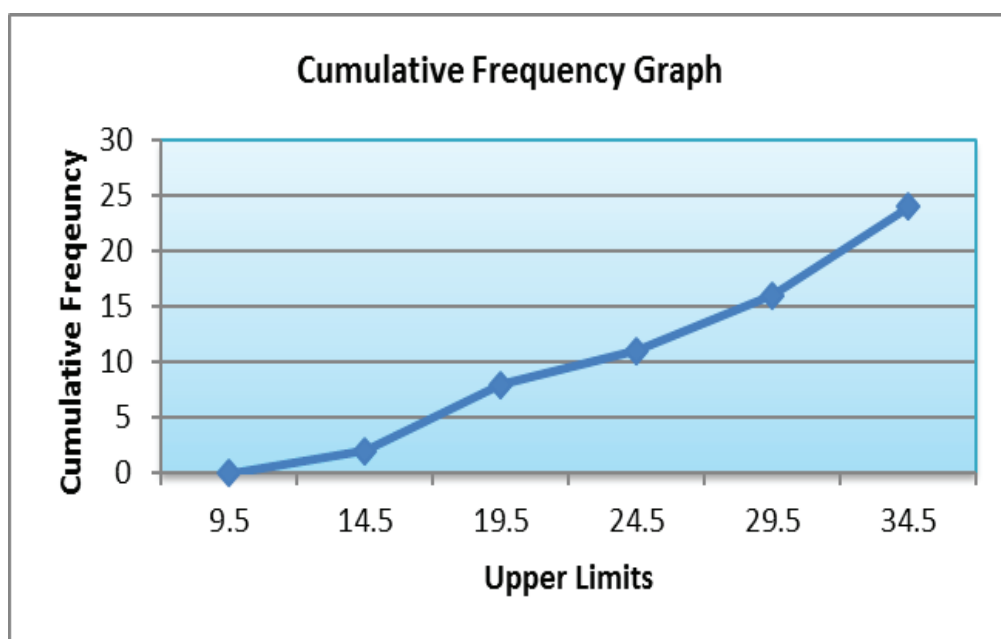


Figure 15.10 : Cumulative Frequency Graph

4. Cumulative Frequency Percentage Curve or Ogive.

Cumulative Frequency Percentage Curve or Ogive is drawn by following similar procedure used for drawing cumulative frequency graph. But an extra step is followed. Here the cumulative frequencies are expressed in terms of cumulative percentages. Thus, cumulative frequency percentage graph is a form of representation of statistical data in terms of cumulative percentages. Let us take the example given above and draw the corresponding cumulative frequency percentage graph.

Class Interval	Frequency(f) (N=24)	Upper Limit	Cumulative Frequency	% of Cumulative Frequency
30-34	8	34.5	24	100
25-29	5	29.5	16	66
20-24	3	24.5	11	45
15-19	6	19.5	8	33
10-14	2	14.5	2	8
5-9 (Extra Class Interval)	0	9.5	0	0

Table 15.8 : Data for Ogive

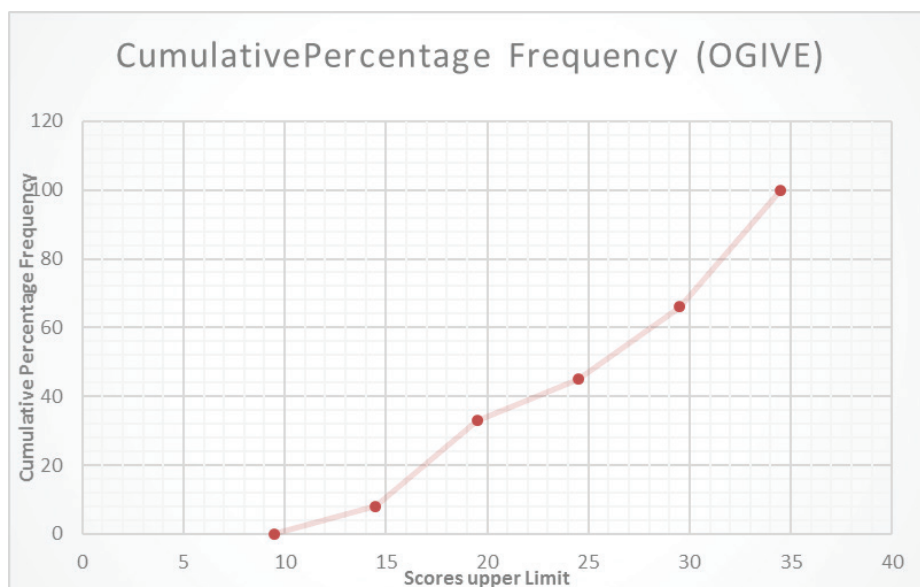


Figure 15.11 : Ogive

Check Your Progress 6

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

12. Write the differences between bar graph and histogram.

.....

.....

.....

13. Differentiate between frequency polygon and histogram.

.....

.....

.....

15.6 LET US SUM UP

In this unit you have studied about the nature of data, methods of organization of data, various types of graphs for data organization and presentation of data. It was already explained to you that in educational researches, statistic is used to collect, organize and interpret data. Data can be categorized a number of ways i.e. quantitative and qualitative data, raw and derived scores, continuous and discrete data, grouped and ungrouped data. When data is arranged into class intervals with frequency, is called grouped data. The arrangement of data in a table in a sequential manner is called frequency distribution. The term 'frequency' refers to the number of cases or objects in a category or class. To develop the frequency distribution, we need to follow certain procedures and steps like: calculation of range, decide the number of class interval, calculation of size of class interval, indication of class intervals, putting tallies, totalling the number of tallies and checking the number of frequencies.

presentation catches one's eye and holds the attention when most careful statistical analysis fails to reveal the nature of distribution. There are various types of graphical representation for discrete/ ungrouped data and continuous/ grouped data. Discrete data cannot be organized into class intervals and data in the form of raw score is called ungrouped data. Line, bar and pie graphs are used for presenting discrete data. These graphical presentations are used for analyzing patterns, trends and comparison purposes. Histogram, frequency polygon and ogive are prepared for continuous data/ungrouped data.

15.7 UNIT END ACTIVITIES

1. On a creativity test 35 students have scored the following marks Prepare frequency distribution table for on 5 size of class-interval.:

67 76 54 22 67 45 48 56 85 55 74
73 59 46 29 40 25 44 77 54 66 27
74 23 55 65 73 83 88 44 38 39 68
44 56

2. Indicate the need and importance of graphical presentation in statistical treatment. Enumerate kinds of graphical presentations.
3. Prepare bar diagram for the following data.

Result	CBSE	MP Board	UP Board	AP Board
Pass	83	78	68	95
Fail	17	12	32	5

4. Differentiate between frequency polygon and histogram. Prepare a histogram, frequency polygon, cumulative frequency polygon and cumulative percentage frequency polygon for the following distribution .

X	10-14	15-19	20-24	25-29	30-34	35-39	40-44
f	2	5	6	9	5	4	3

15.8 REFERENCES AND BIBLIOGRAPHY

- Best. J. W. (1977). Research in Education. Englewood Cliffs, New Jersey: Prentice Hall Inc.
- Garrett, H. E. (1956). Elementary Statistics. Longmans, Green and Co., New York.
- Guilford, J. P. (1965). Fundamental Statistics in Psychology and Education. McGraw Hill Book Company, New York.
- Hannagan, T. J. (1982). Mastering Statistics, The Macmillan Press Ltd., Surrey.
- IGNOU (2017). 'Tabulation and Graphical Representation of Data' in 'Analyzing and Interpreting Learner's Performance' (Block-4). BES-128 :Assessment for Learning .New Delhi
- Lindgren, B. W. (1975). Basic Ideas of Statistics, Macmillan Publishing Co. Inc., New York.

- Tate, M. W. (1955). Statistics in Education, The Macmillan Company, New York.

15.9 ANSWERS TO CHECK YOUR PROGRESS

- 1) A raw score is simply unaltered data from a test or observation and derived score is the numerical description of an individual performance in terms of norms and thus help to compare performance which is obtained after transforming a raw score.
- 2) Continuous data is one which is capable of any degree of subdivision. Continuous data represents measurement and therefore their values cannot be counted but they can be measured. Discrete data is a count that involves only integers. The discrete values can't be subdivided into parts. A discrete series exhibit the real gaps.
- 3) An **ungrouped** data is basically a list of numbers which you first gather from an experiment or research study using research tools. Grouped is the data that has been bundled together in or class intervals. When the sample is small, we normally analyze the data in ungrouped form and if obtained data is large then it has to be grouped.
- 4) Steps for development of frequency distribution table:
 - Calculation of range
 - Calculation of size of class interval
 - Indication of class intervals
 - Putting tallies
 - Totalling the number of tallies
 - Checking the number of frequencies
- 5) In an exclusive class- interval, upper scores of the class-interval are not included in these class-interval. For example, in a class interval of 15-20, 20 is a upper score and it can be excluded. This upper score (i.e. 20) will be included in the next class- interval. In an inclusive class-interval, both the lower and upper scores are included in class interval. This is the most convenient method.
- 6) The graphical representation is important due to the following reasons:
 - Graphical representations are attractive and beautiful.
 - It helps easy visualisation.
 - It facilitates trouble-free interpretation and judgements.
 - It gives a bird's eye view of the entire data.
 - It is easy to construct.

7)

Graph	Diagram
• Constructed on graph paper. • Generally used by researchers for research purposes. • It is drawn on two axes. • Does not have many dimensions as it is independent of choices	• Constructed on plain/normal paper. • Generally used for publicity • There is no restriction of axis • Has many choices and is independent

- 8) Data for discrete variable are obtained on nominal scale or in terms of frequency rather than score. The frequencies are presented in the form of bar diagram, (rectangles with similar width) as used in bar graphs.
- 9) The pie-diagram/pie graph is known as circle graph as we represent the statistical data as a circular figure considering weightage given to the proportions of data. The percentage are shown of circle in terms of angle. The angle of a circle at the center is 360° , which is converted into percentage i.e. 36° is equal to 10 percent and 1 percent is equal to 3.60° . The circle is shown for 100 frequency of a set of data.
- 10) Self-exercise
- 11) Line graphs are one of the common mode of representation of statistical data. A line graph is commonly used to display change over time; as a series of data points connected by straight line segments on two axes. The line graph therefore helps to determine the relationship between two sets of values, with one data set always being dependent on the other set.
- 12) Histogram is essentially a bar graph of a frequency distribution. Bar graph is used for discrete data but histogram is used for continuous data where the statistical data is arranged in class intervals.
- 13) Frequency polygon is a line graph representation of statistical data/frequency distribution. It is a closed figure with many sides. To construct a frequency polygon, the mid points of histogram are joined together and the two end sides are connected to the base line (x axis). As the end points are touching each other then they form a closed figure and hence the name **frequency polygon**.

UNIT 16 ANALYSIS OF QUANTITATIVE DATA-I

Structure

- 16.1 Introduction
- 16.2 Objectives
- 16.3 Measures of Central Tendencies
 - 16.3.1 Mean
 - 16.3.2 Median
 - 16.3.3 Mode
 - 16.3.4 Comparison of Mean, Median and Mode
- 16.4 Measures of Variability
 - 16.4.1 Range
 - 16.4.2 Quartile Deviation
 - 16.4.3 Mean Deviation
 - 16.4.4 Standard Deviation
- 16.5 Measures of Relative Position
 - 16.5.1 Percentile and Percentile Rank
 - 16.5.2 Calculation of Percentile and Percentile Rank
- 16.6 Let Us Sum Up
- 16.7 Unit End Exercises
- 16.8 References and Suggested Readings
- 16.9 Answers to Check Your Progress

16.1 INTRODUCTION

In the previous unit we have learnt about the nature, methods of organization and graphical representation of data. In any research process, when we administer any tool on a group of individuals, we get raw data about a variable under study. To make raw data useful and meaningful, we use some formulas and calculations to arrive at meaningful results which is called statistics. Statistical measures always relate to a sample or group of individuals rather than to a single variable. The statistical measures are of two types-descriptive and inferential. The descriptive statistical measures are used for the description of a particular group and are not used for formulation of a general theory or not applicable to a similar group. The inferential statistical measures obtained from a sample are to be used to the similar group selected from the same sample and are used to draw generalisation. In this unit, we shall focus upon the descriptive statistical measures.

Note: Few sections of the Unit has been partially taken from Unit-14 'Measures of Central Tendency' and Unit -15 'Measures of Dispersion', BES-127 :Assessment for Learning, B.Ed. IGNOU, 2017.

The measures of descriptive statistics most commonly used in educational research are: measures of central tendency, measure of variability and measures of relationship.

In the present unit, we will discuss concept and methods of calculation of measures of central tendency. The next step discussed to know about measures of variability. Therefore, in the next section of this unit, we will discuss concept and methods of calculation of measures of variability and measures of position.

16.2 OBJECTIVES

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

- explain the meaning of measures of central tendency;
- compute mean, median and mode from ungrouped and grouped data and their interpretation;
- explain concept and importance of the measures of variability;
- define and calculate different measures of variability viz. – range, quartile deviation, mean deviation, and standard deviation;
- understand the measures of relative position;
- differentiate between percentile and percentile rank and
- calculate percentile and percentile rank.

16.3 MEASURES OF CENTRAL TENDENCIES

As you are aware that in the process of educational research, we get raw data about the variable under study after the administration of tools and techniques on the selected sample. The raw data is not able to provide any insight or meaning thus one has to use statistics to make it meaningful. The most commonly used statistics is the measures of central tendency.

Central tendency is the central value or average score of a group. **Measures of central tendency** help you to find the middle or the average of a data set. For example, if you have a score of 50 students on an achievement test. In order to understand who are the highest and the lowest scorer?, What is the difference between the highest and lowest scores? what is the average score of the group? and so on. To calculate this, you have to first make a frequency distribution table as explained in the previous unit and then analyze the nature of frequency distribution to answer these questions. For this, you may first measure the central tendency of the distribution or calculate the average or central score of that group. That will help you to compare individuals score with the average score of that group. So, measures of central tendency indicate the level of a group or sample.

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

- **Mode** is the most frequent value in the group.

- **Median** is the middle number in an ordered data set.
- **Mean** is the sum of all values divided by the total number of values.

In the following section, we will discuss about the meaning, procedure of calculation and uses of these measures.

16.3.1 Mean

Mean of a distribution is commonly known as arithmetic average. So, it is popularly known as 'Arithmetic Mean'. It is most familiar, easily understood average and most appropriate measure of a central tendency. When we calculate the mean value of each and every score of the group is considered, that's why Mean is called the most appropriate measure of central tendency. Mean is calculated when :

- The data is distributed symmetrically;
- Data is continuous;
- Data is presented in an interval scale;
- The other statistics are to be calculated later; and
- Focus is to know the center of the gravity of the sample.

Mean can be calculated by adding sum of the observations/scores and dividing it by the total number of cases. If $X_1, X_2, X_3, X_4, X_5, \dots, X_n$ are the N scores or observations. Then the formula of calculating Mean is as follows:

$$\text{Mean (M)} = \frac{X_1 + X_2 + X_3 + X_4 + X_5 + \dots + X_n}{N}$$

$$\text{Mean (M)} = \frac{\sum X}{N}$$

Where: $\sum X$ = Sum of all the scores/values/ observations

N = Total number of cases or size of sample

Mean can be calculated both in ungrouped data as well as grouped data.

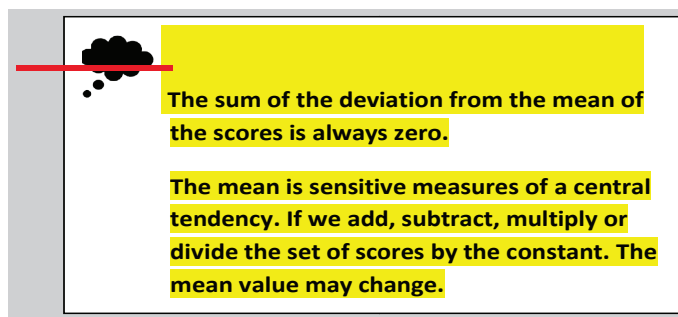
a) Calculation of Mean for Ungrouped Data

When sample size is small ($n < 30$) and data is not sorted or data is raw, we can calculate mean by adding all the raw scores and dividing by the total number of scores.

Example: The score of I.Q. of twelve participants is given as follows: 48, 52, 63, 56, 76, 98, 67, 49, 78, 92, 65, 84. Calculate the mean.

Solution:

$$\begin{aligned}\text{Mean (M)} &= \frac{48 + 52 + 63 + 56 + 76 + 98 + 67 + 49 + 78 + 92 + 65 + 84}{12} \\ &= \frac{828}{12} \\ &= 69\end{aligned}$$



The sum of the deviation from the mean of the scores is always zero.

The mean is sensitive measures of a central tendency. If we add, subtract, multiply or divide the set of scores by the constant. The mean value may change.

Sometime ungrouped data are also organized in a frequency distribution table, which does not involve class interval. Following formula is used for such frequency distribution.

$$\text{Mean (M)} = \frac{\sum fX}{N}$$

Where: $\sum fX$ = Sum of the products of scores and frequency

f = frequency

N = Total frequencies

Example: The scores of an achievement test of 30 students is given as follows. Calculate the mean of the scores .

Score	23	25	27	35	38	40	43	44	49
Frequencies	2	4	3	10	6	3	2	1	1

Solution:

Score (X)	Frequencies(f)	fX
23	2	46
25	4	100
27	3	81
35	10	350
38	6	228
40	3	120
43	2	86
44	1	44
49	1	49
	N=30	$\sum fX = 1104$

Note: These methods are based on an assumption-“the frequency of a class-interval concentrates on its mid-point”. It means that the cases fall in a class-interval score is equal to its mid-point. The class-interval (40-44) has seven frequency. Its mid-point is 42. It is assumed that each individual score is 42. Thus, the total scores obtained will be $42 \times 7 = 294$., while actual scores may be different or approximately will be the same. It is a mater of probability.

$$\begin{aligned}\text{Mean (M)} &= \frac{\sum fX}{N} \\ &= \frac{1104}{30} \\ &= 36.80\end{aligned}$$

b) Calculation of Mean for Grouped Data

This method is commonly used for large samples when the data is organized in frequency distribution table i.e. class intervals and frequencies. There are two methods for calculation of a mean from frequency distribution table.

- Long Method
- Short Method

Let us discuss both the methods one-by-one.

1. Long Method for Calculating Mean

Following formula is used to calculate mean by long method.

$$\text{Mean (M)} = \frac{\sum fX}{N}$$

Where:

- f = Frequency
- X = Mid-point of the class interval
- N = Total frequencies or Total number of cases
- $\sum fX$ = Sum of the products of mid-point and frequency

Following steps can be followed for calculation of Mean by long method:

Step 1: The mid-point of the class intervals are calculated by the following formula.

$$\text{Mid-point} = \frac{\text{Upper score} + \text{Lower score}}{2}$$

Step 2: The mid-point (X) is multiplied with the frequency of the respective class interval (fX).

Step 3: Sum of these products is obtained ($\sum fX$).

Step 4: Sum of products is divided by total number of frequencies.

Let us calculate Mean by long method by following these steps.

Example: Find out Mean of the distribution by using long method.

Class-Interval	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Frequency	5	7	9	12	8	4	5

Class Interval	Mid-Point (X)	Frequency(f)	fX
15-19	17	5	85
20-24	22	7	154
25-29	27	9	243
30-34	32	12	384
35-39	37	8	296
40-44	42	4	168
45-49	47	5	235
		N=50	$\Sigma fX=1565$

$$\begin{aligned}
 \text{Mean (M)} &= \frac{\Sigma fX}{N} \\
 &= \frac{1565}{50} \\
 &= 31.30
 \end{aligned}$$

The mean value is obtained to be 31.30.

2. Short Method for Calculating Mean

The long method has lengthy calculations. Therefore, short method has been developed for this purpose. It is relatively easier and involves less computation. In this method mean of the distribution is assumed to be any of the midpoint of the class intervals. It is also known as assumed mean method. Formula for assumed mean method (short method) is as follows:

$$\text{Mean (M)} = A.M. + \frac{\Sigma fd}{N} \times c.i.$$

Where:

A.M. = Assumed Mean

f = Frequencies

d = Deviation from assumed mean

N = Total number of frequencies

c.i. = Size of the Class Interval

Following steps is used for calculation of Mean in Assumed Mean Method:

Step 1 : Selection/identification of assumed mean

(Assumed mean is generally the mid-point of the class having highest frequency).

Step 2 : Calculate the mid-point of the Class

$$\left[\text{Mid Point} = \frac{\text{Lower Limit of the Class} + \text{Higher Limit of the Class}}{2} \right]$$

Step 3 : Calculation of deviation (d)

$$\left[\text{Deviation (d)} = \frac{\text{Mid point of the Class} - \text{A.M.}}{\text{Size of the Class}} \right]$$

Step 4 : Find out multiplications of frequency and corresponding deviation and place the obtained value in the column headed by fd.

Step 5 : Find the sum of the column, i.e. $\sum fd$

Step 6 : Apply the formula

Let us calculate Mean by using Assumed Mean method by the same example which has been used earlier in the long method :

Under this distribution assumed mean is considered the midpoint of (30-34) i.e. 32 .

Class Interval	Mid-Point (X)	Frequency(f)	d	fd
15-19	17	5	-3	-15
20-24	22	7	-2	-14
25-29	27	9	-1	-9
30-34	32	12	0	0
35-39	37	8	+1	+8
40-44	42	4	+2	+8
45-49	47	5	+3	+15
		N=50		$\sum fd = (-7)$

For calculation, the above mentioned steps are followed and the formula is applied.

$$\text{Mean (M)} = \text{A.M.} + \frac{\sum fd}{N} \times c.i.$$

A.M. = 32, f = 50, c.i. = 5, $\sum fd = (-7)$

$$\begin{aligned} \text{Mean (M)} &= 32 + \frac{(-7)}{50} \times 5 \\ &= 32 - 0.7 \\ &= 31.30 \end{aligned}$$

The mean value obtained is 31.30 by using both methods.

Check Your Progress 1

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

1. Define Mean.

.....

.....

.....

.....

2. Why mean is considered to be the most accurate and sensitive measures of central tendency?.

.....
.....
.....
.....

3. In which situations is calculation of mean suitable?

.....
.....
.....
.....

16.3.2 Median

Median is another measure of central location of a set of data. Median is the point or position while mean is the value. The median is a point on a scale in which half of the scores fall above it and half below it. So median is the central point from which 50 percent cases of the distribution fall above and 50 percent cases below the point. The median is the percentile 50 (P_{50}) and also second quartile (Q_2) of a frequency distribution.

Median is computed in the following situation when:

- scores are presented in ordinal scale (the interval scores are converted on ordinal scale);
- distribution is skewed;
- some scores are missing or incomplete distribution is given; and
- exact mid-point of the distribution is required.

There are two situations for computing median:

- a) Data in the ungrouped form
- b) Data in the grouped form

a) Calculation of Median for Ungrouped Data

For calculating Median is ungrouped data, we use a very simple formula i.e. $(N+1)/2$ th item in the series in order of ascending or descending order. There are two situations in computation of median from ungrouped data:

- When size of group (N) is odd number
- When size of group (N) is even number

The following steps are used for computing median from ungrouped data:

Step 1: The scores are arranged in order of size either in ascending or descending.

Step 2: The mid-point is calculated with the help of formula: mid-term = $(N+1) / 2$ th term

Step 3: Determine the median score in the series.

Step 4: If the group size is even number the median is obtained by taking the average of scores of above and below of $(N+1)/2$ th term.

Let's discuss calculation of Median through examples :

When size of group (N) is odd number: Suppose we have scores of nine students and scores 23, 14, 15, 32, 20, 13, 25, 27 and 18. The scores of these seven cases are arranged in ascending order for obtaining median.

Scores (in ascending order)-13, 14, 15, 18, 20, 23, 25, 27, 32

$$\text{Median} = \frac{N+1}{2}$$

$$= \frac{9+1}{2}$$

$$= \frac{10}{2}$$

$$= 5\text{th item}$$

$$= 20 \text{ is the median score}$$

5th item of the series in ascending or descending order is 20

When size of group (N) is even number: Suppose we have scores of 12 students and scores are 36, 23, 14, 15, 32, 20, 13, 25, 27, 18, 39, 11. The scores of these twelve cases are arranged in ascending order for obtaining median.

Scores (in ascending order) - 11, 13, 14, 15, 18, 20, 23, 25, 27, 32, 36, 39

$$\text{Median} = \frac{N+1}{2} \text{ th item in the series}$$

$$= \frac{12+1}{2} \text{ th item}$$

$$= \frac{13}{2} \text{ th item}$$

$$= 6.5^{\text{th}} \text{ item}$$

$$= \frac{6\text{th item} + 7\text{th item}}{2}$$

$$= \frac{20+23}{2} \text{ [Note: 6th item is 20 and 7th item is 23]}$$

$$= \frac{43}{2}$$

$$= 21.5$$

So, 21.5 is the median score.

b) Calculation of Median When Data are Grouped into frequency Distribution

Now you are aware that the Median is the mid-point of the series. For calculating

Median in grouped data, the assumption made is that ‘the frequency of a class interval is evenly or uniformly distributed over the range of class interval.’ We use the following formula for calculating Median in grouped data:

$$\text{Median} = L + \frac{\frac{N}{2} - cfb}{F} \times \text{c.i.}$$

Where :

L = Lower limit of the median class-interval

N = Total number of the cases

cfb = Cumulative Frequency below the median class-interval

F = Frequency of the median class-interval

c.i. = Size of the Class Interval

Steps followed for calculation of Median:

Step 1 : The frequency distribution table is converted into cumulative frequency table in an upward direction.

Step 2 : Determine the Median class interval which is the class interval that follows N/2th value in the median class-interval.

Step 3 : Calculate exact lower limit of the median class interval by subtracting 0.5 from the lower limit of the median class.

Step 4 : Find out the cumulative frequency below the Median Class (cfb).

Step 5 : Find out F; It is the exact frequency of the Median Class.

Step 6 : All the values are placed into the formula.

Example: Calculate median for the following frequency distribution.

Class Interval	65-69	60-64	55-59	50-54	45-49	40-44	35-39	30-34	25-29
Frequencies	3	5	7	8	10	7	6	2	2

Solution :

Class Interval (C.I.)	Frequencies (f)	Cumulative Frequency (cf)
65-69	3	50
60-64	5	47
55-59	7	42
50-54	8	35
45-49 (Median Class - interval)	10(fm)	27

40-44	7	17 (cfb)
35-39	6	10
30-34	2	4
25-29	2	2
	N = 50	

$$\text{Median} = L + \frac{\frac{N}{2} - cfb}{F} \times \text{c.i.}$$

Where: $N/2 = 25$ $L = 44.5$ $cfb = 17$ $F = 10$ $\text{c.i.} = 5$

$$\begin{aligned} \text{Median} &= 44.5 + \frac{25 - 17}{10} \times 5 \\ &= 44.5 + \frac{8}{10} \times 5 \\ &= 44.5 + 4 \\ &= 48.5 \end{aligned}$$

Check Your Progress 2

- Note: a) Write your answers in the space given below.
b) Compare your answers with those given at the end of the Unit.

4. Complete these sentences.

- The central point of frequency distribution is the
- Median is calculated when data is presented in scale.
- The median is the percentile of a frequency distribution.

5. In which situations is median computed?

.....
.....
.....
.....

6. Find out mean and median for the following ungrouped data .

18 23 45 11 35 26 41 35 19 38

.....
.....
.....
.....

16.3.3 MODE

Mode is the most frequently occurring observation or score in the distribution. In a data series any score which occurs most frequently is known as mode of distribution. On a bar chart, the mode is the highest bar. Mode is calculated

when:

- Data are in the nominal scale;
- Scores are marked like skewed distribution rather than normal distribution;
- Data is incomplete;
- Central tendency is estimated quickly and roughly.

Like mean and median, mode is also calculated for two situations.

- Data in the ungrouped form
- Data in the grouped form into a frequency distribution

a) Calculation of Mode for Ungrouped Data

Calculation of mode in an ungrouped data is known as empirical or crude mode. For ungrouped data, we only need to identify the observation which occurs maximum times. So, the crude mode is that single observation which occurs most frequently.

Mode = Observation with maximum frequency

For example in the data: 6, 8, 8, 9, 3, 4, 6, 7, 8, 3, 9, 8 the value 8 appears the maximum number of times.

Thus, mode = 8

Any data series / distribution may have one mode, or more than one mode. Depending upon the number of modes it can be called unimodal, bimodal, trimodal, or multimodal. Let us understand through this example.

Series A-	Series B-	Series C-
{4, 5, 4, 7, 8, 5, 9, 6, 3, 2, 4}	{4, 5, 4, 7, 8, 5, 9, 6, 8, 7, 10, 3, 2, 8, 4}	{4, 5, 4, 7, 8, 5, 9, 6, 8, 7, 10, 3, 7, 8, 4}
4 appears three times, all the rest appear only once.	4 appears three times, as does 8.	4, 7 and 8 appears three times.
Mode[A]={4}	Mode[B]={4, 8}	Mode[C]={4, 7, 8}
Series A has one mode. Therefore, it is unimodal series.	Series B has two modes. Therefore, it is bimodal series.	Series C has three modes. Therefore, it is trimodal series.

b) Calculation of Mode for Grouped Data

Calculation of Mode in group data is known as the true mode. It can be calculated by employing following methods:

- Indirect Method
- Direct Method

Indirect Method: If we have calculated mean and median value of the distribution, the mode can be calculated by using the following formula:

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

Let us take an example and calculate Mode in indirect method.

Example: If for some frequency distribution mean is 32.5 and median is 37. Calculate the mode.

$$\begin{aligned}\text{Mode} &= (3 \times \text{Median}) - (2 \times \text{Mean}) \\ &= (3 \times 35.5) - (2 \times 37) \\ &= 106.5 - 74 \\ &= 32.5 \text{ (Answer)}\end{aligned}$$

Mode in Direct Method:

In a grouped frequency distribution, the mode value can be easily obtained by observing the frequencies of the class-interval. The mid-point of the class-interval which has the highest frequency is mode value of the distribution. It is unrealistic to calculate the Mode by looking only at the frequencies. A direct technique therefore needs to be employed to calculate the Mode. For this, the following method can be used to calculate Mode.

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times \text{c.i.}$$

Where,

L = Lower Limit of the Modal Class

c.i. = Size of the class interval

f₁ = Frequency of the modal class

f₀ = Frequency of the class preceding of the modal class

f₂ = Frequency of the class succeeding of the modal class

(Source: NCERT, 2010)

Let us calculate Mode in direct method by using the above example.

Example: Calculate Mode of the following data in direct method.

Class Interval	65-69	60-64	55-59	50-54	45-49	40-44	35-39	30-34	25-29
Frequencies	3	5	7	8	10	7	6	2	2

Solution: In the above distribution highest frequency 10 falls in (45-49) class-interval, hence it is the mode class-interval. So its lower limit is 44.5 and f₁=10, f₀=7 & f₂=8 are frequencies of the modal, preceding and succeeding class-intervals respectively. These values are put in the formula to calculate mode value.

$$\begin{aligned}
 \text{Mode} &= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times \text{c.i.} \\
 &= 44.5 + \frac{10 - 7}{2 \times 10 - 7 - 8} \times 10 \\
 &= 44.5 + \frac{3}{5} \times 10 \\
 &= 44.5 + 6 \\
 &= 50.5 \text{ (Answer)}
 \end{aligned}$$

16.3.4 Comparison of Mean, Median and Mode

You have already studied the concept, processes of calculation, uses and limitations of Mean, Median and Mode in the previous sections. In this section, we will be comparing these three central tendencies.

Table 16.1: Comparison of Mean, Median and Mode

Aspects	Mean	Median	Mode
Nature of Data	Equal Interval Scale	Ordinal Scale	Nominal Scale
Concept	Mean is a point on the scale of measurement obtained by dividing the sum of all the scores by their total numbers.	Median is the mid-point of the distribution, below and above which 50 percent of the cases lie.	Mode is the score of the distribution which occurs most frequently.
Accuracy	Mean is the most accurate measures of central tendency.	Median is an approximate measures of central tendency.	Mode is a rough estimate of measures of central tendency.
Use	Only when all the scores of the distribution are known and distribution is symmetrical.	Can be calculated in case few scores are not known and if the distribution is skewed.	Can be calculated in case the absence of few scores and if the distribution is skewed.
Values	There will be only one Mean in the distribution.	There will be only one Median in the distribution.	There may be more than one Mode in the distribution.

(Source: 16.3.4 Comparison of Mean, Median and Mode has been adopted from BES-127, block 4, unit 14, pp53)

Check Your Progress 3

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

7. Define Mode.

.....

8. Write any three differences between mean, median and mode.

.....

9. In which situations is calculation of mode suitable?.

.....

10. What is the median of the following set of scores? 8, 19, 5, 12, 10, 14, 17

A. 19 B. 5 C. 12 D. 14

11. What is the mode of the following set of scores? 8, 19, 5, 12, 10, 8, 14, 17, 19, 8

A. 8 B. 5 C. 12 D. 17

12. The most frequently occurring number in a set of values is called the _____.

A. Mean B. Median C. Mode D. Range

13. The _____ is the value you calculate when you want the arithmetic average.

A. Mean B. Median C. Mode D. Range

16.4 MEASURES OF VARIABILITY

Measures of central tendency describe the typical value and do not provide a complete picture of the data. You may find different samples or populations which may have identical measures of central tendency yet differ from one another in other important ways. You may understand it through the following example:

Group A (Score)	Group B (Score)	Group C (Score)
10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 30	0, 0, 0, 0, 0, 0, 0, 0, 0, 80, 140	20, 20, 20, 20, 20, 20, 20, 20, 20, 20, 20
Mean = 20	Mean = 20	Mean = 20

From the above example, you might have noticed that there is no difference between three groups in terms of a mean of the sample groups but if you look closely at the data, you will find a variation within the groups. When you compare the group scores individually you find that group C is homogenous while group B is heterogeneous. Therefore, in order to get a complete picture of data, you need to know the dispersion or variation that exists between the data.

In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution. Measures of variability define how far away the data points tend to fall away from the center. We talk about variability in the context of a distribution of values. A low dispersion indicates that the data points tend to be clustered tightly around the center. High dispersion signifies that they tend to fall further away.

Just as there are multiple measures of central tendency similarly there are several measures of variability like

- 1) Range
- 2) Quartile Deviation
- 3) Mean Deviation
- 4) Standard Deviation

Let us discuss concept, properties and process of computation about all the four types of measures of variability one-by-one.

16.4.1 Range

The difference between the highest score and the lowest score of a set of data is the range. Range is computed when

- only total spread is required;
- information of extreme score is required; and
- Quick and crude measure are needed.

Range = (Highest Score-Lowest Score)

Let us calculate range of the following data:

45 65 23 43 46 78 98 105 56 67 76 74 49

Range of the distributing is $(105-23=82)$

It is very simple and quick measure of variability. Since, it is based on two extreme scores and does not give any information about the variation of intermediate or middle score. Therefore, it is not sensitive measure. It has been illustrated with the help of an example:

Series A 34 38 43 49 56 59 65 65 70 72 84

Series B 34 34 34 36 36 32 34 33 35 35 84

The range of these groups are as follows:

Group A Range = $84-34= 50$

Group B Range= 84–34=50

The range of both the groups are the same but group (A) is heterogenous and group (B) is homogenous while this nature is not revealed by the range.

Therefore, it is considered as a very crude measure of variability. If we change the middle scores high and low score are kept the same, the range may not change. Thus, it is not a sensitive measure of variability.

16.4.2 Quartile Deviation

In measure of variability Quartile Deviation is an improvement over the range. It is also known as ‘semi-interquartile range’, as the one half of range from Q_3 to Q_1 is used as measure of variability. As you have already acquainted with the interval or the difference between the highest and the lowest scores is known as ‘range’, in a similar way distance between first and third quartile divided by two is known as ‘**Quartile Deviation**’. Therefore, it may be expressed as:

$$QD \text{ or } Q = \frac{Q_3 - Q_1}{2}$$

The first Quartile (denoted by Q_1) or lower quartile has 25% of the items of the distribution below it and 75% of the items are greater than it. The third Quartile (denoted by Q_3) or upper quartile has 75% of the items of the distribution below it and 25% of the items above it. Thus, Q_1 and Q_3 denote the two limits within which central 50% or half of the data lies. So, the formula of quartile deviation can be expressed as follows:

$$QD = \frac{P_{75} - P_{25}}{2}$$

Q_1 and Q_3 are calculated by using the following formula:

$$Q_1 \text{ or } P_{25} = L + \left(\frac{\frac{N}{4} - Cf_B}{F} \right) c.i.$$

And

$$Q_3 \text{ or } P_{75} = L + \left(\frac{\frac{3N}{4} - Cf_B}{F} \right) c.i.$$

Note: The method of calculation of percentiles are basically the same as that used in calculating median. In calculation of median we find out $N/2$ term whether in computation of Q_1 and Q_3 we find out $N/4$ and $3N/4$ terms respectively.

Let's calculate quartile deviation for the following data.

Example:

Class Interval	65-69	60-64	55-59	50-54	45-49	40-44	35-39	30-34	25-29
Frequencies	3	5	7	8	10	7	6	2	2

Solution:

Class Interval (C.I.)	Frequencies (f)	Cumulative Frequency (cf)
65-69	3	50
60-64	5	47
55-59	7	42 (Q3)
50-54	8	35
45-49	10	27
40-44	7	17 (Q1)
35-39	6	10
30-34	2	4
25-29	2	2
	N = 50	

First, we calculate Q_1 or P_{25} :

$$Q_1 = L + \left(\frac{\frac{N}{4} - Cf_B}{F} \right) c.i$$

Where

L = Lower limit of class interval in which Q_1 falls = 39.5 ($N/4 = 50/4 = 12.5$ falls in class interval 40-44)

Cf_B = Cumulative Frequency below the $N/4$ class-interval = 10

F = Frequency of the $N/4$ class-interval = 7

c.i. = Size of the Class Interval = 5

Putting the values in the formula for Q_1

$$\begin{aligned}
 Q_1 &= 39.5 + \frac{12.5 - 10}{7} \times 5 \\
 &= 39.5 + \frac{2.5}{7} \times 5 \\
 &= 39.5 + 1.78 \\
 &= 41.28
 \end{aligned}$$

Similarly, Q_3 or P_{75} can calculate

$$Q_3 = L + \left(\frac{\frac{3N}{4} - Cf_B}{F} \right) c.i.$$

Where

L = Lower limit of the class interval in which Q_3 falls=54.5
($3N/4=3 \times 50/4=37.5$ falls in class interval 55-59)

Cf_B = Cumulative Frequency below the $3N/4$ class-interval=35

F = Frequency of the $3N/4$ class-interval=7

$c.i.$ = Size of the Class Interval=5

Putting the values in the formula for Q_3

$$\begin{aligned} Q_3 &= 54.5 + \frac{37.5 - 35}{7} \times 5 \\ &= 54.5 + \frac{2.5}{7} \times 5 \\ &= 54.5 + 1.78 \\ &= \mathbf{56.28} \end{aligned}$$

$$\text{Therefore, } QD = \frac{Q_3 - Q_1}{2} = \frac{56.28 - 41.28}{2} = 7.50$$

$$QD = 7.50$$

Interpretation of Quartile Deviation

Quartile deviation is easy to calculate and interpret. It is independent of the problem of extreme values and, therefore, it is more representative and authentic than range. In distribution where we prefer median as a measure of central tendency, the quartile deviation is also preferred as measure of dispersion. However, both the measures are not suitable to algebraic operations, because both do not consider all the values of the given distribution. In case of symmetrical distribution, mean and median are equal and median lies at an equal distance from the two quartiles i.e.

$$Q_3 - \text{Median} = \text{Median} - Q_1$$

In case of non-symmetrical distribution, two possibilities may arise:

I. $Q_3 - \text{Median} > \text{Median} - Q_1$ (Positive Skewed Curve)

II. $Q_3 - \text{Median} < \text{Median} - Q_1$ (Negative Skewed Curve)

Check Your Progress 4

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

14. What do you understand about measures of variability?

.....

15. 'Range is a very crude measure of variability'. Explain.

.....

16. Define Quartile Deviation.

.....

16.4.3 Mean Deviation

The quartile deviation does not provide the awareness about the variability of the total distribution. It represents the variation of the middle 50 percent scores. Therefore, mean deviation and standard deviation are improvement over the quartile deviation and range. Here, all the scores are considered in calculating mean and standard deviation. The mean deviation (MD), also known as average deviation (AD). The mean deviation or average deviation is the arithmetic mean of all the deviations from the mean. While calculating mean deviation, no account is taken of signs and, therefore, all deviations whether having plus or minus sign are treated as positive only.

The following formulae may be used to calculate the Mean Deviation (MD)

$$MD = \frac{\sum |x|}{N} \text{ (For ungrouped data)}$$

$$MD = \frac{\sum |fx|}{N} \text{ (For grouped data)}$$

Where:

X = Score

M = Mean

$|x| = |X - M|$ = Deviation of scores from their mean regardless of sign

N = Total number of observations or size of sample

$\sum f \cdot |x|$ = Sum of product of f and $|x|$ disregarding sign

Let us illustrate the computation method of mean deviation for grouped and ungrouped data:

A) Calculation of Mean Deviation for Ungrouped Data

Step 1: Calculate mean of the data

Step 2: Compute deviation as $(X-M)$

Step 3: Obtain sum of variation regardless of sign

Step 4: Divide modulus sum of deviation by number of observations

Example: Calculate the mean deviation for the following data:

X ; 25, 28, 20, 30, 35, 34, 28, 26, 16, 24

Solution: First we may compute the mean of the distribution:

$$\begin{aligned} \text{Mean} &= \frac{25 + 28 + 20 + 30 + 35 + 34 + 28 + 16 + 14 + 20}{10} \\ \text{Mean} &= \frac{240}{10} \\ &= 24 \end{aligned}$$

Score (X)	$x = X - M$	$ x $
25	1	1
28	4	4
20	-4	4
30	6	6
35	11	11
34	10	10
28	4	4
16	-8	8
14	-10	10
10	-14	14
		$\sum x = 72$

$$\begin{aligned} \text{MD} &= \frac{\sum |x|}{N} \\ &= \frac{72}{10} \\ &= 7.2 \end{aligned}$$

The mean deviation of the distribution is 7.20.

B) Calculation of Mean Deviation for Grouped Data

Example: Calculate the mean deviation of the following distribution:

Class-Interval	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Frequency	5	7	9	12	8	4	5

Solution:

Class Interval	Mid-Point (score)(X)	Frequency(f)	fX	Deviation $ x = X - M $ (M=31.30)*	$f \cdot x $
15-19	17	5	85	14.30	71.5
20-24	22	7	154	9.30	65.10
25-29	27	9	243	4.30	38.7
30-34	32	12	384	0.7	8.4
35-39	37	8	296	5.7	45.6
40-44	42	4	168	10.7	42.8
45-49	47	5	235	15.7	78.5
		N=50	$\Sigma fX=1565$		$\Sigma f \cdot x$ =350.6

$$MD = \frac{350.6}{50} = 7.01$$

The mean deviation is obtained to be 7.01. The following steps are used for calculating mean deviation:

Step 1: Mean is calculated of the data by a long method (calculation of mean score you have already learnt)

$$\begin{aligned} \text{Mean (M)} &= \frac{\Sigma fX}{N} \\ &= \frac{1565}{50} = 31.30 \end{aligned}$$

Step 2: Deviation from the mid-point and mean are calculated {i.e. (X-M)=(17-31.30)=(-14.3)}

Step 3: The deviation is multiplied by the corresponding frequency (f). The product is summed up without considering signs ($\Sigma f \cdot |x|$)

Step 4: The mean deviation is calculated by using the formula $MD = \frac{\Sigma |fx|}{N}$

In calculating mean deviation signs $\pm$ are ignored. This is not logical because upper and lower scores are given same treatment. So, the method does not fulfil the assumption of algebraic properties and hence it is not possible to use this method in higher statistics. Therefore, Standard deviation is an improvement over the mean deviation.

16.4.4 Standard Deviation

Standard deviation (SD) is commonly used as a measure of variability, because

it is the most consistent and stable index of variability. It is a positive square root of the mean of squared deviations from mean. The problem of signs which is there in the calculation of mean deviation is taken care by squaring the deviations. The commonly acknowledged symbol for the standard deviation is the Greek letter 'σ' (known as Sigma). Standard deviation is used when:

- We need the most reliable measure of variability.
- Mean is used as measure of central tendency.
- We want to compute higher statistics, viz. correlation, 't' ANOVA, ANCOVA, etc.

a) Calculation of Standard Deviation for Ungrouped Data

The standard deviation for ungrouped data is computed by using the formula

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

Where

x = deviation of raw score (X-M)

N = Total number of scores or observations

Let us illustrate the use of this formula for computing standard deviation of the following data:

X: 25, 28, 20, 30, 35, 34, 28, 16, 14, 20

Solution:

Step 1: First compute the mean of the distribution:

$$\text{Mean} = \frac{25 + 28 + 20 + 30 + 35 + 34 + 28 + 16 + 14 + 20}{10}$$

$$\begin{aligned}\text{Mean} &= \frac{240}{10} \\ &= 24\end{aligned}$$

Step 2: Calculate deviation of the scores from the mean

Step 3: Obtain squares of the deviations

Step 4: Calculate the sum of squares of the deviations.

Step 5: Put the values in the formula

Score (X)	$x = X - M$	x^2
25	1	1
28	4	16
20	-4	16

30	6	36
35	11	121
34	10	100
28	4	16
16	-8	64
14	-10	100
10	-14	196
		$\sum x^2 = 666$

Then we put the values in the following formula:

$$\sigma = \sqrt{\frac{\sum x^2}{N}}$$

$$\sigma = \sqrt{\frac{666}{10}} = \sqrt{66.6} = 8.16$$

b) Calculation of Standard Deviation for Grouped Data

Following are the formulas that are used to calculate S.D. in various situations.

i.)
$$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$

where $(x = X - M)$ (For grouped data)

(Note: This formula is generally used when value of mean is in whole number).

ii.)
$$\sigma = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

(Note: This formula is generally used when value of mean is in fraction).

iii.)
$$\sigma = \sqrt{\frac{\sum fd}{N} - \left(\frac{\sum fd}{N}\right)^2}$$

(Note: This formula is generally used as short method for calculating S.D.).

Where,

f : frequency

X : raw data

x : deviated value from mean

d : deviation from assumed mean

c.i.: class-interval

N : total frequency

Example : Calculate the standard deviation of the following distribution:

Class-Interval	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Frequency	5	7	9	12	8	4	5

Note: In subsection 16.3.1 we have calculated the mean for the same distribution. The mean is 31.30. As mean value (31.30) is not a whole number, therefore we are using formula number (iii) that is short method for calculating S.D.

Solution:

Class Interval	Mid-Point (X)	Frequency(f)	d	fd	fd ²
15-19	17	5	-3	-15	45
20-24	22	7	-2	-14	28
25-29	27	9	-1	-9	9
30-34	32	12	0	0	0
35-39	37	8	+1	+8	8
40-44	42	4	+2	+8	16
45-49	47	5	+3	+15	35
		N=50		Σfd=(-7)	Σfd²= 141

$$\begin{aligned}
 \sigma &= i \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \\
 &= 5 \sqrt{\frac{(141)}{50} - \left(\frac{-7}{50}\right)^2} \\
 &= 5 \sqrt{2.82 - (0.0196)} \\
 &= 5 \sqrt{2.80} \\
 &= 5 \times 1.67 \\
 &= 8.35
 \end{aligned}$$

To calculate S.D, we used the following steps:

Step 1: The assumed mean is taken as mid-point of middle class interval (AM=32).

Step 2: The deviation (d) is obtained by subtracting assumed mean from the midpoints of the class intervals.

Step3: The deviations(d) are multiplied by corresponding frequency of the class interval (fd)

Step 4: The (fd) is multiplied by corresponding (d) of the class interval to obtain fd²

Step 5: The sum of fd and fd² are calculated.

Step 6: The values are put in the formula.

Importance of S.D: The main purpose of standard deviation is to understand how spread out a data set is. A lower standard deviation means that the data is more clustered around the mean and hence the data set is more consistent. A high standard deviation means data has wide range of value. Standard deviation is the most popular and robust measure of spread. Standard deviation is the master measure of dispersion and amenable to algebraic operations. S.D. is also used in testing difference between two groups i.e. 't' test. The between and among difference is tested with the help of S.D.

Check Your Progress 5

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

17. In which conditions do we calculate standard deviation?

.....

18. Calculate mean deviation and standard deviation for the following set of data: 25, 28, 20, 30, 35, 34, 28, 16, 14, 20

.....

19. Write suitable words in the following sentences:

- The most accurate measure of variability is.....
- The crudest measure of variability is.....
- The middle 50 percent cases of variability indicated by.....
- are used along with the mean.

16.5 MEASURES OF RELATIVE POSITION

In the previous section we have discussed about the measure of central tendency and variability. The measures of central tendency and variability are used to analyze the nature of distribution of data. To some extent mean is used to interpret the score of an individual. The score may be above or below the mean of the data. General type of interpretation is made like average, above average or below average, but the specific place or rank in a group cannot be located.

For example: A teacher conducted a class test of 100 marks for 40 students and the mean of the students' scores is 70. Student A, B, C and D obtained 75, 65, 40 and 30 marks respectively. With the help of mean we can say only that students A and B performance are average and students C and D performances are below average. So, the interpretation is that performances of students A and B & that of students C and D is same even though there is 10 marks difference in scores.

Another example is that a student has scored 60 marks in Mathematics and 55 marks in Social Studies. The mean of Mathematics score is 70 and Social Studies mean is 40. It may be said that his/her performance in Mathematics is below average and in Social Studies is above average. So even he/she got higher marks in Mathematics as compared to Social Studies but his performance in the class is better in Social Studies. Now, we get only very superficial information about his/her performance. If we want to judge someone's ability in Mathematics and Social Studies, then we need to find his/her relative position in the group. May be by obtaining 55 marks in Social Studies he/she got first position in the class or may be by obtaining 60 marks in Mathematics he/she got last position in the class. Therefore, it is necessary to know that among all the scores in Social Studies, what is the position of the 55 marks and where it stands in the group. The place/position with regard to his/her performance is a more meaningful interpretation of his/her scores. The place or position or rank is calculated by transforming raw scores into percentiles.

16.5.1 Percentile and Percentile Rank

In measures of central tendencies, you have already studied about median. It is the midpoint of the series, also we can say that median is that point in a frequency distribution below and above which 50% of the cases lie. Similarly, first quartile i.e. Q_1 and third quartile i.e. Q_3 are those points below which lie 25% and 75% measures respectively. Similarly, you can calculate the points below which any percent of score e.g. 10%, 15%, 35%, 67%, 85%, 99%, etc. lies. These points are known as percentiles and we may represent the percentile using notations viz., P_{10} , P_{15} , P_{35} , P_{67} , P_{85} , P_{99} , etc. respectively.

The percentile is a relative value in a group of an individual in terms of percent. It locates the position of an individual in the group, in terms of percent of the individuals below or above him/her. We may define percentile and percentile rank in the following words:

Percentile	Percentile Rank
The percentile of a given distribution is the point on the score scale below which percent of scores fall.	The percentile rank of a given point on a score scale is the percentage of measures in the whole distribution which are below that given points.

16.5.2 Calculation of Percentile and Percentile Rank

a) Calculation of Percentile (P_k) in a Frequency Distribution

The method of calculation of percentiles are basically the almost same as that is used in calculating median.

The formula for calculating percentile is as follows:

$$P_p = L + \left(\frac{P_n - cf_B}{f_a} \right) \times i \quad \dots\dots(1)$$

Where,

P = Percentile

p = Percentage of distribution desired e.g. 20%, 35%, etc.

L = Exact lower limit of the class interval upon which P_p lies

P_n = Part of N

cf_B = Cumulative Frequency below L of the class-interval

f_a = Actual frequency in which the class-interval P_p falls

Example: Calculate P_{10} , P_{25} , P_{60} and P_{70} for the following data:

Solution:

Class Interval (C.I.)	Frequencies (f)	Cumulative Frequency (cf)
65-69	3	50
60-64	5	47
55-59	7	42
50-54	8	35
45-49	10	27
40-44	7	17
35-39	6	10
30-34	2	4
25-29	2	2
	$N = 50$	

$$P_{10} = 10\% \text{ of } 50 = 5 \rightarrow = 34.5 + \left(\frac{5-4}{6} \right) \times 5 = 35.33$$

$$P_{25} = 25\% \text{ of } 50 = 12.5 \rightarrow = 39.5 + \left(\frac{12.5-10}{7} \right) \times 5 = 41.28$$

$$P_{60} = 60\% \text{ of } 50 = 30 \rightarrow = 49.5 + \left(\frac{30-27}{20} \right) \times 5 = 50.25$$

$$P_{70} = 70\% \text{ of } 50 = 35 \rightarrow = 49.5 + \left(\frac{35-27}{8} \right) \times 5 = 54.50$$

The percentile points P_{10} , P_{25} , P_{60} and P_{70} are 35.33, 41.28, 50.25 and 54.50 respectively. It is also a transformation of raw scores 35, 41, 50 and 54 into percentile ranks 10, 25, 60 and 70 respectively.

b) Calculation of Percentile Rank (PR)

So far you have understood, how percentiles have been calculated directly from any given frequency distribution. Now we are going to calculate percentile rank (PR) of X , Y , Z scores or the position on a scale of 100 to which the individuals scores lie. The following formula may be used for calculating the PR :

$$PR = \frac{100}{N} X \left(Cf_B + \frac{X-I}{c.i.} f \right)$$

Where,

X = Score for which PR is to be calculated.

Cf_B = Cumulative Frequency below l

l = lower limit having X

f = frequency of C.I. having X

c.i. = Class Interval

N = Total frequency

Example

Calculate the PR for the score 42 for the distribution of scores given in the above table.

Solution: In the given table

$$X = 42, Cf_B = 10, l = 39.5, f = 7, i = 5, N = 50$$

$$\begin{aligned} \text{PR of } 42 &= \frac{100}{50} X \left(10 + \frac{(42 - 39.5)}{5} \times 7 \right) \\ &= 27 \\ \therefore \text{PR of } 42 &\text{ is } 27 \end{aligned}$$

In testing and interpreting the test scores, percentiles are very useful. Whenever you want to compare two individuals, it is always better and advised to compare them on the basis of their PR positions and not on the basis of scores they have obtained.

Another method for estimating percentiles and percentile ranks is the graphical presentation which is known as cumulative percentage frequency curve which has been earlier discussed in the unit 15 'Organization and Interpretation of Data' of this course.

Check Your Progress 6

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

20. Define percentile.

.....
.....
.....
.....

21. Differentiate between percentage and percentile.

.....
.....
.....
.....

16.6 LET US SUM UP

In this unit you have learnt about certain concepts and methods of calculating of central tendency, variability and percentile. **Central tendency** is the central value or score of a group. The three most common measures of central tendency are the mode, median, and mean. Mean of a distribution is commonly understood as arithmetic average. Median is the point while mean is the value. The median is a point on a scale in such a way that half of the scores fall above it and half below it. Mode is the most frequently occurring observation or score in the distribution.

In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution. Measures of variability define how far away the data points tend to fall from the center. Range, mean deviation, quartile deviation and standard deviation are the most common measures of variability. Standard deviation (SD) is commonly and most widely used to measure of variability, because SD is the most consistent and stable index of variability. Deviations of all scores taken from their mean and then square root of their average, is known as, standard deviation. The percentile is a relative value in a group of an individual in terms of percent. It locates the position of an individual in the group, in terms of percent of the individuals below or above him/her .

16.7 UNIT END EXERCISES

1. Differentiate between mean, median and mode.
2. Calculate Mean, Median Mode for the following distribution.

X	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
f	3	7	10	10	15	8	5	2

3. Calculate quartile deviation and standard deviation fir the following data.

X	30-39	40-49	50-59	60-69	70-79	80-89	90-99
f	1	4	14	20	22	12	2

4. Calculate P_{20} , P_{40} and P_{60} for the following distribution.

X	20-25	25-30	30-35	35-40	40-45	45-50	50-55	55-60
f	4	6	13	15	12	10	8	2

16.8 REFERENCES AND SUGGESTED READINGS

- Blommers, P. and Lindquist, E. F. (1959). Elementary Statistical Methods. Hough-ton Mifflin Company, Boston.
- Garrett, H. E. (1956). Elementary Statistics. Longmans, Green and Co., New York.
- Guilford, J. P. (1965). Fundamental Statistics in Psychology and Education. McGraw Hill Book Company, New York.
- Hannagan, T. J. (1982). Mastering Statistics, The Macmillan Press Ltd., Surrey.

- IGNOU. (2017). 'Measures of Central Tendency' (Unit -14) in Analyzing and Interpreting Learner Performance (Block-4). BES-127: Assessment for Learning, B.Ed. IGNOU
- IGNOU. (2017). 'Measures of Dispersion' (Unit -145) in Analyzing and Interpreting Learner Performance (Block-4). BES-127: Assessment for Learning, B.Ed. IGNOU
- Lindgren, B. W. (1975). Basic Ideas of Statistics, Macmillan Publishing Co. Inc., New York.
- Tate, M. W. (1955). Statistics in Education, The Macmillan Company, New York.

16.9 ANSWERS TO CHECK YOUR PROGRESS

1. Mean is the sum of all values divided by the total number of values.
2. When we calculate the mean value of each and every score of the group is considered, that is why Mean is called the most appropriate measure of central tendency.
3. Mean is calculated when:
 - The data is distributed symmetrically;
 - Data is continuous;
 - Data is presented in an interval scale;
 - The other statistics are to be calculated later; and
 - Focus is to know the center of the gravity of the sample.
4. (i) median (ii) ordinal (iii) 50
5. Median is computed in the following situation when:
 - scores are presented in ordinal scale;
 - distribution is skewed;
 - some scores are missing or incomplete distribution is given; and
 - exact mid-point of the distribution is required.
6. Self- Exercise
7. Mode is the most frequently occurring observation or score in the distribution. In a data series any score which occurs most frequently is known as mode of the distribution.
8. Refer 16.4.4
9. Mode is calculated when:
 - Data are in the nominal scale;
 - Scores are marked like skewed distribution rather than normal distribution ;
 - Data is incomplete;
 - Central tendency is estimated quickly and roughly.

10. C
11. A
12. C
13. A
14. Measures of variability define how far away the data points tend to fall away from the center.
15. Range is considered as a very crude measure of variability because if we change the middle scores high and low score are kept the same, the range may not change.
16. Distance between first and third quartile divided by two is known as 'Quartile Deviation'. It may be expressed as:

$$QD \text{ or } Q = \frac{Q_3 - Q_1}{2}$$

17. Standard deviation is used when:
- We need the most reliable measure of variability.
 - Mean is used as measure of central tendency.
 - We want to compute higher statistics, viz. correlation, 't' ANOVA, ANCOVA, etc.
18. Self-Exercise
19. a) standard deviation b) Range
c) Quartile deviation d) Mean deviation
20. The percentile of a given distribution is the point on the score scale below which percent of scores fall.
21. The percentile is a relative value in a group of an individual in term of percent. It locates the position of an individual in the group, in terms of percent of the individuals below or above him/her.

UNIT 17 ANALYSIS OF QUANTITATIVE DATA-II

Structure

- 17.1 Introduction
- 17.2 Objectives
- 17.3 The Concept of Correlation
- 17.4 Types of Correlation
- 17.5 Co-efficient of Correlation and Interpretation
- 17.6 Methods of Computing Co-efficient of Correlation
 - 17.6.1 Spearman's Rank Order Co-efficient of Correlation (ρ)
 - 17.6.2 Pearson's Product Moment Co-efficient of Correlation (r)
 - 17.6.3 Comparison between Rank Order Correlation and Product Moment Correlation
- 17.7 Normal Probability Curve
 - 17.7.1 Properties of Normal Probability Curve
 - 17.7.2 Deviation from Normalcy
 - 17.7.3 Applications of Normal Probability Curve
- 17.8 Let Us Sum Up
- 17.9 References and Suggested Readings
- 17.10 Answers to Check Your Progress

17.1 INTRODUCTION

In the previous units, we have discussed about the analysis of data in research study with reference to a single variable, primarily with techniques of calculating descriptive statistical measures like mean, median, mode, quartile deviation, standard deviation, percentile etc. The data in which we secure measures of one variable for each individual is called a **univariate** distribution. If we have measures on two variables of each individual in a study, the joint presentation of the two sets of scores is called a **bivariate** distribution.

Suppose in a research study you may want to study the relationship between two variables. For example, you want to analyse the relationship between height and weight, intelligence and achievement, environmental awareness and environmental behaviour, etc. So, you have to analyse, whether change in one will bring a change in another or not. You can do so by finding a relationship between them. In simple words, we can say that the statistical tool which helps us to study the relationships between two or more than two variables is called correlation. The study of correlation between two or more variables is important in research. So, in the first section of the unit you will study about the concept of correlation, its types and procedures for calculation of the co-efficient of correlation.

Note: Few sections of the Unit has been partially taken from Unit-16 'Correlation: Importance and Interpretation and Unit -17 'Nature of Distribution and Its Interpretation', BES-127 :Assessment for Learning, B.Ed. IGNOU, 2017

In the next section of this unit, we will discuss the concept, need and properties of the Normal Probability Curve (NPC). If we take any variable or any event of the nature whether educational, psychological or otherwise, it occurs on a definite pattern and follows a Natural Probability Theory in its occurrence. This pattern in graphical terms is known as Normal Probability Curve and is bell shaped.

17.2 OBJECTIVES

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

- define correlation and co-efficient of correlation;
- calculate co-efficient of correlation according to the nature of scores and their distribution;
- interpret the results obtained i.e. interpret the Co-efficient of correlation;
- explain the concept of Normal Distribution and Normal Probability Curve;
- describe the properties of the Normal Probability Curve;
- explain the concept of deviations from normalcy in data ; and
- apply the concept of normal distribution to research problems.

17.3 THE CONCEPT OF CORRELATION

If you observe your daily routine, you may realise that when we spend more time in walking/running/exercising then we tend to burn more calories. Similarly, during summer when the temperature goes up the ice-creme sale increases, as the height of mountain increases the temperature goes down. From these examples you can infer that there is some connection or association between the two pairs of traits or variables like time spent in running vs calories, temperature Vs ice-cream sale, height of mountain vs temperature. This type of association between variables is called correlation where the quantity of one variable increases/decreases as a result of it the quantity of other variable increases/decreases accordingly.

According to Tuttle, “Correlation is an analysis of the co-variation between two or more variables”. How with the change in one variable, brings change in the other related variables changes is known as correlation. In other words, we can say that the Correlation is a statistical technique that helps us to understand if there is any relationship between the two pairs of variables. It also helps us understand the direction of correlation and also the intensity of correlation i.e. how strong relationship is?

Correlation is only co-variation; it must be taken as causation. This means, it is not a change brought in variable by manipulating another variable. Therefore, the correlation should never be interpreted as cause-and-effect relationship. Thus, the correlation is simply a degree of relationship between the variables of a bivariate data (two variables). The correlation between the two variables X and Y simply means that when the value of one variable is found to change in one direction, the value of the other variable is found to change either in the same direction (i.e. positive change) or in the opposite direction (i.e. negative change), but change is certain.

17.4 TYPES OF CORRELATION

Correlation is usually classified into the following three types:

- (i) Positive Correlation
- (ii) Negative Correlation
- (iii) Zero Correlation

(i) **Positive Correlation:** If both the paired variables vary in the same direction with change, then, they are said to be positively correlated. It means that if one of them increases, the other related variables also increases and if one of them decreases, the other also decreases. Both of them have similar a direction of change. For example, height is positively correlated to weight, as the diameter of a circle increases so does its circumferences. The academic achievement is positively correlated with motivation or intelligence. Similarly, you may find a positive correlation between reading or study habits and achievement, attendance and achievement, etc. This positive correlation can further be explained with the following example.

Example: Following are the marks of eight students in science and maths subjects.

Student	Marks in Science	Marks in Maths
1	25	35
2	20	30
3	35	45
4	28	39
5	45	50
6	36	47
7	50	65
8	18	25

Now you can see that the marks of science and maths achieved by all the eight students are moving in the same direction i.e. increasing/decreasing in the same order. Thus, we say achievement of students in science and maths is positively correlated. Positive correlation can be expressed by the graphical presentation. If the data points make a straight line going from the origin out to high x- and y-values, then the variables are said to have a positive correlation.. Figure 17.1 shows an example of a perfect positive correlation between two variables. Perfect positive correlation means when the quantities of two variables increase and decrease in equal proportion in the same direction.

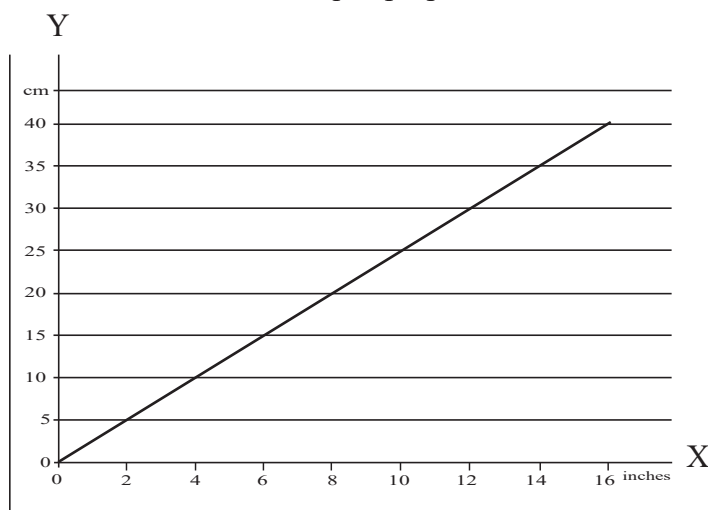


Figure 17.1: Perfect Positive Correlation between Two Variables

- (ii) **Negative Correlation:** If the two variables under study vary in the opposite direction with any change, they are said to be negatively correlated. It means that if one of them increases, the other related variable will decrease and if one of the related variable decreases, the other will increase. Both the variables follow opposite direction in change. For example, the supply and the prices of commodities. You know, if the commodity is available in abundance, the prices falls and if it is in shortage, the prices will rise. The same could be the case with the absentee and achievement of students. The more the students absent themselves from the school, the lesser will be the achievement and vice versa. This negative correlation can further be explained with the following example.

Example: Here the speed of a car and amount of time taken to reach a destination.

Speed (in KMs)	60	50	40	30	20	10
Time (in Minutes)	10	20	30	40	50	60

Now you can see that the speed and the time taken to reach the destination move in the opposite direction i.e. When one variable increases the other one decreases. Thus, we can say that speed and time are negatively correlated.

As we discussed earlier that correlation can be shown through graph. If the line goes from a high value on the y-axis down to a high value on the x-axis, the variables have a **negative correlation**. Figure 17.2 shows perfect negative correlation between two variables. In perfectly negative correlation quantities of two variable increases or decreases in the same proportion but in the opposite direction.

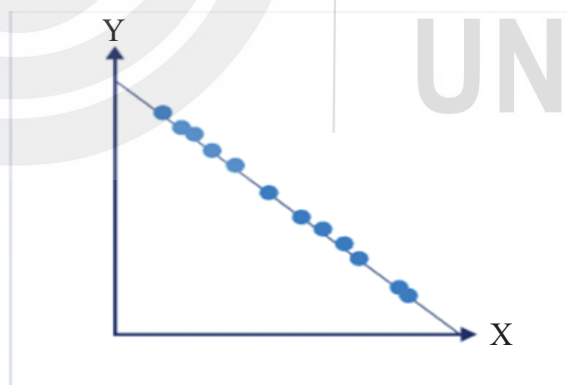


Figure 17.2: Perfect Negative Correlation between two Variables

- (iii) **Zero Correlation-**When there is no relationship observed or established between two variables under study i.e. the change in one variable cannot establish the change in other variable in a definite pattern. It may first increase time; second time decrease and third time may not change at all. In such cases, we can say no definite change or Zero correlation. For example, it is often seen that the child's intelligence and his/her weight do not affect each other. That is in general the intelligence neither increases with the weight of the body nor does it decrease. Thus, there is no relation between body weight and intelligence of a student. This type of relation is called zero relation.

Check Your Progress 1

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

1. What do you mean by the term correlation?

.....
.....
.....
.....

2. What do you understand by positive and negative correlation?

.....
.....
.....
.....

3. Suggest one example of each of the following:

a) Positive correlation.

.....
.....
.....
.....

b) Negative correlation.

.....
.....
.....
.....

c) Zero Correlation

.....
.....
.....
.....

17.5 CO-EFFICIENT OF CORRELATION AND INTERPRETATION

The degree of association or the degree of relationship between two variables is measured quantitatively in the form of an index which is termed as co-efficient of correlation.

Co-efficient of correlation is a single number that tells us to what extent the two variables are related and to what extent the variations in one variable change with the variations in the other.

It is determined to find the relationships because it is simple to understand and convenient to express. It is a constant value.

Symbol of co-efficient of correlation

The coefficient of correlation is always symbolized either by r or ρ (Rho). The notion ' r ' is known as product moment correlation co-efficient or Karl Pearson's Coefficient of Correlation. The symbol ' ρ ' (Rho) is known as Rank Order Correlation Coefficient or Spearman's Rank Correlation Coefficient. Sometimes it is also written as r_{xy} which means coefficient of correlation between x and y variables.

The measurement of the correlation between two variables results in a maximum value that ranges from -1 to $+1$, through zero. The ± 1 values denote the **perfect coefficient of correlation**.

Interpretation of Coefficient of Correlation

The coefficient of correlation is interpreted in verbal description. The following points should be considered while interpreting the correlation value :

- Firstly, the sign of the correlation coefficient should be taken into consideration. Positive sign tells us about the existence of a positive correlation between two variables and negative sign indicates a negative correlation.
- Secondly value of correlation coefficient may be understood as:

Value Correlation	Interpretation
± 1	Perfect Positive/negative Correlation
$\pm .90$ to $\pm .99$	Very High Positive/Negative Correlation
$\pm .70$ to $\pm .90$	High Positive/Negative Correlation
$\pm .50$ to $\pm .70$	Moderate Positive/Negative Correlation
$\pm .30$ to $\pm .50$	Low Positive/Negative Correlation
$\pm .10$ to $\pm .30$	Very low Positive/Negative Correlation
$\pm .00$ to $\pm .10$	Negligible Positive/Negative Correlation

Table17.1: Interpretation of Coefficient of Correlation

There are various types of correlation like rank order correlation, product moment correlation, biserial correlation etc. Suppose we calculate the correlation between the marks obtained in history and geography of a class of 40 students and the correlation coefficient was found to be $+ 0.17$. It means that there is positive correlation between the marks obtained by 40 students in history and geography and correlation is negligible. (using above table 17.1).

17.6 METHODS OF COMPUTING CO-EFFICIENT OF CORRELATION

In the case of ungrouped data of bivariate distribution, the following two methods are used to compute the value of co-efficient of correlation.

1. Rank order Co-efficient of Correlation or Spearman's Rank Order Co-efficient of Correlation.
2. Pearson's Product Moment Co-efficient of Correlation.

17.6.1 Spearman's Rank Order Coefficient of Correlation (ρ)

Rank Order Coefficient of Correlation is used only when the data is in an ordinal scale and it is calculated with the help of ranks. It may be defined as the correlation between the ranks assigned to the individuals in two traits or attributes. Individuals may be ranked in order of merit for traits like intelligence, creativity, achievement, adjustment etc. This is the most widely used method as it is easy to compute. It is applicable for sample size ($N < 30$) for quick results regarding the relationship of two variables. This is denoted by the symbol rho (ρ). Rank Order Coefficient of Correlation (ρ) is calculated by using the following formula.

$$(\rho) = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

Where (ρ) = Rank Order Correlation Coefficient

N = Total number of set of scores in either set or Number of pairs of scores.

D = Differences of ranks of each pair

In the rank order correlation method usually, scores are given and these are converted into ranks. The following procedure is used in assigning rank on the basis of scores.

- The number of ranks are equal to the number of cases or size of sample (N).
- The data, being discrete, is converted into Rank form.
- Start ranking from the highest to the lowest. Highest score is given rank value as 1 and the next is given the rank 2 and so on.
- When there are two individuals getting the same score, in that situation then their positional ranks are added and an average rank is given to each individual. For example, two students have scored highest score. The first and second two ranks are to be consumed, the average $(1+2)/2=1.5$ is assigned to each equal score. The third rank will be assigned the next lower score. Let's take some scores of students for better understanding.

Here score obtained by 8 students in a test are given:

Student	A	B	C	D	E	F	G	H
Score	11	12	14	16	14	18	15	10
Rank	7	6	4.5	2	4.5	1	3	8

In this example after 3rd rank student C and E have equal score 14, the next 4th and 5th rank are to be consumed, the average $(4+5)/2=4.5$ rank is assigned to each of the students i.e. students 'C' and 'E' and next 6th rank is assigned to the next lower score 12. The total ranks are 8 which is equal to the size of the sample.

- When three individuals get the same score, then their positional ranks are added and average is taken. For example, after 3rd rank three students have the same score. The next three ranks 4, 5 and 6 are to be consumed but the average $(4+5+6)/3=5^{\text{th}}$ rank is assigned to each student and equal score to

be assigned. The seventh rank will be assigned the next lower score.

Now let us calculate the coefficient of correlation

Example: A group of 10 students administered Maths and Science tests, their obtained scores are given in the following table. Calculate the coefficient of correlation.

Student	A	B	C	D	E	F	G	H	I	J
Maths Scores (X)	32	28	35	36	22	20	30	18	38	32
Science Scores(Y)	27	25	26	29	15	17	25	16	28	26

To find the rank order correlation coefficient among these two variables, let us find the ranks and form another table for computations.

Student	Maths Scores(X)	Science Scores(Y)	Rank Maths (R_1)	Rank Science (R_2)	$D = R_1 - R_2$	D^2
A	32	27	4.5	3	1.5	2.25
B	28	26	7	5	2	4
C	35	26	3	5	-2	4
D	36	29	2	1	1	1
E	22	15	8	10	-2	4
F	20	17	9	8	1	1
G	30	25	6	7	-1	1
H	18	16	10	9	1	1
I	38	28	1	2	-1	1
J	32	26	4.5	5	-0.5	0.25
						$\sum D^2 = 19.50$

Table 17.2 : Table for Computation of Rank Order Coefficient of Correlation

In this example following steps are followed in calculation of '(p)'.

- First find the ranks of Maths scores. Rank is allotted from highest to lowest scores i.e. 38 is the highest score so first rank is allotted to it and in the same way ranks are given. Whereas two students have scored same marks 32 thus ranks 4 and 5 each are to be consumed. The average rank 4.5 is assigned to both the students.
- Similarly find the ranks for science scores. Rank 1 is assigned to the highest score 29 then rank 2 to score 28 and in the same way ranks are allotted. Here three students have obtained the same score 25. Therefore, three ranks 4, 5 and 6 are to be consumed. The average of these ranks $(4+5+6)/3=5$ is assigned to all the three students.
- Now find the difference of Ranks of each of the score pairs for Maths and

$$\text{Science D} = R_1 - R_2$$

- iv. In the next column find the squares of the difference of ranks D^2 .
- v. Apply the values in the formula to find out rank order coefficient of correlation:

The formula is given here:

$$(\rho) = 1 - \frac{6 \sum D^2}{N(N^2 - 1)}$$

In this example

$$\sum D^2 = 19.50$$

$$N = 10$$

$$N^2 = 100$$

$$(\rho) = 1 - \frac{6 \times 19.50}{10(100 - 1)}$$

$$(\rho) = 1 - \frac{117.0}{990}$$

$$= 1 - 0.12$$

$$= 0.88$$

The rho- coefficient is 0.88. It shows very high positive correlation between maths and science performance of the students.

17.6.2 Pearson's Product Moment Coefficient of Correlation (r)

Pearson's Coefficient of correlation is determined when the data of variables consists of scores paired in some meaningful way. There are several ways of calculating 'r' in such cases.

(a) Deviation Score Method

In this method, coefficient of correlation is calculated on the basis of deviations of the scores from their means and is defined by the following formula.

$$r_{xy} = \frac{\sum xy}{n \sigma_x \sigma_y}$$

Where r_{xy} = Correlation coefficient between x and y

x = Deviation of first variable scores from their mean ($X - \bar{X}$)

y = Deviation of second variable scores from the mean ($Y - \bar{Y}$)

n = Total number of pairs of scores

σ_x = Standard Deviation of first variable scores

σ_y = Standard Deviation of second variable scores.

Example: A group of twelve students has been administered intelligence and creativity tests. Their obtained scores are given in the following table. Calculate Pearson's Product Moment Coefficient of Correlation (r) .

Student	A	B	C	D	E	F	G	H	I	J	K	L
Intelligence Quotient (I.Q.) Scores (X)	25	28	35	36	22	20	30	18	38	32	40	36
Creativity Scores(Y)	20	25	26	29	15	17	25	16	28	26	38	35

Solution:

Stu- dent	I.Q.(X)	Cre- ativ- ity (Y)	$x=(X-\bar{X})$ Deviation	$y=(Y-\bar{Y})$ Deviation	x^2	y^2	xy
A	25	20	-5	-5	25	25	25
B	28	25	-2	0	4	0	0
C	35	26	5	1	25	1	5
D	36	29	6	4	36	16	24
E	22	15	-8	-10	64	100	80
F	20	17	-10	-8	100	64	80
G	30	25	6	0	36	0	0
H	18	16	-12	-9	144	81	108
I	38	28	8	3	64	9	24
J	32	26	6	1	36	1	6
K	40	38	10	13	100	169	130
L	36	35	6	10	36	100	60
	$\sum X=360$	$\sum Y=300$		Total	$\sum x^2=670$	$\sum y^2=566$	$\sum xy=542$

Table 17.3: Table for Calculation of Product Moment Coefficient of Correlation by Deviation Score Method

In this method following steps are used,

Step 1: Calculate means of both the scores

Step 2: Calculate deviation of each score from means as $x=(X-\bar{X})$ and $y=(Y-\bar{Y})$

Step 3: Calculate deviation squares in each case and sum them up i.e. $\sum x^2$ and $\sum y^2$

Step 4: Calculate standard deviation of both the group of scores i.e. σ_x and σ_y

Step 5: Find 'xy' and their sum $\sum xy$

By substituting the corresponding values

$$\text{Mean } (\bar{X}) = \sum X/N = 360/12 = 30$$

$$\text{Mean } (\bar{Y}) = \sum Y/N = 300/12 = 25$$

$$\text{S.D.}(\sigma_x) = \sqrt{\frac{\sum x^2}{N}} = 7.47$$

$$\text{S.D.}(\sigma_y) = \sqrt{\frac{\sum y^2}{N}} = 6.86$$

$$\begin{aligned} r_{xy} &= \frac{\sum xy}{n\sigma_x\sigma_y} \\ &= 542 / (12 \times 7.47 \times 6.86) \\ &= 542 / 614.93 \\ &= 0.88 \end{aligned}$$

The person product moment of coefficient is 0.88. It shows very high positive correlation between Intelligence and Creativity.

(b) Computation of 'r' by Raw Score Method

In the raw score method you have to substitute the values of deviations in raw score form and value of standard deviations in raw score form then the following formula is used for the calculation of coefficient of correlation from raw scores:

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}}$$

Let us take an example for calculating 'r' using this formula and find out the coefficient of correlation. In following example, raw scores and the steps for calculating 'r' by Raw Score Method have been provided.

Example : Find the coefficient of correlation between Age and Glucose Level of ten students .

Table 17.4: Table for Computation of 'r' by Raw Score Method

Sub- ject	Age (X)	Glucose level (Y)	X ²	Y ²	XY
A	20	68	400	4624	1360
B	18	65	324	4225	1170
C	30	62	900	3844	1860
D	35	70	1225	4900	2450
E	40	80	1600	6400	3200
F	28	62	784	3844	1736
G	55	84	3025	7056	4620
H	62	99	3844	9801	6138
I	70	95	4900	9025	6650
J	60	88	3600	7744	5280
	$\sum X=418$	$\sum Y=717$	$\sum X^2=20602$	$\sum Y^2=61463$	$\sum XY=34464$

The following steps are used for calculating 'r' by the Raw Score Method

- (i) Find the square of the numbers in the X column, and put the result in the X^2 column.
- (ii) Find the square of the numbers in the Y column, and put the result in the Y^2 column.
- (iii) Multiply X and Y together to fill the XY column. For example, row 1 would be $20 \times 68 = 1360$
- (iv) Find sum of all raw scores of $\sum X$ and $\sum Y$; sum of squares of raw scores of $\sum X^2$ and $\sum Y^2$ then find sum of multiple of raw scores of each pair i.e. $\sum XY$
- (vi) Now compute Coefficient of Correlation by substituting each value in the given formula.

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{(N\sum X^2 - (\sum X)^2)(N\sum Y^2 - (\sum Y)^2)}}$$

$$r = \frac{10(34464) - (418)(717)}{\sqrt{(10(20602) - (418)^2)(10(61463) - (717)^2)}}$$

$$r = \frac{344640 - 299706}{\sqrt{[206020 - 17424][614630 - 514089]}}$$

$$r = \frac{44934}{\sqrt{[188596][100541]}}$$

$$= 44934/137701.23$$

$$= 0.32$$

The Pearson Product Moment Coefficient is 0.32. It shows low correlation between age and glucose level in individuals.

17.6.3 Comparison between Rank Order Correlation and Product Moment Correlation

You have already studied concept, processes of calculation of Rank Order Correlation and Product Moment Correlation in the earlier sections. In this section, we will compare both of them.

Table 17.5: Comparison between Rank Order Correlation and Product Moment Correlation

S.No.	Rank Order Correlation	Product Moment Correlation
1.	Denoted by (ρ) Rho Greek letter.	Denoted by r.
2.	Applicable for small size sample up to 30 only.	Applicable for small as well as large sample.
3.	Used in bivariate free distribution.	Used in bivariate linear distribution.

4.	Considered non-parametric technique of correlation	Considered as parametric technique of correlation.
5.	Requires data on an ordinal scale.	Used for data on an interval scale.
6.	Significance of Rho-correlation is not ascertained.	Significance of Pearson's coefficient is tested for statistical inferences.
7	Rho- correlation can not be used for regression and prediction.	Extensively used for prediction purpose by developing regression analysis.

Check Your Progress 2

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

4. Compute correlation of the marks of 10 students in rank order method.

Scores (Math) : 30, 37, 70, 45, 69. 55, 40, 45, 55, 63

Scores (English) : 33, 40, 64, 39, 52, 40, 36, 44, 52, 60

.....

.....

.....

5.. Compute correlation of the marks of 15 students in Pearson's product moment deviation score method.

Scores (Hindi) : 35, 37, 60, 40, 62. 55, 40, 43, 55, 50, 65, 62, 53, 45, 49

Scores (Social science) :33, 40, 61, 39, 52, 50, 36, 44, 50, 60, 70, 68, 55, 49, 45

.....

.....

.....

6. List any two differences between Rank Difference Correlation and Product Moment Correlation.

.....

.....

.....

17.7 NORMAL PROBABILITY CURVE

Let us look around our nature and we find that most of the things are distributed in a very predictable fashion. In the wildlife, we see, from the smallest creature like ants to very large creatures like elephants, hippopotamus in

the forest. But along with these two small and big categories, medium-sized creatures are also found in sufficient numbers. If we talk about human development, there is also a difference in the height, weight and size of people in the world. You might have observed that in your class, mostly boys or girls have average height and there are a few boys or girls with very tall or very short. Similarly, it is common to find majority of individuals with average intelligence level as compared to a few of them with high and low intelligence level. Similar trend can be observed in most of the events/ phenomena. You would find that maximum number of cases/frequencies is concentrated at the centre or central value of a distribution and frequencies gradually taper off symmetrically on both the sides of this value.

For better understanding, let's discuss this through an example. Look at the following hypothetical frequency distribution of height of 400 boys,

Height (in feet and inches)	Frequency
5.3	1
5.4	8
5.5	20
5.6	45
5.7	78
5.8	96
5.9	78
5.10	45
5.11	20
5.12	8
6.0	1

Table 17.6: Frequency Distribution of the Height of 400 boys

Let us draw the bar graph and the frequency polygon with the help of the above distribution, which looks graphs as shown in the figure 17.3 and 17.4 respectively.

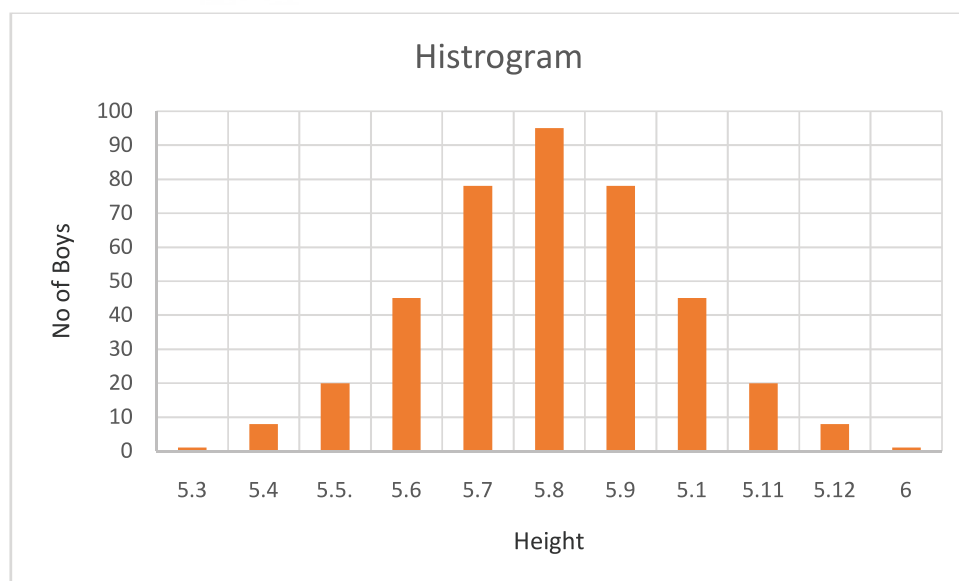


Figure 17.3: Bar Graph for the Data given in Table 17.6

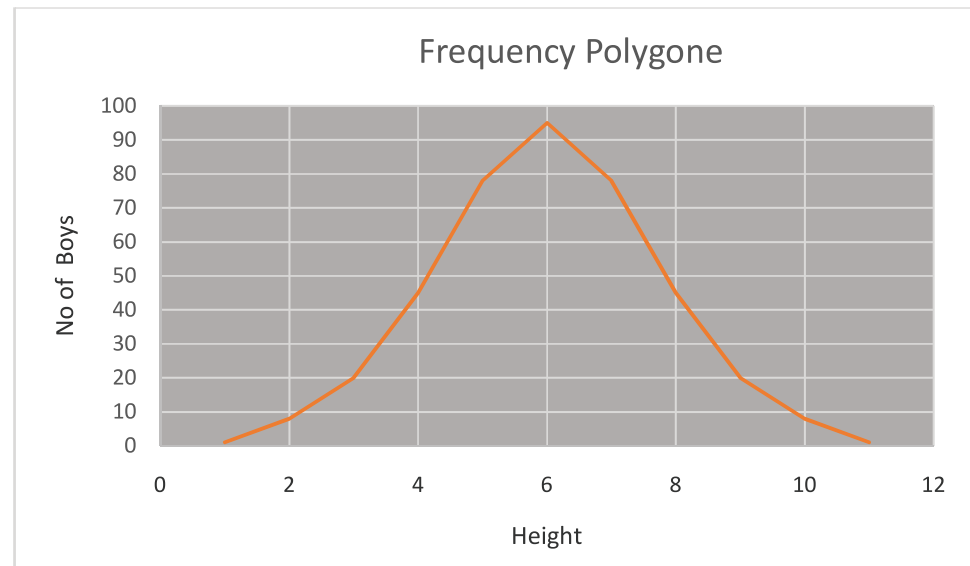


Figure 17.4: Frequency Polygon of the Data given in Table 17.6

By observing these graphs, we may say that most students' height is between 5 feet 7 inches to 5 feet 9 inches. The shape of the curve is just look like a 'Bell' and is symmetrical on both the sides.

If you compute the values of Mean, Median and Mode, you will find that these three central tendency value are approximately the same ($M = Md = Mo = 5.8$).

Meaning: The Normal Probability Curve is based upon the law of probability (the games of chance) discovered by French Mathematician Abraham Demoivre (1667-1754) in the eighteenth century. It is a 'Bell' shaped curve in which all the three measures of central tendency (Mean, Median, mode) fall at one point.

17.7.1 Properties of Normal Probability Curve (NPC)

By now you must have understood the meaning of normal distribution from the above discussion that it is a peculiar distribution and has its specific properties. Some of the properties of a normal distribution are:

- 1. The Curve is Symmetrical:** It is a symmetrical curve around the mean i.e., if we draw a curve with mean as a variant, then the number of cases above the mean value and the number of cases below the mean value will be equal. The curve will be symmetrical around the vertical axis. In other words the left and right values to the middle or central point are mirror images, as shown in figure 17.5.
- 2. The Bell Shaped Curve:** When we draw a normal probability curve for distribution, it is always a bell shaped curve irrespective of the nature of the variable for which it is being drawn as shown in figure 17.5.
- 3. The Normal Curve is Bilateral:** The 50% area of the curve lies to the left side of the maximum or central ordinate and 50% lies to the right side. Hence the curve is bilateral as shown in figure 17.5.
- 4. The Mean, Median and Mode at the Same Point:** The values of the mean, median and mode computed for a normal distribution have the same value.

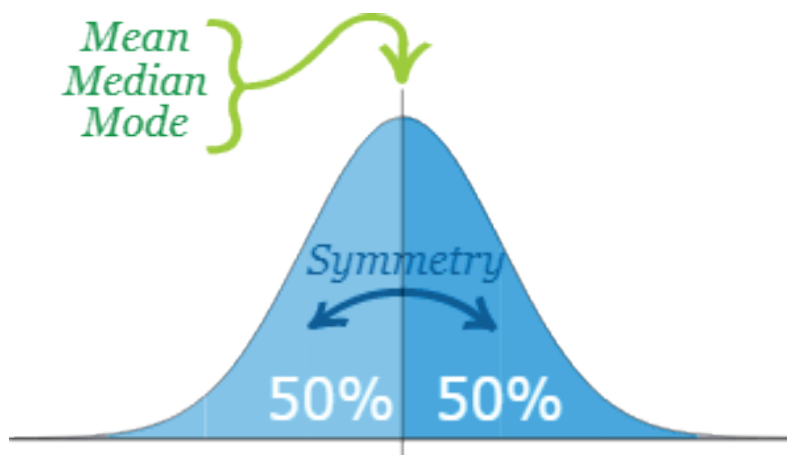


Figure 17.5: Bilateral Symmetry and Mean, Median and Mode Coincide in the Normal Probability Curve

5. **The Maximum Ordinate Occurs at the Centre:** The maximum height of the ordinate always occurs at the central point of the curve that is, at the mid-point. The ordinate at the mean is the highest ordinate and it is denoted by Y_0 . The height of the normal curve is maximum at the mid-point or the mean value.
6. **Equi-distant from the Mean/Median:** The first and the third quartiles (i.e., Q_1 and Q_3) are equidistant from the mean or median in a normal distribution. We can say that: $Q_3 - \text{Median} = \text{Median} - Q_1$

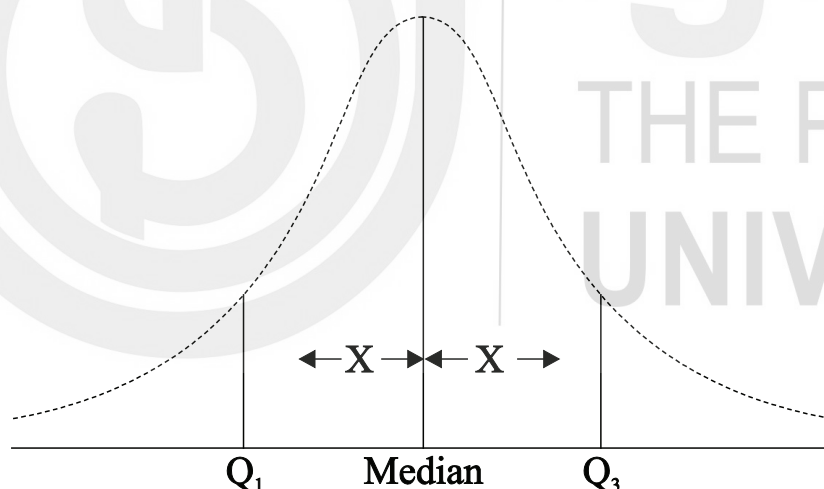


Figure 17.6 : Quartiles in the Normal Distribution Curve

If x is the distance between the third quartile and the median, then the distance between the median and the first quartile will also be the x .

7. **The Curve is Asymptotic:** The distribution is continuous and is never ending on either side. Hence the normal distribution curve has no boundaries in either direction and the curve never ends or touches the baseline, and also never touch it.
8. **It is Unimodal:** Since there is only one point where the frequency is maximum in a normal distribution (i.e., Mean), so it will always be a unimodal distribution. Normal distribution has only one mode.
9. **Normal Curve is a Smooth Curve:** It is a smooth curve and not a histogram. It is moderately peaked. The kurtosis of the normal curve is 0.263.

10 The Total Area of NPC is Divided into $\pm$ Standard Deviations: To find the deviations from the point of departure (mean), standard deviation of the distribution(σ) is used as a unit of measurement. The percentage of distribution area under a normal curve is given by the percentages given below in different cases. The percentages of area around mean are:

- i. Mean to + 1 S.D. (1σ) = 34.13%
Mean to - 1 S.D. (-1σ) = 34.13%
+ 1 σ to - 1 σ = 68.26%
- ii. Between Mean to + 2 S.D. (σ) = 47.72%
Mean to -2 S.D. (σ) = 47.72%
+ 2 σ to -2 σ = 95.44%
- iii. Mean to + 3 S.D. (σ) = 49.86%
Mean to -3 S.D. (σ) = 49.86%
+ 3 σ to -3 σ = 99.72%

In the following curve we have shown the areas under different conditions.

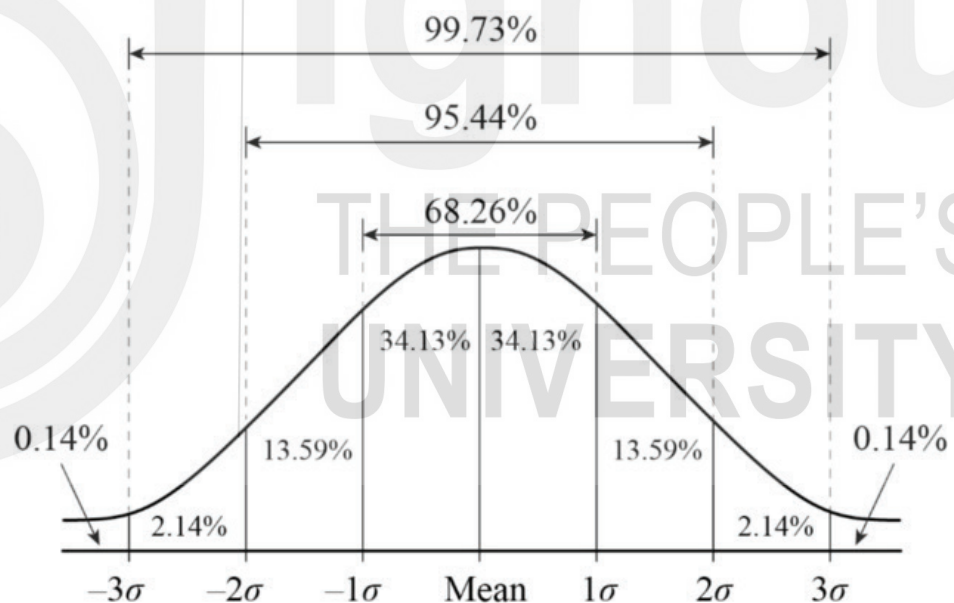


Figure 17.7 : Areas under Normal Probability Curve

The percentage of cases in a given normal curve is maximum around the centre i.e. around the mean and go on decreasing as you move away from the Mean. Secondly, more than two third cases, i.e., 68.26% of the total area of the curve fall between ± 1 standard deviation. If we find the number of cases between ± 2 standard deviation, they come out to be 95.44% and between ± 3 standard deviation are 99.72% of the cases.

11. Transform Raw Data to Standard Scores: A “Z” score results from conversion of a raw score into a number indicating how many standard deviation units the raw score is below or above the mean of the distribution. This makes use of the mean and standard deviation to describe a particular score. In a NPC the standard score Z is calculated using the formula:

$$z = \frac{x - \mu}{\sigma}$$

Where,

Z is the “z- score” or the standard score

X is the value which is to be standardized

μ is the standard mean of the distribution

σ is the standard deviation of the distribution.

17.7.2 Deviation from Normalcy

It is not that all the variables occur in the form of a normal distribution. There are two main ways in which a distribution can deviate from a normal distribution and are as follows:

(i) lack of symmetry (called skew) and (ii) pointiness (called kurtosis). The deviations from normalcy can be measured by:

Skewness: NPC is scattered symmetrically around its mean at the centre but sometimes the spread is not equal on both sides of the mean this is called skewness. In a perfectly normal distribution the skewness is zero. Skewness measures the symmetry of a normal distribution.

Skewness (sk) = $3(\text{Mean} - \text{Median}) / \text{Standard Deviation}$

The mean, median and mode coincide to give the NPC a perfect shape so that the left and the right sides of the curve are balanced. If these three statistics do not coincide in distribution then the symmetry is disturbed and the distribution appears to shift either to the left or to the right. When the scores are shifted at the lower side, the spread of the curve is gradually towards the right side, this is called **positive skewness** (look at figure 17.8). The **negative skewness** occurs when the scores are clustered towards the higher end of the distribution, consequently, the spread of the curve is more towards the left (look at the figure 17.8).

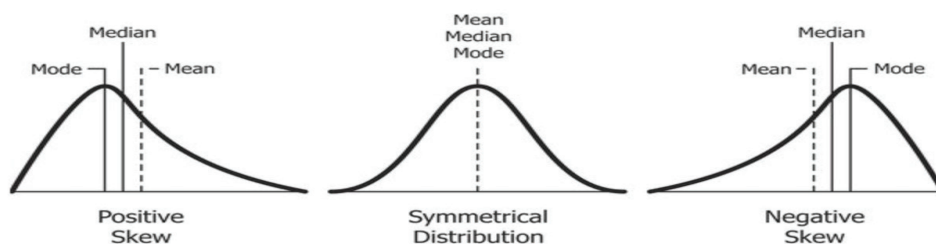


Figure17.8: Skewness in the Curves

Kurtosis: Kurtosis refers to how pointy a distribution is. It is the measure of whether the data are heavy-tailed or light-tailed relative to normal distribution. The NPC is a **mesokurtic** curve. Distributions are generally described as **platykurtic** (relatively flat), **leptokurtic** (relatively peaked), or somewhere in the middle i.e., mesokurtic as shown in figure 17.9. A distribution with positive kurtosis has many scores in the tails (so called

heavy-tailed distribution) and is pointy. This is known as leptokurtic distribution as shown in figure 17.9. In contrast, a distribution with negative kurtosis is relatively thin in the tails (has light tails) and tends to be flatter than normal. This distribution is called platykurtic as shown in figure 17.9. Ideally, we expect the our data to be normally distributed

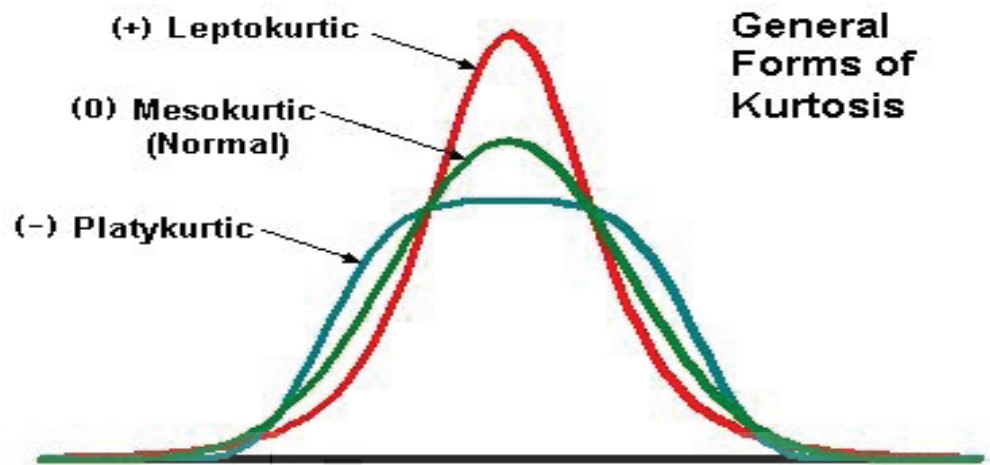


Figure 17.9 : Showing Forms of Kurtosis

17.7.3 Applications of Normal Probability Curve

Normal curve represents a model distribution. It can be used as a model to:

- **To compare various distributions:** It helps to find out whether the distribution is normal or not, if not, in what way it diverges from the normal. Statistical techniques are used for the analysis of data. Parametric test requires a normal distribution so before using parametric statistical techniques we have to find out whether data is normally distributed or not. In this case NPC is used as model/reference curve to compare distributions.
- **For grouping purposes :** In researches sometimes we have to categorise data in different groups based on some traits. A group of individuals may be conveniently grouped into certain categories as A, B, C, D, E (very good, good, average, poor, very poor) in terms of some trait (assumed to be normally distributed), with the help of a normal probability curve.
- **For determining the percentage of cases in a normal distribution within given limits:** By using NPC we can determine number of cases falls within given limit.

Example : In a normal distribution of 500 scores of the students with $M = 24$ and $\sigma = 8$, what percentage of cases lie between 16 and 32.

Here first of all we have to convert both the scores 16 and 32 into a standard score each.

$$Z = \frac{x}{\sigma}$$

x = deviation of the score (X-M)

σ = Standard deviation.

$$\text{Z Score for 16} = \frac{x}{\sigma} = \frac{16-24}{8} = \frac{-8}{8}$$

$$= -1\sigma$$

$$\text{Z score for 32} = \frac{x}{\sigma} = \frac{32-24}{8} = \frac{8}{8} = +1\sigma$$

In NPC, it is found that 34.13 cases fall between mean and -1σ and 34.13 cases fall between mean and $+1\sigma$. So $\pm\sigma$ covers 68.26% of cases. So that 68.25% of the cases fall between 16 and 32.

So, NPC helps us to study variables used in behavioural researches because they tend to be normally distributed, It also helps to classify the distribution in various categories and determine the percentage of cases that fall in a particular range. NPC have great significance and a number of applications in educational researches. Here we have discussed a few of them. You will learn about other uses of NPC in further studies.

Check Your Progress 3

Note: a) Write your answers in the space given below.

b) Compare your answers with those given at the end of the Unit.

7. Write any three characteristics of NPC.

.....

.....

.....

8. What do you understand by the term 'Skewness' and 'Kurtosis'?

.....

.....

.....

9. What is the importance of NPC in the educational researches?

.....

.....

.....

17.8 LET US SUM UP

You have learnt about the concept of correlation, the process of calculation of coefficient of correlation, main properties and applications of NPC.

Further, Correlation is simply a degree of relationship between the variable of bi-variate data. Co-efficient of correlation is a single number that tells us to what extent the two variables are related and to what extent the variations in one variable change with the variations in the other. The correlation is usually of three types: Positive Correlation, Negative Correlation and Zero Correlation. If we take any variable or any event of nature whether educational, psychological or otherwise, it is found to observe a definite pattern and follows a *Natural Probability Theory* in its occurrence. This pattern in graphical terms is known as Normal Probability Curve, which is bell shaped.

NPC helps us to study variables used in educational researches because they tend to be normally distributed. It helps to classify the distribution in various categories and also to determine the percentage of cases in a particular range. NPC has great significance and a number of applications in educational researches.

17.9 REFERENCES AND SUGGESTED READINGS

- Aggarwal, Y. P. (1990), 'Statistical Methods – Concepts, Applications and Computation', Sterling Publishers Pvt. Ltd., New Delhi.
- Anderson, Richard L. (1952), 'Statistical Theory in Research', McGraw hills Book Co., New York.
- Best John W and James V. Kohn (2004), 'Research in Education', Prentice Hall of India Pvt. Ltd., New Delhi.
- Downie, N. M. (1983), 'Basic Statistical Methods', HarperCollins College Div., New York.
- Ferguson, G. A. (1974), 'Statistical Analysis in Psychology and Education', McGraw Hill B.K. Co., New York.
- Garrett. H. E. & Woodworth, R. S.(1969), 'Statistics in Psychology and Education'. Vakils, Feffer& Simons Pvt. Ltd., Bombay.
- Guilford, J. P. & Benjamin, F.(1973), 'Fundamental Statistics in Psychology and Education', McGraw Hill B.K. Co., New York.
- IGNOU. (2017). Correlation: Importance and Interpretation' (Unit -16) in Analyzing and Interpreting Learner Performance (Block-4). BES-127: Assessment for Learning, B.Ed. IGNOU, 2017
- IGNOU. (2017). 'Nature of Distribution and Its Interpretations ' (Unit -145) in Analyzing and Interpreting Learner Performance (Block-4). BES-127: Assessment for Learning, B.Ed. IGNOU, 2017
- Kerlinger, F. N.(1993), 'Foundations of Behavioural Research: Educational and Psychological Inquiry'. Holt, Rinchart& Winston, New York.
- Loether, H. L. & McTavish, D. G. (1988), 'Descriptive and Inferential Statistics: An Introduction', 3rd ed. Boston: Allyn& Bacon.
- Sonedecor, G. W. & Cochran, W. G.(1967), 'Statistical Methods' (6th ed.), Iowa State University Press, Ames, Iowa.

17.10 ANSWERS TO CHECK YOUR PROGRESS

1. Correlation is the association or relationship between the variation of two variables.
2. Coefficient of correlation is a number that tells us the extent to which the two given variables are related or the change in one variable is accompanied by a change in the other variable.
3. A coefficient of correlation can vary from a value of $+1.00$ to -1.00 , through zero.
4. Self exercise
5. Self exercise
6. Refer 17.6.3
7. The normal curve is symmetrical. The normal curve is unimodal. Mean, median and mode coincide.
8. Skewness: NPC is scattered symmetrically around its mean at the centre but sometime the spread is not equal on both sides of the mean this is called skewness. In a perfectly normal distribution the skewness is zero.

Kurtosis: Kurtosis refers how pointy a distribution is. It is measure of whether the data are heavy-tailed or light-tailed relative to normal distribution. The NPC is a mesokurtic curve.
9. Refer 17.7.3