Indira Gandhi
National Open University
School of Education

BES-127
Assessment for Learning

Block

4

ANALYSING AND INTERPRETING LEARNER'S PERFORMANCE

Programme Coordinators: **Prof Saroj Pandey, (From April, 2017) and Dr. Gaurav Singh, SOE, IGNOU, New Delhi**

Course Coordinator: **Dr. Niradhar Dey, SOE, IGNOU, New Delhi**

COURSE PREPARATION TEAM

Course Writers

Unit 13 Mr. Ajith Ku. C.
SOE, IGNOU, New Delhi

Unit 14 Dr. Niradhar Dey
SOE, IGNOU, New Delhi

Unit 15 Dr. Asheesh Srivastava
Dept. of Education,
Visva-Bharati, Vinaya Bhavana

Unit 16 and 17 Prof. N.K. Gupta
Division of Educational Research,
NCERT, New Delhi

Content Editing
Prof. N. K. Dash
SOE, IGNOU, New Delhi

Language Editing
Prof. D. Venkateshwarlu
SOE, IGNOU, New Delhi

Proof Reading
Dr. Niradhar Dey
SOE, IGNOU, New Delhi

PRINT PRODUCTION

Prof. Saroj Pandey
Director
SOE, IGNOU, New Delhi

Mr. S.S. Venkatachalam
Assistant Registrar (Publication)
SOE, IGNOU, New Delhi

November, 2017

ISBN: 978-81-266-

Further information on the Indira Gandhi National Open University courses may be obtained from the University's Office at Maidan Garhi, New Delhi-110068.

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

Laser Typeset : Rajshree Computers, V-166A, Bhagwati Vihar, Uttam Ngr. (Near Sec.-2, Dwarka), N.D.59

Printed at :

BES 127	**ASSESSMENT FOR LEARNING**
BLOCK 1	**Evaluation in Teaching-Learning Process**
Unit 1	Concept and Purpose of Evaluation
Unit 2	Perspectives of Assessment
Unit 3	Approaches to Evaluation
Unit 4	Issues, Concerns and Trends in Assessment and Evaluation
Block 2	**Techniques and Tools of Assessment and Evaluation**
Unit 5	Techniques of Assessment and Evaluation
Unit 6	Criteria of a Good Tool
Unit 7	Tools for Assessment and Evaluation
Unit 8	ICT Based Assessment and Evaluation
Block 3	**Learner's Evaluation**
Unit 9	Teacher Made Achievement Tests
Unit 10	Commonly Used Tests in Schools
Unit 11	Identification of Learning Gaps and Corrective Measures
Unit 12	Continuous and Comprehensive Evaluation
Block 4	**Analysing and Interpreting Learner's Performance**
Unit 13	**Tabulation and Graphical Representation of Data**
Unit 14	**Measures of Central Tendency**
Unit 15	**Measures of Dispersion**
Unit 16	**Correlation: Importance and Interpretation**
Unit 17	**Nature of Distribution and Its Interpretation**

BLOCK 4 ANALYSING AND INTERPRETING LEARNER'S PERFORMANCE

Introduction to the Block

This Block consists of five units interpreting the performance of the learners quantitatively and qualitatively by using various statistical techniques. The Block begins with the Unit, '**Tabulation and Graphical Representation of Data'** (Unit-13) which deals with the conceptual discussions of the use of Educational Statistics in learner's assessment of performance. The Unit also eleborates on organization of data in frequency distribution and graphical presentation. The Unit also discusses various scales of measurement and its use in educational assessment.

Second Unit of this Block, titled, **'Measures of Central Tendency'** (Unit-14) elaborately discusses the techniques of measuring the central tendency, its interpretation and its use in teaching-learning process. The major statistics used for calculating central tendency such as Mean, Median and Mode have been discussed in this Unit with suitable examples. This Unit will help the learners to select appropriate statistics for calculating measures of central tendency.

The third Unit of this Block, entitled, **'Measures of Dispersion'** (Unit-15) which deals with the concept of measuring variability/dispersion of the data. The Unit, particularly 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 fourth Unit of this Block, **'Correlation : Importance and Interpretation'** (Unit-16) deals 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, product movement method, and scatter diagram correlation method have been discussed in this Unit. This Unit will particularly help you to calculate the degrees of relationship and interpret the performance of the learners.

The last Unit (Unit-17) of this Block, **'Nature of Distribution and Its Interpretation'** deals with the normal distribution and the distributions deviate from the normality such as Skewness and Kurtosis. As a teacher, you must have realized that the knowledge of an individual's position in the group and the need to categorise the group according to the level of ability of its nature is crucial. For such problems, the Normal Probability Curve is helpful for the teachers.

UNIT 13 TABULATION AND GRAPHICAL REPRESENTATION OF DATA

Structure

13.1 INTRODUCTION

In schools we deal with different kinds of data, of which the one pertaining to students is of utmost importance as it is utilized to assess and grade children on the basis of their performance. Apart from assessing the children, data serve numerous purposes such as comparing the learning performance among students, modifying teaching-learning activities, evaluating the student achievement and so on. So as to carry out these activities, data pertaining to children's performance need to be collected. Generally data are collected in the form of numerical or alphabetical form. The process of collecting, organizing, interpreting and analyzing the data is termed as "statistics". In this unit, we shall discuss the uses of educational statistics in assessment, various kinds of data, method of organizing them and the different scales of measurement.

13.2 OBJECTIVES

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

- describe the term statistics and its use in assessing and evaluating children's performance;
- state the meaning and nature of data;

- explain the importance of data organisation;
- present data in frequency tables;
- compare various methods of representing data;
- draw various types of graphs;
- explain the use of different graphical methods of representing data; and
- recognize scales of measurement.

13.3 USE OF EDUCATIONAL STATISTICS IN ASSESSMENT AND EVALUATION

Imagine an eighth standard classroom wherein 50 students got admitted in the year 2014, 48 in 2015 and 46 in the year 2016. The class teacher of that particular class organized the same data in a tabular form as shown in the Table 13.1.

Table 13.1: Year-wise admitted students

Admitted Year	No. of Students
2014	50
2015	48
2016	46

What can we understand from it? Generally the schools or any educational organization deal with different kinds of data related to students and administrative aspects, and such data need to be collected, organized and interpreted to make valuable decisions that have long term effect. Many a times, the schools find difficult to deal with large volume of data concerning their organizations and in such situations, the mathematical technique, 'statistics' help them deal with the data. **The mathematical process of collecting, organizing, interpreting and analyzing the data are termed as "statistics".** In statistics, data related to any individual/organization/behavior etc. are expressed in numerical form.

The word statistics is derived from the Latin word 'Status', Italian word 'Statista', German word 'Statistik', and French word 'Statiique' , which all means 'political state'. It provides data concerning the various attributes of state/country that help in successful administration.

Seligman defines "**Statistics is the science which deals with the methods of collecting, classifying, presenting, comparing and interpreting numerical data collected through some light on any sphere of enquiry**" (quoted from Pilai, 2008). In fact, educationists and psychologists use statistics widely to study human behavior. At the same time, statistics also help a teacher analyze and judge students' performance. In this Unit, we will study statistics in the context of assessment and evaluation. Before that, let us recapitulate some basic concepts of statistics. There are five types of statistics, which are described in Figure 13.1.

Descriptive Statistics: It involves the methods of collecting, classifying, organizing, picturing and summarizing observations or information's from data and generating statistical interpretations/ conclusions about the data. It is concerned about particular group of individuals.

Inferential Statistics: It involves the methods of using information from samples to arrive at conclusions about the population. Thus inferential statistics is concerned about samples selected form the population.

Inductive Statistics: It involves the method of generating statistical interpretations about one group/individual based on the interpretations made on another group/individual. It is sometimes called predictive statistics.

Analytical Statistics: It involves methods of arriving statistical interpretations comparing two or more sets of data.

Applied Statistics: It involves the application of statistics to study real life situations such as educational growth, economic status, political matters, etc

Figure 13.1: Types of Statistics

Statistics is used in many subjects like Mathematics, Science, Psychology, Commerce, Economics, Social Sciences, Geography, Agriculture, Business Management, etc. Being a teacher trainee, you are more concerned with its use in education. Can you identify few situations where statistics is made use of in education? Why don't you try the activity given below for better understanding?

Activity 1

Select a topic of your choice and teach it using two methods, say constructivist method and concept attainment model. Conduct an achievement test in both cases and analyse the results. Which method do you find effective? Prepare a report on the activity conducted.

....

....

....

....

....

....

....

....

What do you observe as you complete the activity? How will you draw conclusion about the achievement of the students taught through two methods? Definitely, in order to draw conclusion, you might require the use of statistics. In education, you may find number of instances, where statistics is profusely used. This illustrates the significance of statistics in education. In the activity given above, you may use 'correlation', a statistical concept to make judgment on the effectiveness of teaching methods. Similarly, statistical techniques like mean, median, mode, standard deviation, etc are widely used in analyzing and interpreting data. A few such applications are discussed below. These concepts/ techniques have been discussed in the further Units (Units 14, 15 & 16) of this Block.

- 'Raw scores' of children are converted to 'standard scores' to formulate inferences. At times even raw scores are also used to arrive at conclusion.
- 'Mean', 'median', 'mode' are used to find the average of observations.
- 'Range', 'Quartile deviation', 'Average deviation' and 'Standard deviation' are used to find the extent of variation from the mean scores of observations.
- 'Normal probability curve' helps us to understand the nature of particular group of students.
- 'Correlation' helps to find out the relationship between the groups with regard to certain variable.
- 'Analysis of variance' again helps teacher to compare performance of students and 'critical ratio' between means and test the significance of difference.

By this time, you understand that statistics plays a pivotal role in education. Without statistics, it is difficult for the teacher to interpret and judge the learning performance of children. Not only learning performance, performance of students in co-curricular/extracurricular activities analysed and interpreted using statistics. For example during sports meet, running time of students participating in 100m race marked in numerical terms. As we know, assessment and evaluation are two critical aspects of teaching –learning process. Assessment is the process of collecting scores and evaluation is the interpretation of that score. For example, a teacher conducts a unit test for a particular unit. The marks against each student depict the assessment while assigning them rank is evaluation. The whole process involved is statistics. The following are the utility of statistics in assessment and evaluation.

- It helps teacher analyse and interpret scores. Usually raw scores are collected from students performance in examinations, drawing test, sports activities, etc. Raw scores itself has no meaning and they are converted to standard scores/derived scores and interpreted to judge the students performance.
- It helps teachers compare the scores of different groups within the school or outside. A teacher can evaluate the achievement of students belonging to different classes using statistical procedures. Comparison can also be done among different organisations.
- It helps teacher construct standardized achievement test. The statistical procedures are followed at various stages of test construction, especially during item analysis.
- Statistics helps to determine the individual differences among children. As we are aware, no two individuals are alike and they differ in intelligence,

aptitude, attitude, personality, etc. Individual differences are measured with the help of various tests and the result is interpreted using statistical techniques.

- When the standardised tests are administered on children, the results obtained are analysed and used for providing counselling and guidance services.

- It helps a teacher predict the future performance of children. As discussed each individual is different from the other. A child who has scored high marks in an intelligence test may perform well in other fields. Accordingly the teacher would be able to predict his/her future career.

- It helps a teacher make selection, categorisation and promotion of students. The tests conducted at school level serve different purposes such as for promoting children to higher classes, for selecting to various arts/sports activities, and for categorising them based on their performance.

- It helps a teacher compare the functioning and working of his/her organisation with that of other. We have observed certain schools performing better compared to others which is revealed through statistical analysis.

- One of the most important applications of statistics is its use in educational research.

Let us attempt a question now.

Check Your Progress 1

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

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

1. What are the uses of educational statistics in assessment and evaluation?

....

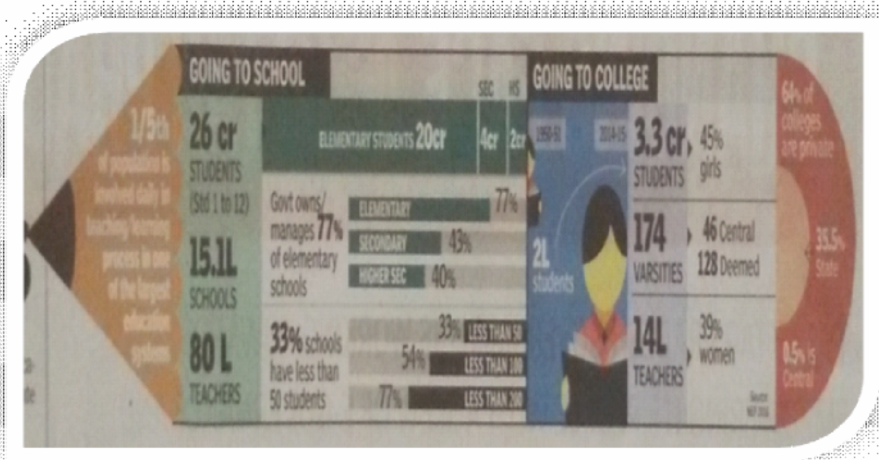
....

....

....

13.4 MEANING AND NATURE OF DATA

Joseph, a secondary school teacher was going through the newspaper 'Times of India' (June 18, 2016) and found interesting news about the 'draft education policy'. It says that, there are 26 crore students (1 to 12 standard), 15.1 lakh schools and 80 lakh teachers. The newspaper also had information related to other aspects of education, as shown in the figure 13.2. What does it imply? In daily life, people obtain different information either through newspapers, television, internet and other media. It is from such information meaningful interpretations are made.



Source: Times of India, (June 18, 2016)

Figure 13.2: Data in News Paper

Let us take another example. A teacher of eighth standard marks the attendance of children present on a particular day (the total number of students is 55). She asks children to raise their hands as she calls out their roll numbers. While she calls out roll numbers 3, 34 and 42 none of them raise their hands. In such a situation, the teacher would presume that, these children are absent on that particular day. So we have the information that the classroom mentioned is eighth standard, there are a total of 55 students and three students are absent. She can further enquire about the reasons of absenteeism. This is how interpretations are made from the available information. So whether it is newspaper, television or classroom, the information are collected and valid interpretations are generated. So in simple terms, the facts concerning situations/ individual/group from which conclusions are drawn are termed as data.

Data is the plural form of the word 'datum' which means 'fact'

Evidence or fact which describes group or a situation and from which conclusion is drawn is called data (Biswal & Dash, 2009). In statistics data represents any kind of information obtained from an experiment, observation, interview, or through any investigatory procedures. In the context of schools, the information about the total number of students, the number of teachers, the periods allotted for each subject, number of absentees each day, the marks scored by children in term end examinations and assignments, participation in co-curricular activities, mode of transportation in reaching schools, etc. represents data. These information (or data) would help a teacher in many ways such as, to judge her children's learning performance, recognize talents of children in extra curricular activities providing guidance and counselling services and so on. At this point, attempt the activity given below, and formulate a conclusion.

Activity 2

Distribute the following sheet to your peers and ask them to fill it. What conclusions would you about the socio-economic status of the peers arrive at?

S. No	**Name of the Peer**	**Occupation**		**Annual Income**		**Education**	
		Father	Mother	Father	Mother	Father	Mother

Let us explore a few more basic concepts of statistics and data in special. Data is any kind of information from which conclusions are made. Data are classified into two; primary data and secondary data. A brief description of them is given in figure 13.3:

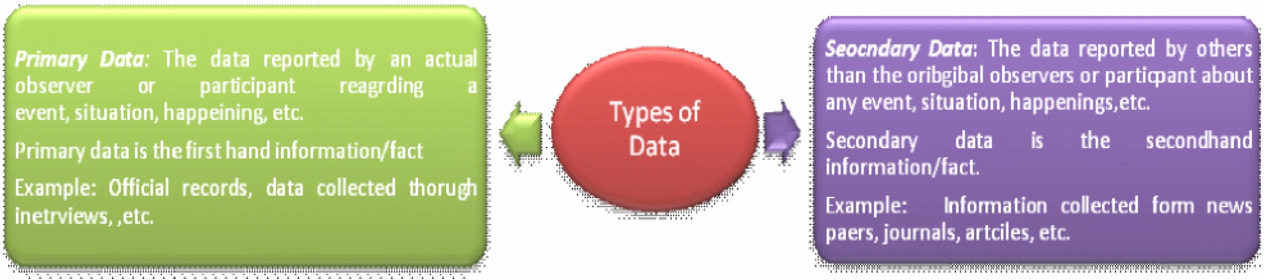


Figure 13.3 : Types of Data

Suppose you are planning to select 20 students from a total strength of 400, studying in seventh standard, to be trained for a drama competition for state level. In such a situation we have two sets of students i.e. the total 400 students and 20 students to be selected from the 400. You may come across similar situations action researches/educational researches. In educational research, two basic concepts; population and sample are very important. Let us briefly understand these two concepts from Figure 13.4.

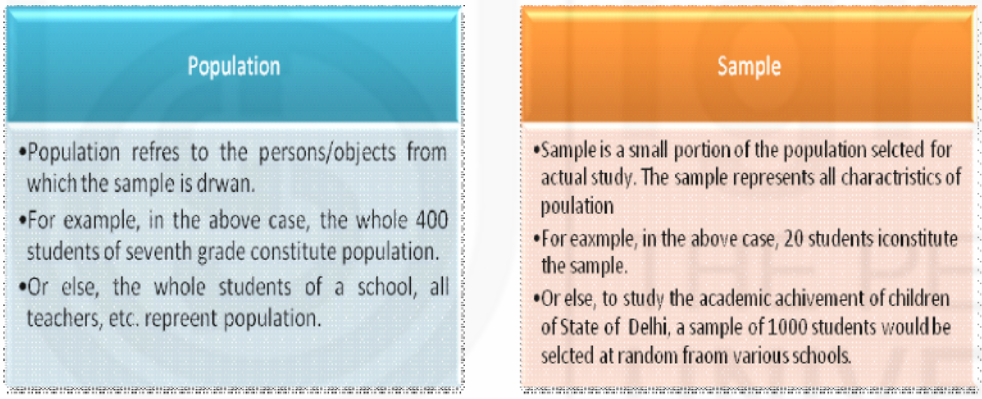


Figure 13.4 : Population and Sample

Having understood population and sample, let us explore the concept "score", a common term, frequently used in educational research, statistics and school settings. Score refers to the numerical description of the performance of any test, for example, the score secured by a student in social science term-end examination. There are two types of scores. They are described in Figure 13.5.

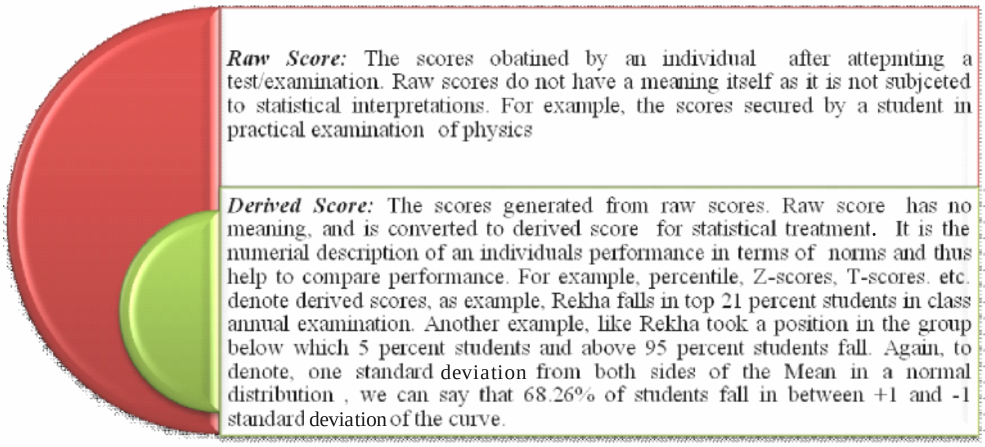


Figure 13.5: Raw and Derived Score

Check Your Progress 2

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

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

2. Differentiate primary and secondary data.

....

....

....

13.4.1 Discrete/Raw and Group Data

We have already discussed that data are the evidences or facts that describe a person/group from which conclusions are drawn. Generally the data are collected from the population. For example, number of students who failed in Class – X examination during the academic year 2015. Here the number of students failed (say for example 23 students) represents the data. Apart from primary and secondary data, data could be qualitative or quantitative; or continuous or discrete. Let understand each of these concepts.

Series: Series is the sequence of number that has some relationship with each other. For example, the weight of students may be 40kg, 41.5, kg and so on.

Qualitative Data/Series: The data (facts, Items. events, persons, phenomena, etc) expressed in qualitative terms are called qualitative data. Qualitative data are not measurable on a scale. For example, gender of students, type of school, mode of transportation of children, etc.

Quantitative Data/ Series: The data expressed in numerical format are called quantitative data. Such data are measurable and countable. For example, the data showing the children who passed tenth grade, children's attendance in a particular day, etc.

Continuous Series/Data: The data expressed in a sequence form are called continuous data. There will not be any gap in between the numbers. Continuous data are expressed as fractions. For example, the height of the children may be 5'7" or 5'4" and so on.

Discrete Series/Data: If the data expressed have gaps in between, such data are called discrete data. Discrete data are represented as whole numbers and not as fractions. For example, number of children in a particular class, number of periods in a day, etc.

13.5 ORGANIZATION/GROUPING OF DATA: IMPORTANCE OF DATA ORGANIZATION AND FREQUENCY DISTRIBUTION TABLE

You may recall the annual sports meet of your school. Suppose in the long jump event for junior girls, there were 15 students and each student got three chances. The competition was organized in such a manner that, each student attempted their first chance and the same was repeated till each participant completed their third chance. Rajendra Kumar, the mathematics teacher, was entrusted with the responsibility of entering the attempts made by each student. Let us see how he had done it.

1) Student 1 : 3.2m, 3.2m & 3.3m
2) Student 2 : 3.5m, 3.4m & 3.5m
3) Student 3 : 3.4m, 3.4m & 3.3m
4) Student 4 : 3.6m, 3.7m & 3.5m
5) ……………………
6) ……………………

Rajedra Kumar continued marking attempts made by 15 students. Now the question is, how will you determine the winner of the competition? The answer is very simple. Just look for the student, who had jumped maximum distance. But the question is, will you be able to identify the winner so quickly. The answer may probably be ‘no’. Why is it so? In this case, Rajendra Kumar had marked student attempts in such a way that, it is difficult for someone to interpret so quickly. Here you may feel that, the winner can be easily identified as the number of students are comparatively less, but think of a situation, where there are large numbers of participants. So, there arises the need for organizing scores, measurements, attempts, etc, so that it will help one to interpret data easily.

Let us discuss another example. A teacher conducts a unit test in social science for 45 students. If the teacher arranges marks secured by each student in serial order, it may be difficult to identify the students who scored high mark, low mark, the students who failed and so on. In this case, the marks obtained will be converted to classes (for example, students who scored marks in between 10-20, 20-30 and so on), and then it is presented in a table so that one may be able to recognize the performance effortlessly. **The arrangement of data in a table in sequential manner is called frequency distribution.** The term ‘frequency’ refers to the number of cases or objects in a category or class. For example, suppose 8 students scored mark below 35 in the term examination, then the ‘8’ represents frequency. The data may be arranged in a systematic way as follows (Biswal & Dash, 2009):

- In the form of statistical table
- In the form of rank order, and
- In the form of frequency distribution.

Let us elaborate our discussion on frequency distribution. Suppose, Ms Radhika, a physical education teacher is interested in measuring the weight of the children studying in seventh standard. To do the same, she arranges a weighing machine and calls students as per their roll numbers. She weighs each student’s weight using the weighing machine and notes down them on a piece of paper. Her recordings students weights are shown in Table 13.2:

Table 13.2: Discrete Data

Roll No	Name of the Student	Weight
01	Student A	32
02	Student B	38
03	Student C	31
04	Student D	27
05	Student E	39
06	Student F	36
……..	…………..	………..

The recording was done for 48 students of the class. Now the question is "can we locate the student who is underweight/overweight'? How many students have weight between 30kg-35kg? To answer such questions, we may convert the data into frequency distribution. So frequency distribution is the arrangement of scores and frequency of their occurrences or frequency distribution refers to the tabulation of quantitative data in class intervals which vary in size. From frequency distribution, it is easy to comprehend and understand the general trend of the group and accordingly interpretation can be made. Based on the interpretations appropriate corrective measures may be possible. Let us see the frequency distribution in the above case, where the weights of 48 children are recorded. The frequency distribution is presented in Table 13.3.

Table 13.3 : Grouped Data

Class Interval	Frequency
25-29	12
30-34	31
35-39	3
40-44	2

From the frequency distribution, it is evident that the students studying in seventh standard has a few student whose weight lies between 25-29 kg while majority of them has weight between 30-34kg. Many such interpretations can be made from the frequency distribution. Why don't you try it as an activity?

> ***Activity 3***
>
> Analyse the frequency distribution given in Table-3 and note down the interpretations that can be made out of it.
>
> ....
>
> ....
>
> ....

While discussing frequency distribution, you might notice that, so as to develop the frequency distribution, we need to follow certain procedures and steps. So now let us discuss the procedure followed in developing frequency distribution. For that, imagine a class test conducted by Mr. Ram Vallabh, English teacher of class sixth. The marks scored by 30 students are shown below:

2	22	31	24	15	48	43	28	6	9	38	29	26
37	33	32	26	34	34	40	17	42	40	29	18	35
39	27	25	32									

The above data represents the scores of students in class test and we are going to create the frequency distribution of this particular data. In order to make the frequency distribution the following procedure are carried:

1. Determining range
2. Determining size of class interval
3. Writing contents of frequency distribution

4. Writing class intervals
5. Putting tallies
6. Totalling the number of tallies
7. Checking the number of frequencies

Let us discuss each of these steps briefly.

1. ***Determining range:*** Range is the difference between highest and lowest score in the set of data. In the above case '48' is the highest score and '2'

   Therefore, Range = Highest Score-Lowest score

   Range = 48-2 = 46

2. ***Determining size of class interval:*** Generally the size of class interval is calculated using the formula;

   Size of the Class Interval = Range/No of classes desired

   In this case, Size of class interval = 46/10 = 4.6?5

> Each class has certain number size and is decided based on the following assumptions. But each individual has the freedom to follow these suggestions or not.
>
> - If the number of scores is more than 500, the number of class intervals should be within 25 to 50
> - If the number of scores is more than 200-500, the number of class intervals should be within 25 to 40
> - If the number of scores is more than 100-200, the number of class intervals should be within 15 to 25
> - If the number of scores less than 100, the number of class intervals should be within 5 to 15

3. ***Writing contents of frequency distribution:*** It is sure that, the frequency distribution is made out contents and the same being written in three columns as shown below:

1	2	3
Class Interval	Tallies	Frequency

4. ***Writing class intervals:*** In order to write the class interval, the lowest score and size of class interval is used. In the above case, the lowest score is '2' and size of class interval is '5' and hence the first class interval would be 2-6. The next class interval would be 7-11, 12-16 and so on.

5. ***Putting tallies:*** As the task of writing class interval is completed, 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 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~~

6. ***Totalling the number of tallies:***

As you complete the task of tallying, the next task is to aggregate them. In this example, the total number of tallies against each class interval is given below:

Class Interval	**Tallies**	**Frequency**
1-4	I	1
5-9	II	2
10-14	0	0
15-19	III	3
20-24	II	2
25-29	~~IIII~~ II	7
30-34	~~IIII~~ I	6
35-39	IIII	4
40-44	IIII	4
45-49	I	1

7. ***Checking the number of frequencies:***

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 (Σf) is 30, wherein 'Σ' stands few sigma and 'f' for frequency.

Class Interval	**Tallies**	**Frequency**
1-4	I	1
5-9	II	2
10-14	0	0
15-19	III	3
20-24	II	2
25-29	~~IIII~~ II	7
30-34	~~IIII~~ I	6
35-39	IIII	4
40-44	IIII	4
45-49	I	1
	Σf=30=N	

Check Your Progress 3

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

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

3. Discuss the steps involved in development of frequency distribution.

....

....

....

13.6 GRAPHICAL REPRESENTAION OF DATA: TYPES OF GRAPHS AND ITS USE

Many a times, numerical data is complex and difficult to interpret and understand. This is true for common people and investigators and in our context for teachers. For example, in the above case, the marks scored by students have been converted to frequency distribution. But at times frequency distribution may not serve the purpose as it is not appealing and 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 could be easily interpreted and understood by any one. 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 a geometric figure drawn on scale.

13.6.1 Importance of Graphical Representation

The graphical representation is important due to the following reasons:

- Graphical representations are attractive and beautiful.
- It helps easy visualisation and appealing to eyes.
- Graphical representation 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, both are different. Now let us look for the major difference between graph and diagram forms Figure 13.6.

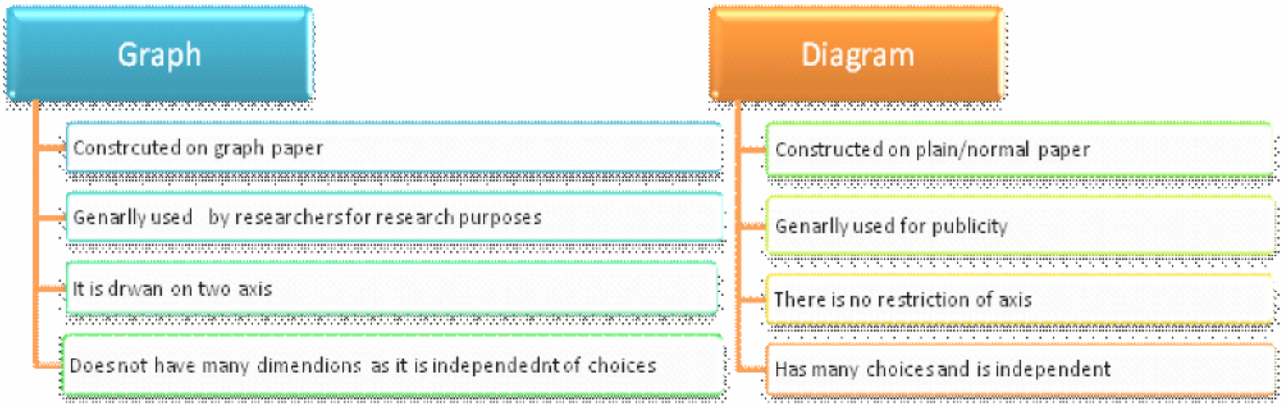


Figure 13.6: Graph and Diagram

13.6.2 General Principles of Drawing a Graph

When a teacher/investigator aspires to convert his/her data into graphical format, he/she needs to keep in mind the following procedures/principles.

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 the flexibility to dilate between 60% to 80%.
5. The graph generally has four quadrants as shown in Figure 13.7. But educationists/psychologists usually use the (++) quadrants to utilize maximum space of the graph paper.

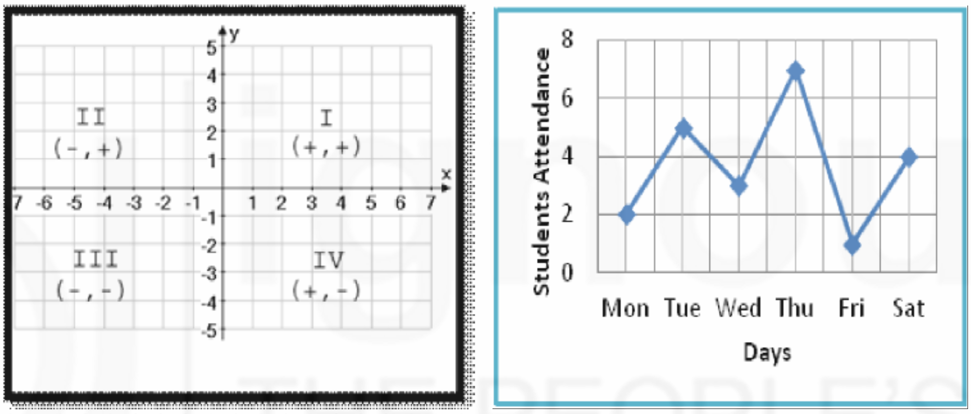


Figure 13.7: Drawing a Graph

13.6.3 Types of Graph/Graphical Representations

In the previous section, we have discussed what graphical representation, difference between graphs is and diagram, how bar graph is drawn, etc. In this section, we will discuss various types of graphical representation/graph and you will find them useful as you teach your students. There are various types of graphical representations for grouped and ungrouped data. Let us discuss those representations separately.

Graphical Representation of Grouped Data

The data in the form of raw scores is called ungrouped data while data organized in the frequency distribution is called grouped data. The following are the different types of graphs/diagrams which we use when the data is ungrouped.

i. Pictograph or pictogram

ii. Bar graph or bar diagram

iii. Circle or Pie graphs/diagram

iv. Line graphs

Now let us discuss each of these graphs in details.

Pictograph or Pictogram

A picture is said to be worth 100 times more meaningful than the words spoken or written. Pictures have the quality to convey ideas in a more meaningful way. Thus the statistical data can be translated into pictures. The pictorial representation of statistical data is known as pictograph or pictogram. Let us discuss an example for pictograph. Consider a school, having the following number of girls in various classes as shown below:

Table 13.4 : Data for Pictogram

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

In order to construct the pictograph, we will assign a picture for girl. For example, the picture selected in this case is (a human face). After that, we will convert the above data into pictograph. The pictograph would be like the one given below. Remember that, to draw pictograph, we need to select a scale, for example the scale selected in this case is 5 girls for one picture.

Table 13.5 : Pictogram

S.No	Class	No. of Girls
1	8A	
2	8B	
3	8C	
5	8D	
6	8E	

Pictograph has both merits and limitations. The merit of pictograph is that, it is visually appealing and easy to comprehend. Anyone can easily make out number of girls present in each class without any difficulty. In the above example, the limitation of pictograph was not much visible as we have total number of girls, which is divisible by 5. If the number is not an exact multiple of five, we would have faced difficulty to represent them as pictures. For example, if the number of girls were 23, it would be difficult to represent them. Such complications are very minimal in the forthcoming representations.

Bar graph or bar diagram

In pictograph we have observed that, pictures were used to represent statistical data. Instead of pictures, bar (rectangles with similar width) are used in bar graphs. Thus the mode of representation of statistical data using bars is known as bar graph or bar diagram. The following are the steps used for constructing 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 now apply these steps to draw the bar graph for the data given above. Here the number of girls studying in various eighth grades is given. In order to draw bar graph, number of girl's students is taken on the y axis and the corresponding grades are selected on the x axis. The resulting bar graph is given below:

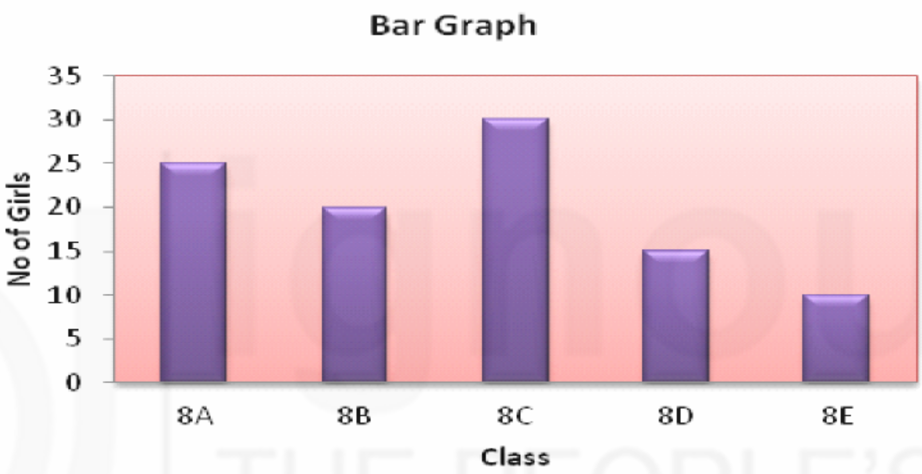


Figure 13.8 : Bar Graph

The next question is to how to interprete bar graph. In this case we can say that, the class 8C has more 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 4

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

....

....

....

Circle or Pie graphs/diagram

The pi-diagram/pie graph is known as circle graph as we represent the statistical data as circular figure considering weightage given to the proportion of data. Many a times, we are interested to discern percentage of statistical data and in such cases pie-diagram the most popular graphical representation is used.

Thus the percentage break-ups are represented in pi- diagrams. 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.

Adya, the teacher in charge to organize arts festival of Bhawan public school has collected the details of children who wish to participate in picture drawing competition. As per the instruction many students have registered themselves for it. The details of their registrations are given in Table 13.6:

Table 13.6 : Data for Pie graph

S.No	Class	No of Students
1	10	16
2	9	29
3	8	36
4	7	13
5	6	41
	Total	**135**

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 π (2 pie). 2π is equal to 2×180^0 $= 360^0$. Thus the whole circle represents 360^0. Thus we will represent the the total sample i.e. 135 through a circle having 360^0. Let us see how it is done.

In the above example, there are 16 students from 10^{th} class registered for drawing competition. Thus, first we will find out the percentage of students out of 135.

i.e. proportion out of 135= $(16 \div 135) \times 100 = 11.85\%$ = 12%

Now we will find the proportion of 10^{th} class students in 360^0.

i.e. proportion out of 360^0 $=(16 \div 135) \times 360° = 42.66^0$.

Thus in the circle 42.66% represents 10^{th} class students registered for drawing competition. In a similar way we will calculate the student belonging to rest of the classes.

Class 9

Proportion out of 135= $(29 \div 135) \times 100 = 21.48\%$ = 21% (Approx.)

Proportion out of 360^0 $=(29 \div 135) \times 360° = 77.33^0$.

Class 8

Proportion out of 135= $(36 \div 135) \times 100 = 26.66\%$ = 27% (Approx.)

Proportion out of 360^0 $=(36 \div 135) \times 360° = 96.00^0$.

Class 7

Proportion out of 135= $(13 \div 135) \times 100 =$ **9.62%** = 10% (Approx.)

Proportion out of 360^0 $=(13 \div 135) \times 360° = 34.66^0$.

Class 6

Proportion out of 135= $(41 \div 135) \times 100 = 30.37\%$s = 30% (Approx.)

Proportion out of 360^0 $=(41 \div 135) \times 360° = 109.33^0$.

Note that, if we add the proportions of different changes out of 360^0, we will get 42.66+77.33+96.00+34.66+109.33 = 359.98 = 360^0. Thus the number of students registered can be represented in percentages. The next step is transferring the percentage breakdowns into different sectors of a circle. For that, draw a circle with a compass. Then depict percentage corresponding to class 10 using the protractor. The same process is repeatedfor each class. The final pie-diagram is shown in Figure 13.9:

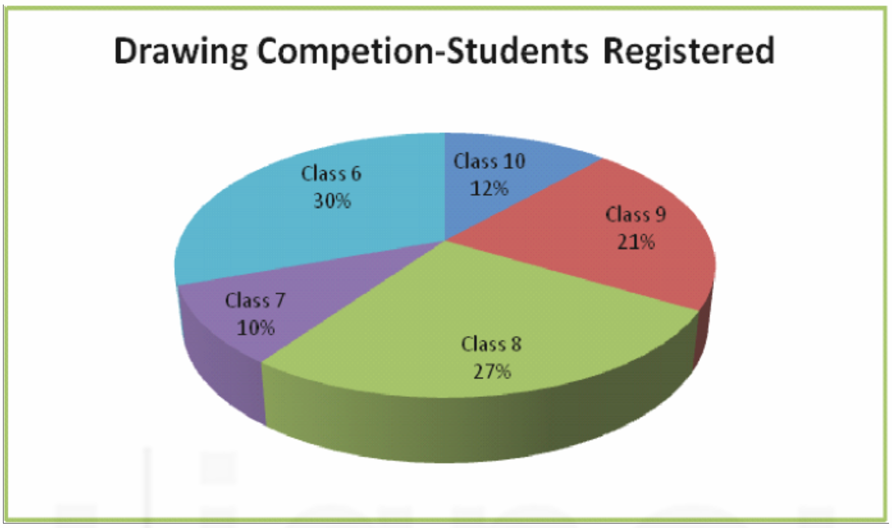


Figure 13.9 : Pie Graph

Check Your Progress 4

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

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

4. What is pie-graph? Develop pie-graph for the following data.

S.No	Class	No of Students from Urban Area
1	10A	35
2	10B	20
3	10C	42
4	10D	11
5	10E	34

Line graphs

Line graphs are one of the common modes of representations of statistical data. In a line graph the relationship between two variables are illustrated in a graph. 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 by choosing an example. Below, given are the details of absentees in a particular class.

Table 13.7 : Data for Line graph

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

To represent the data given in Table 13.7 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 I 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:

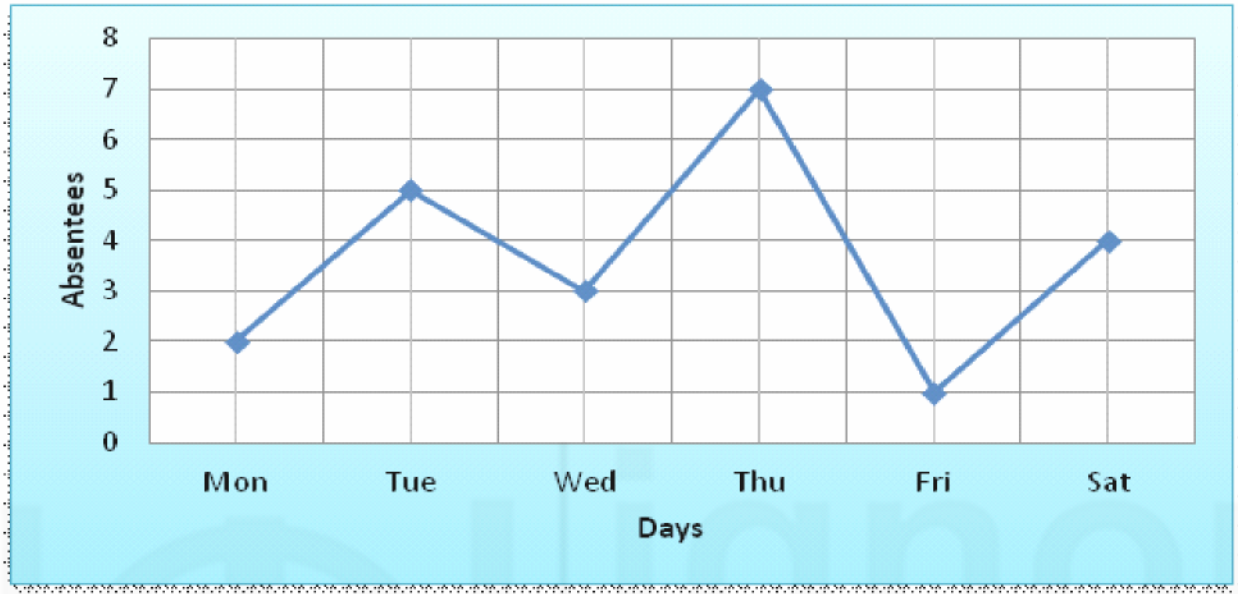


Figure 13.10: Line Graph

Graphical Representation of Grouped Data

When the raw scores are arranged in frequency distribution, the data obtained is called grouped data. The following are the graphical representations of grouped data.

1. Histogram or column diagram
2. Frequency polygon
3. Cumulative Frequency Graph
4. Cumulative Frequency Percentage Curve or Ogive.

Now let us discuss each of these representations with examples.

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. But histogram is used when the statistical data is arranged in class intervals. Here the frequency is represented using vertical adjacent rectangles. Generally the class interval is depicted in x axis and frequency on y axis. Thus the base of the rectangle represents the class interval and height its frequency. Thus 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 construction of histogram. For that, **Table 13.8 : Data for Histogram** the data given in Table 13.8 will be used.

Table 13.8 : Data for Histogram

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

To construct 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 in 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 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 13.11:

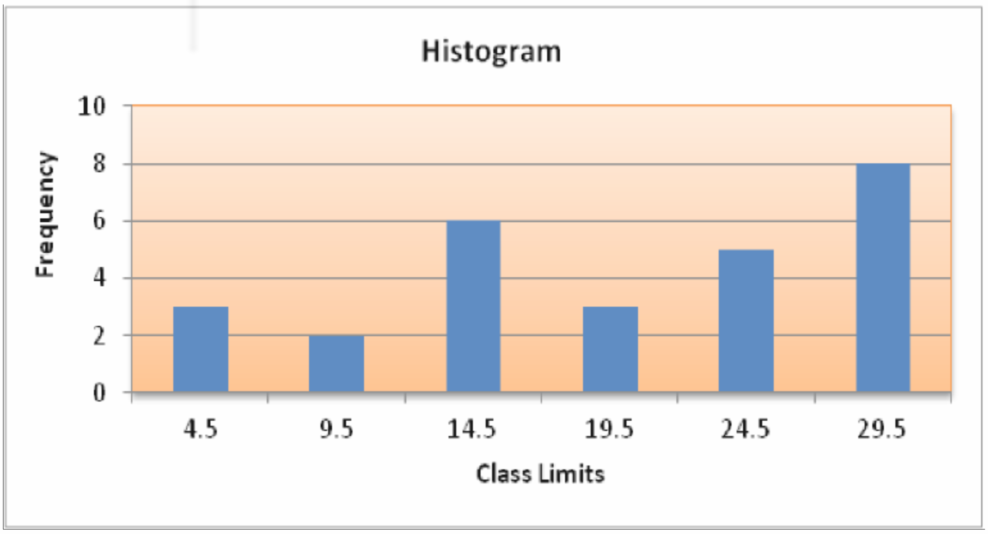


Figure 13.11 : Histogram

Frequency Polygon

What do you mean by polygon? Polygon is a closed figure with many sides. So, as you draw a frequency polygon get a many sided closed figure. Let us explore a more specific definition of frequency polygon. Frequency polygon is a line graph representation of statistical data/frequency distribution. To construct 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 touch themselves form a closed a figure and hence the name frequency polygon. Now let us look at steps followed in construction of frequency polygon using the same data discussed in the previous section. The same data is reproduced in Table 13.9.

Table 13.9 : Data for Frequency Polygon

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

- To draw 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 points 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 create a closed polygon.

The frequency polygon for the given frequency distribution is given in Figure 13.12:

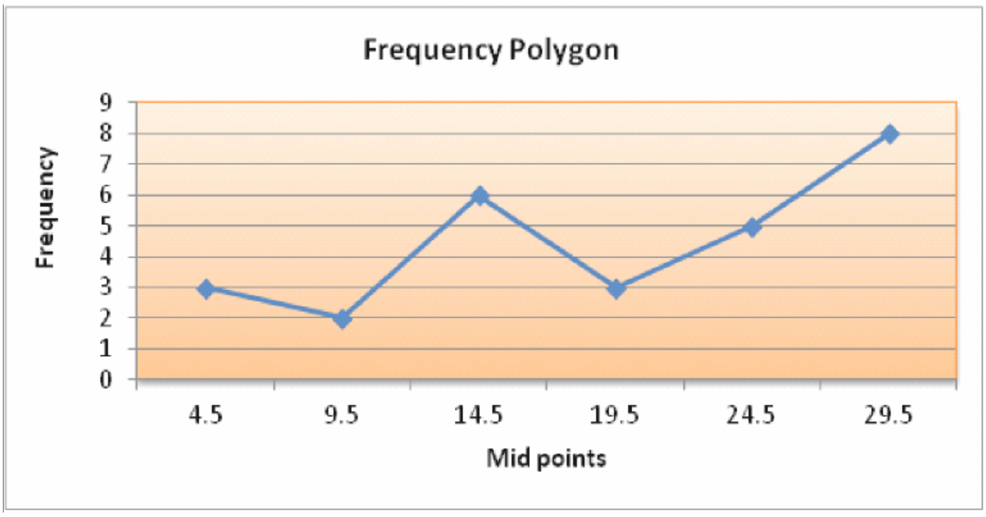


Figure 13.12 : Frequency Polygon

Cumulative Frequency Graph

The third method of representing grouped 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 13.10 in a cumulative frequency graph.

Table 13.10 : Data for Cumulative Frequency Graph

Class Interval	Frequency(f)	Upper Limit Frequency	Cumulative
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

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 13.13.

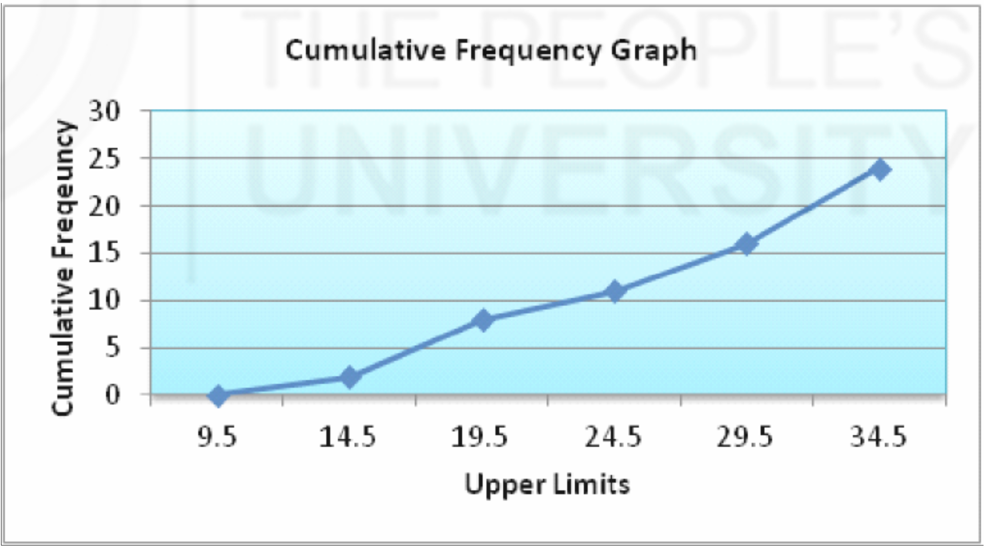


Figure 13.13 : Cumulative Frequency Graph

Cumulative Frequency Percentage Curve or Ogive.

Cumulative Frequency Percentage Curve or Ogive is drawn by following the 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.

Table 13.11 : Data for Ogive

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

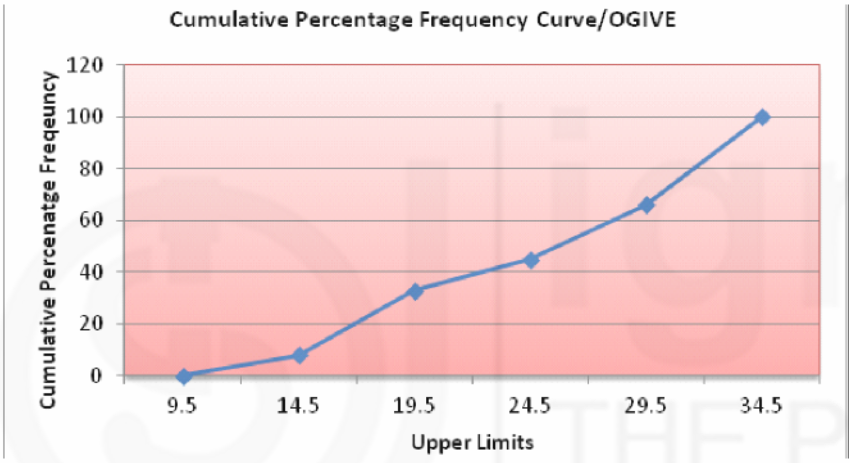
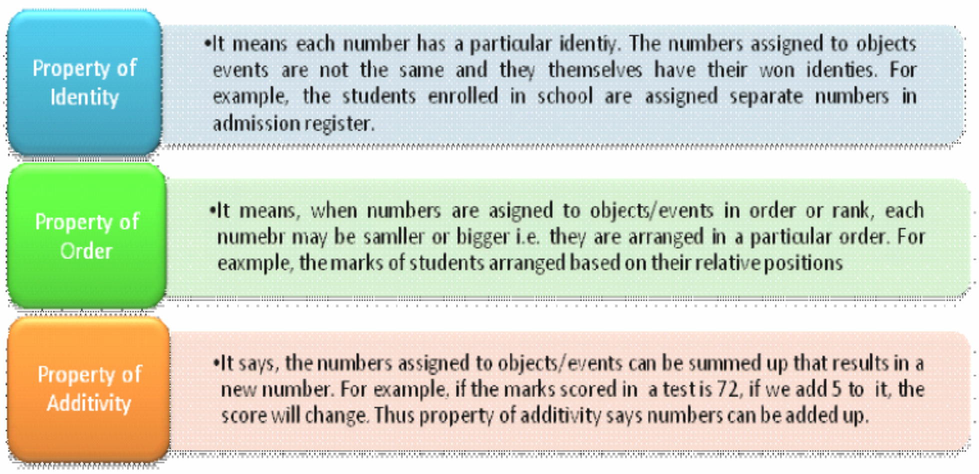


Figure 13.14 : Ogive

13.7 SCALES OF MEASUREMENT

Measurement is quantitative in nature. It is the process of assigning numbers to objects or events based on accepted norms and rules. For example, conducting a unit test and assigning scores/marks to children's performance is measurement. While assigning numbers to object/events, certain properties need to be considered. These properties are known as the different types of scales of measurement. The following are the properties:



Classification of Scales of Measurement

To measure variables one may use different instruments. If it is time, stop watch is used to measure academic achievement, achievement test is administered and so on. Thus the nature of measurements depends on the variables involved in the measurement process but all kinds of measurement can be classified into certain categories. These categories are called scales. There are four types of scales of measurement, based on the properties described above. They are:

1. Nominal Scale
2. Ordinal Scale
3. Interval Scale
4. Ratio Scale

A brief description of these scales of measurement is given below:

Nominal Scale: Measurement using nominal scale simply names or categorizes the responses. In nominal scale, numbers or symbols are assigned to represent individual/objects/events/categories, etc. Nominal scales employ the property of identity.

Example: Numbers assigned for jersey of players in a school football team, writing M/F to represent girls and boys, etc.

Ordinal Scale: In ordinal scale, the measurements are arranged on the basis of any particular order or rank ranging from lowest to highest or vice versa. Thus the objects/events are ranked. Ordinal in ordinal scale refers to 'order' and it provides direction in addition to nominal information. Ordinal scale employs the property of order. At the same time, ordinal scale applies the property of identity too.

Example: Ranking students on the basis of marks they scored, ordering children based on their difference in heights.

Interval Scale: Interval scales are numerical scales with equal intervals having the same interpretations. Interval scales do have order but with equal interval between them. For example, the interval between 40 and 50 degree and 80 and 90 in Fahrenheit has the same meaning. Interval scales are characterized by absence of 'zero' even if the measurement scales have it. For example, measurement of temperature in Fahrenheit scale has zero but '0' doesn't mean temperature is zero. Interval scales employ the property of identity, order and additivity.

Ratio Scale: This is the most informative scale of measurement. Ratio scales are the interval scales with presence of absolute zero. For example, in Kelvin scale, it has a zero point, which means temperature measurement '0' refers to zero itself.

Example: The measurement of money comes under ratio scale. There can have chances that a person may have Rs10 or Rs 1lakh or No money at all (The 'zero' case).

Check Your Progress 5

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

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

5. Discuss briefly the various scales of measurement with examples.

....

....

....

13.8 LET US SUM UP

School is a rich source of information useful for educational research and results of such researches are used for improving educational practices and implementing innovative measures. In educational researches, statistics is used to collect, organise and interpret information. There are various methods to accomplish these tasks. In this Unit, we have touched upon various aspects of statistics such as use of statistics in educational assessment and evaluation, the meaning and nature of data and its types, various data organizing techniques, frequency distribution and the scales of measurement. All these would enable you to assess and improve your teaching –learning practices.

13.9 REFERENCES AND SUGGESTED READINGS

Garrett, H.E. (1956). *Elementary statistics*. Longmans, Green and Co: New York.

Guildford, J.P. (1965).*Fundamental statistics in psychology and education*. McGraw Hill Book Company: New York.

Rath, R.K. (1999). *Fundamentals of educational statistics and measurement*. Taratarini Pustakalaya: Orissa.

Asthana H.S. & Bhushan, B. (2007). *Statistics for social sciences*. Prentice Hall of India, Pvt. Ltd. :New Delhi.

Sahre, B.K. (2005). *Statistics in psychology and education*. Kalyani Publishers: Ludhiana.

Best J.W. (1977). *Research in education*. Englewood Cliffs, New Jersey: Prentice Hall Inc.

Bhatacharya S. (1968). *Foundation of education and educational research*. Acharya Book Depot: Baroda.

Koul L. (1984). *Methodology of educational research*. Vikas Publishing House Pvt. Ltd: New Delhi.

Pilai, RSN and Bhagavati (2008). Statistics – Theory and Practice (7th ed.), New Delhi : S. Chand and Company Ltd.

Wiersma, W. (1975).*Research methods in education: An introduction*. F. E: Peacock Publishers Inc: Illionois.

13.10 ANSWERS TO CHECK YOUR PROGRESS

1. To analyse students achievement scores, prepare progress report and report cards. This can also be used in conducting action research.
2. Primary data are collected by the teacher directly dealing with the students. Like achievement score of the students on the test which the teacher has taken. Secondary data are usually taken from various sources which has published in certain books, reports, encyclopaedias, etc.
3. Self exercise content section (refer 13.5).
4. Self exercise.
5. Nominal, ordinal, interval and ratio scales. For examples, refer section 13.7.

UNIT 14 MEASURES OF CENTRAL TENDENCY

Structure

14.1 INTRODUCTION

As a teacher, we always deal with the data gathered from assessment of student's performance in continuous as well as terminal examinations. Many a time, we get engaged with analyzing the data/score of the students and compare them with other students or groups. We also prepare the progress report of the students in which we use certain statistics to describe the performance of the students in a meaningful and understandable way. We analyze the individual score of the students as well as the group performance of the sections, classes and schools. We also compare the performance of students between sections, classes and schools.

Keeping in view the above, the present Unit discusses the concept, process and methods of calculation of measures of central tendency of the scores. In the previous Unit, you have studied about the nature of data. In this Unit, you will recapitulate the concept of individual and the group data. Further the Unit will make you acquainted with various methods used for calculation of measures of central tendency i.e. Mean, Median and Mode and also make you understand the interpretation and use of the Mean, Median, and Mode. The Unit also makes you acquainted with establishing relationships and comparisons among Mean, Median and Mode.

14.2 OBJECTIVES

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

- differentiate between individual and group data;
- explain the meaning of measures of central tendency;
- compute mean from ungrouped and grouped data and interpret it;
- compute and interpret the combined mean;
- compute median from ungrouped and grouped data and interpret it;
- compute mode from ungrouped and grouped data and interpret the it;
- discuss the uses and limitations of mean, median and mode;
- discuss the relationship of mean, median and mode;
- compare the mean, median and mode for their relative importance in a given context; and
- select an appropriate measure of central tendency as per the nature of data and purpose.

14.3 INDIVIDUAL AND GROUP DATA

In the previous Unit, you have studied the nature of data, its organization, making frequency distribution table and converting raw scores into group scores. To recapitulate what you have studied earlier, we can say that the raw or the individual score keeps very little meaning because it lacks value of interpreting the score with other scores. At the same time, without knowing other individual scores it is difficult to analyze the order and the position of an individual score in the group. For example, let Sohan scored 50 in a test of Mathematics. What does the score mean? How many students are below or above the score of Sohan? Whether Sohan is an average, below average or above average student? These questions can't be answered simply from the individual score. To make the individual score more meaningful, we can convert it into group scores. Group scores can be presented in terms of average score, equal intervals, and also ordering it both in ascending and descending order. Individual scores can be presented as group score in terms of frequency distribution table. The individual (raw score) and the group score have been presented in the following table.

Table 14.1: Individual/Raw score and the group score

Individual/Raw Score			Group Score	
Name	**Identification**	**Score**	**Class Interval**	**Frequencies**
Sohan	X_1	45	55 - 59	1
Ritu	X_2	20		
Zakir	X_3	11	50 - 54	2
Suresh	X_4	33		
Sania	X_5	53	45 - 49	2
Manju	X_6	57		
Rehan	X_7	46	40 - 44	2
Kanak	X_8	24		
Roni	X_9	29	35 - 39	2
Deekshya	X_{10}	30		
Kranti	X_{11}	19	30 - 34	3
Samir	X_{12}	36		
Subrat	X_{13}	42	25 - 29	1
Manjit	X_{14}	38		
Abdul	X_{15}	21	20 - 24	4
John	X_{16}	50		
Karan	X_{17}	17	15 - 19	2
Harihar	X_{18}	23		
Mamta	X_{19}	43	10 - 14	1
Deep	X_{20}	33		
	N = 20			**N = 20**

The above table shows the individual score which can also be called as raw score and it is converted to group score which is presented in Class Intervals with their frequencies. Getting raw score is the first step to go forward and interpreting it in different meaningful ways, mean, median, or mode of the scores. The mean, median and the mode can be calculated in various methods depending upon the nature of the data. In the next section, we will discuss the concept of Measures of Central tendency and how to calculate the central tendency.

Activity 1

Referring table 14.1, convert the following individual/raw scores into group scores by representing it in class intervals and frequencies.

Individual/raw Scores:

X_1 – 20; X_2 – 18; X_3 – 52; X_4 – 45; X_5 – 37; X_6 – 38; X_7 – 24; X_8 – 17; X_9 – 48; X_{10} – 40; X_{11} – 50; X_{12} – 52; X_{13} – 33; X_{14} – 26; X_{15} – 31; X_{16} – 14; X_{17} – 44; X_{18} – 29; X_{19} – 56; X_{20} – 63; X_{21} – 67; X_{22} – 55; X_{23} – 30; X_{24} – 27; X_{25} – 35; X_{26} – 41; X_{27} – 31; X_{28} – 19; X_{29} – 54; X_{30} – 22

(N = 30)

....

....

14.4 MEASURES OF CENTRAL TENDENCY

As a teacher, you might have engaged in analyzing variety of data relating to school as well as performance of the students in different subjects. Again, you might have come across to analyse the data for preparing the report card and progress of the students. In such cases, quite a time, we need to use certain statistical techniques to make the individual score meaningful by comparing it with the group scores. The most common statistics that we use in school is the measures of central tendency. Central tendency is the central value or score of a group. For example, let the score of 30 students in a particular subject is given, we can calculate the average score of that group and then we can compare the individual score with the central value/score of that group. Again, central tendency also helps us to compare the central values of two different groups it may be within different sections of a same class, between two different classes in the same school and also between same standards in two different schools. We use measures of central tendency for making inter and intra group comparisons. The most commonly used measures of central tendency are:

- Mean
- Median
- Mode

14.4.1 Scales of Measurement and Measures of Central Tendency

In Unit-13, you have studied the scales of measurement. We usually use four types or scales of measurement, i.e. nominal, ordinal, interval and ratio scale. In educational measurement, generally, we use nominal, ordinal and equal interval scale. Ratio scale is used in measurement of physical sciences where there is a concept of absolute zero point in measurement. But in educational measurement, we consider zero point measurement as the relative zero point. The use of different methods of calculating measures of central tendency depends upon the nature of the data and its scales of measurement.

Data on Nominal Scale: If data is on nominal scale which is mostly qualitative in nature, we can simply count the number of cases in each category and obtained frequencies. In such cases, if the data is in nominal scale, we can use the statistics 'The Mode' as the measure of central tendency. It supports the definition of the mode, i.e., 'mode is the most frequent item/score of the group'.

Data on Ordinal Scale: Ordinal scale implies that the scores are in an order, either ascending or descending. When data is arranged in rank order the measure of central tendency found by locating a point that divides the whole distribution into two equal halves. In such cases, we use the statistics 'median' as the measure of central tendency. It supports the definition of median, i.e., 'median is a point on the scale of measurement below and above which lie exactly 50 percent of the cases. It may be noted that median is defined as a point and not as a score or any particular measurement.

Data on Equal Interval Scale: When data is presented in class intervals and there is scope to convert the data into class intervals, we can best use statistics 'Mean' as the measure of central tendency. Mean is the most popular measures of central tendency which is highly used in schools. Mean of a distribution of scores may be defined as the point on the scale of measurement obtained by dividing the sum of all the scores by the number of scores. Mean is also called as the most appropriate measures of central tendency.

In the next section, we will discuss the methods of calculating Mean, Median and Mode as the measures of central tendency.

14.5 THE MEAN

As discussed above, mean is calculated when the data is presented or have the scope to present it in interval scale. It is also popularly known as 'Arithmetic Mean'. Mean can be calculated by adding sum of the observations/scores by dividing by total number of cases. The formula of calculating Mean is as follows:

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

Where, ΣX = Sum of all the scores/values

N = Total number of cases

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

14.5.1 Calculation of Mean for Ungrouped Data

When only raw data is given, Mean can be calculated by adding all the raw scores and dividing it by the total number of the raw scores.

Example: The scores of ten students in a class is given as follows. Calculate the Mean of the scores.

Scores: 15, 37, 22, 11, 40, 29, 32, 45, 20, and 30. (N = 10)

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

['X' is the individual score and 'N' is total number of cases]

= (15+37+22+11+40+29+32+45+20+30) / 10

= 281 / 10

= 28.1 (Answer)

14.5.2 Calculation of Mean for Grouped Data

We can calculate Mean in three situations, i.e.

1. When the scores and the frequencies are given
2. When data is arranged in frequency distribution table i.e. Class Intervals as well as Frequencies are given.
3. When data is arranged in frequency distribution table i.e. Class Intervals as well as Frequencies are given. (by using Assumed Mean Method is also called as short method)

1. Calculation of Mean when Scores and Frequencies are given

We can calculate Mean by using a simple formula when scores and frequencies are given. Let us solve with an example.

Example: Calculate mean of the following data:

Score	25	31	33	42	46	51	55	58	60	72
Frequencies	3	7	9	12	13	6	4	3	2	1

Solution:

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

?fX = Summation of frequencies multiplied with the scores

N = Total frequencies

Score (X)	Frequencies (f)	fX
25	3	75
31	7	217
33	9	297
42	12	504
46	13	598
51	6	306
55	4	220
58	3	174
60	2	120
72	1	72
	N = 60	**$\sum$fX = 2583**

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

= 2583 / 60 = 43.05 (Answer)

2. Calculation of Mean when Data are given in Class Intervals with Frequencies

In group data, when the class intervals as well as frequencies are given, we can calculate Mean by using the same formula. This method is also called as long method for calculating Mean. The formula for calculating Mean is as follows:

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

X = Mid-point of the Class Interval

f = Frequencies

N = Total number of cases ($\sum$ f)

Example : Compute the Mean for the following frequency distribution.

Class Interval	40-44	35-39	30-34	25-29	20-24	15-19	10-14
Frequencies	3	5	10	14	8	6	4

Solution :

Score (X)	**Frequencies (f)**	**Mid Point (X)**	**fX**
40-44	3	42	126
35-39	5	37	185
30-34	10	32	320
25-29	14	27	378
20-24	8	22	176
15-19	6	17	102
10-14	4	12	48
	N = 50		**$\sum$ fX = 1335**

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

= 1335 / 50

= 26.7 (Answer)

3. Calculating Mean by using Assumed Mean Method

In group data, when the data is presented in class intervals with frequencies, we can calculate Mean by using long method as well as assumed mean or short method. Assumed mean method is widely used to calculate mean in this situation as because to avoid lengthy calculations of multiplications of mid-points of class intervals with their corresponding frequencies. In assumed mean method, we assume a class, assuming that the mean lies in that class. We follow certain steps to calculate Mean in assumed mean method. The following formula is used to calculate Mean in assumed mean method:

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

A.M. = Assumed Mean

f = Frequencies

d = Deviation from assumed mean

N = Total number of frequencies

c.i. = Size of the Class Interval

Steps followed for calculating Mean in Assumed Mean Method:

Step 1 : Calculation 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

$[Mid\ point = \frac{Lower\ Limit\ of\ the\ Class + Higher\ Limit\ of\ the\ Class}{2}]$

Step 3 : Calculation of deviation (d)

[Deviation (d) can be calculated by using the method like : $d = \frac{Mid\ point\ of\ the\ Class - A.M.}{Size\ of\ the\ Class}$

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 an example:

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

Class Interval	50-54	45-49	40-44	35-39	30-34	25-29	20-24	15-19	10-14
Frequencies	2	3	5	7	10	6	4	2	1

Solution :

Score (X)	**Mid Point (X)**	**Frequencies (f)**	**Deviation (d)**	**fd**
50-54	52	2	+4	8
45-49	47	3	+3	9
40-44	42	5	+2	10
35-39	37	7	+1	7
30-34	32 (A.M.)	10	0	0
25-29	27	6	-1	-6
20-24	22	4	-2	-8
15-19	17	2	-3	-6
10-14	12	1	-4	-4
		N = 40		**$\sum$fd = 10**

For calculating the values, the above mentioned steps are used.

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

A.M. = 32

N = 40

c.i. = 5

$\sum$fd = 10

$$Mean\ (M) = 32 + \frac{10}{40} \times 5$$

= 32 + 1.25

= 33.25 (Answer)

Note: *After solving such questions you will see that you always get values +1, +2, +3 etc. on one side and -1, -2, -3 etc. on the other side of the class interval chosen for Assumed Mean. In fact it becomes just a mechanical process, after some time, to find deviations i.e. first putting Zero (0) against the column of Assumed Mean and putting +1, +2, +3 towards class intervals with bigger score limits and -1, -2, -3 towards class intervals shown with smaller score limits. This may help you save time as well. (B.Ed. ES-333, IGNOU, 2010)*

14.5.3 Calculation of Combined Mean

As like calculation of Mean is very much important for the part of a teacher, accordingly calculation of Combined Mean is also equally important. Quite a time, you might have come across the situation that in a particular class, there may be many sections and you have to analyse their scores in different subjects. The number of students in each sections are also sometime different. In such cases, you can calculate a combined mean of the entire groups if the data relating to mean score and the Number of students in each group is given. Similarly if we have the means for various schools and the district mean is required, it would also call for computing the combined mean. Combined Mean can be calculated by using the following formula:

$$Combined\ Mean\ (M.comb.) = \frac{N1M1 + N2M2 + N3M3 + ---}{N1 + N2 + N3 + ---}$$

N1 = Number of the first group

N2 = Number of the second group

N3 = Number of the third group

M1 = Mean of the first group

M2 = Mean of the second group

M3 = Mean of the third group

Let us calculate mean by using the above formula.

Example : There are three sections in Class-IX of a school, section A, B and C. The Mean score in Mathematics of each sections with the number of students are given as follows. Find the combined mean of the students in Mathematics.

Sections	**A**	**B**	**C**
Mean Scores	**82**	**65**	**70**
Number of Students	36	33	41

Solution:

$$Combined\ Mean\ (M\ comb.) = \frac{N1M1 + N2M2 + N3M3}{N1 + N2 + N3}$$

[N1 = 36; N2 = 33; and N3 = 41]

[M1 = 82; M2 = 65; and M3 = 70]

$$Combined\ Mean\ (M\ comb.) = \frac{(36 \times 82) + (33 \times 65) + (41 \times 70)}{36 + 33 + 41}$$

$$= \frac{2952 + 2145 + 2870}{110}$$

$$= \frac{7967}{110}$$

= 72.43

From the above example, you may notice that the combined mean is 72.43, where as the individual mean of different sections are 82 in section A, 65 in section B and 70 in section C. Now you can compare between average performance of students in a section with the combined mean of all the sections. If someone erroneously adds the mean scores of all the sections and divide it by the number of sections, the average will come as 72.33, which is not same as the calculated combined mean. It is therefore, for getting the multi group average, better to calculate their combine mean by using the appropriate formula, not simply by calculating the average.

14.5.4 Use of Mean

Mean is called as the most appropriate measure of central tendency as it considers the value of each and every score of the group for calculating the mean. Mean is referred because of its high reliability and its applicability to inferential statistics. Mean provides a clear idea about how the scores are varied from the central value. Mean is used when:

- The data is distributed symmetrically, i.e. distributions are not marked skewed.
- We wish to know the centre of gravity of a sample.
- Central tendency with greater stability is wanted.
- Other statistics (standard deviations, coefficient of correlation etc.) for inferential purposes are to be calculated.
- Group performances are to be calculated with accuracy and precision.
- Comparing the intra and the inter group students.

14.5.5 Limitations of Mean

Besides the use of Mean it has also its own limitations. The limitations can be listed as follows:

- A single extreme score (may be lowest or highest) may influence the value of mean. As example, let the daily wages of five persons are Rs. 100.00; Rs.150.00; Rs. 170.00; Rs.160.00 and Rs. 3000.00. The average daily wages comes as Rs.716.00. In this case, the only one extreme score influence the group average. A person who is getting Rs.100.00 per day, is calculated as Rs. 716.00 as average. This is not justified.
- There are situations, where mean may not provide meaningful information.
- Works only when all values are equally important.

Check Your Progress 1

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

b) Compare your answer with the one given at the end of the Unit.

[Put ( ) mark in appropriate answer]

1. Mean is used when data is presented in :

   a) Nominal Scale b) Ordinal Scale c) Equal Interval Scale

2. Mean is the appropriate measure of central tendency when the data is :

   a) Symmetrical b) Positively Skewed c) Negatively Skewed

3. Mean is the most reliable method of finding out the central tendency, because :

   a) It is easy to calculate

   b) Each score is given weightage to calculate mean

   c) Can calculate mean by assumed mean method

   d) It is highly used in research

4. Combined Mean is calculated, when :

   a) Comparison is needed among the groups

   b) Multi group comparison is needed in case the numbers of students in the groups are unequal.

   c) Comparison of mean scores is needed between two or more schools.

   d) All of the above.

14.6 THE MEDIAN

You have studied in the section 14.4.1 of this Unit that Median is used when the scores are presented in ordinal scale, i.e. possible to organize either in descending order or ascending order. Median is defined as it is the 'mid point of the series below and above which lie exactly 50 percent of the cases'. For calculating Mean, scores are given importance whereas for calculating Median, number of items are given importance. It may be noted that median is defined as a point on the series and not as a score or any particular measurement. The score is identified indirectly calculating from the numbers. As like Mean, Median can also be calculated by using both ungrouped and grouped data.

14.6.1 Calculation of Median for Ungrouped Data

For calculating Median in 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 certain changes in calculating Median, when the number of items in the series is in odd or even numbers.

Median in Ungrouped Data (When number of items are odd)

We use the following formula for calculating Median when the total number of items in the group are odd:

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

[Where, N is total number of cases]

Let us calculate it with an example:

Example : Find the Median of the following scores:

Scores : 9, 22, 32, 19, 12, 26, 15 (N = 7)

Solution :

Step 1 : Arrange the scores either in ascending or descending order (Let us arrange it in ascending order).

Step 2 : Use the formula for calculating Median

Step 3 : Determine the Median score from the item

Scores (in ascending order) : 9, 12, 15, 19, 22, 26, 32

N = 7

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

$$= \frac{7+1}{2} \; th \; item$$

$$= \frac{8}{2} \; th \; item$$

= 4^{th} item

= 19 (Answer)

4^{th} item of the series in ascending or descending order is 19.

Median for Ungrouped Data (When number of items are even)

We use the following formula for calculating Median when the total number of items in the group are even:

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

[Where, N is total number of cases]

Let us calculate it with an example:

Example : Find the Median of the following scores:

Scores : 9, 22, 32, 19, 12, 26, 15, 25 (N = 8)

Solution :

Step 1 : Arrange the scores either ascending or descending order (Let us arrange it in ascending order).

Step 2 : Use the formula for calculating Median

Step 3 : Determine the Median score from the item

Scores (in ascending order) : 9, 12, 15, 19, 22, 25, 26, 32

N = 8

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

$$= \frac{8+1}{2}\ th\ item$$

$$= \frac{9}{2}\ th\ item$$

= 4.5^{th} item

$$= \frac{4th\ Item + 5th\ Item}{2}$$

$= \frac{19+22}{2}$ *[**Note**: 4^{th} item is 19 and 5^{th} item is 22]*

$$= \frac{41}{2}$$

= 20.5 (Answer)

4.5^{th} item of the series is 20.5

14.6.2 Calculation of Median for Grouped Data

You have studied that the Median is the mid point of the series in order below and above which 50 percent of the cases lie. For calculating Median in grouped data, the assumption made is that frequencies are evenly distributed within the class interval. The following formula we use for calculating Median in grouped data:

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

Where :

L = Lower limit of the median class

N = Total number of the cases

cfb = Cumulative Frequency below the median class

fm = Frequency of the median class

c.i. = Size of the Class Interval

Steps followed for calculation of Median :

Step 1 : Find out the Median Class (the class which follows N/2th value; can be seen from the cumulative frequency).

Step 2 : Calculate Lower Limit of the Median Class by subtracting 0.5 from the Lower Limit of the Median Class.

Step 3 : Find out cfb; is the cumulative frequency below the Median Class.

Step 4 : Find out fm; is the exact frequency of the Median Class.

Step 5 : Find out the size of the class

Step 6 : Apply the formula to calculate Median

Example : Calculate median for the following frequency distribution

Class Interval	90-94	85-89	80-84	75-79	70-74	65-69	60-64	55-59	50-54	45-49
Frequencies	2	3	3	4	7	12	9	4	2	2

Solution :

Class Interval (C.I.)	**Frequencies (f)**	**Cumulative Frequency (cf)**
90-94	2	48
85-89	3	46
80-84	3	43
75-79	4	40
70-74	7	36
65-69 (Median Class)	**12 (fm)**	**29**
60-64	9	**17 (cfb)**
55-59	4	8
50-54	2	4
45-49	2	2
	N = 48	

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

N/2 = 24

L = 64.5

cfb = 17

fm = 12

c.i. = 5

$$Median = 64.5 + \frac{24 - 17}{12} \times 5$$

$$= 64.5 + \frac{7}{12} \times 5$$

$= 64.5 + \frac{35}{12}$

$= 64.5 + 2.92$

= 67.42 (Answer)

Special Case for Calculation of Median

There may arise a special situation where there are no cases within the interval containing the Median. Let us take an example.

Example : Find the Median for the following frequency distribution :

Class Interval	26-28	23-25	20-22	17-19	14-16	11-13	8-10	5-7
Frequencies	2	2	7	6	0	9	7	1

Solution :

Class Interval (C.I.)	**Frequencies (f)**	**Cumulative Frequency (cf)**
26-28	2	34
23-25	2	32
20-22	7	30
17-19	6	23
14-16 (Median Class)	**0**	**17**
11-13	9	17
8-10	7	8
5-7	1	1
	N = 34	

If you examine the above table for computing median you come across $\frac{N}{2}$ = $\frac{34}{2}$ = 17 against two Class Intervals showing cumulative frequencies. If one calculates Median mechanically, there is the possibility to arrive at a wrong answer. So, it will not be so simple to apply the formula and get the results. Also, if you calculate cumulative frequency from above, another class interval (17-19) may also show 17 as cumulative frequency. We will solve this question in two alternative ways (i) conceptually and (ii) empirically.

Solution 1 :

On counting number of cases from top and bottom sides we find that there are 17 cases above 16.7 and 17 cases below 13.5. We come across two points (instead of one) below and above which lie exactly fifty percent cases as the class interval 14-16 is void (without any frequency). So extending the assumption to this class, (that frequencies are evenly distributed within each class interval). We may add half of this void class interval to either side and find median.

So Median = 13.5 + 3/2 = 13.5 + 1.5 = 15 (Answer)

Or Median = 16.5 – 3/2 = 16.5 – 1.5 = 15 (Answer)

Solution 2 :

<table>
<tr><th>C.I.</th><th>f</th><th>c.f.</th><th>Modified C.I.</th><th>f</th><th>c.f.</th></tr>
<tr><td>26-28</td><td>2</td><td>34</td><td>26-28</td><td>2</td><td>34</td></tr>
<tr><td>23-25</td><td>2</td><td>32</td><td>23-25</td><td>2</td><td>32</td></tr>
<tr><td>20-22</td><td>7</td><td>30</td><td>20-22</td><td>7</td><td>30</td></tr>
<tr><td>17-19</td><td>6</td><td>23</td><td>15 to19.5</td><td>6</td><td>23</td></tr>
<tr><td>14-16</td><td>0</td><td>17</td><td rowspan="2">10.5 to15
(Median Class)</td><td rowspan="2">9 (fm)</td><td rowspan="2">17</td></tr>
<tr><td>11-13</td><td>9</td><td>17</td></tr>
<tr><td>8-10</td><td>7</td><td>8</td><td>8-10</td><td>7</td><td>8
(cfb)</td></tr>
<tr><td>5-7</td><td>1</td><td>1</td><td>5-7</td><td>1</td><td>1</td></tr>
<tr><td></td><td>N = 34</td><td></td><td></td><td>N = 23</td><td></td></tr>
</table>

Alternatively, we modify the class interval where N/2 may fall. The class interval with zero frequency which is affecting calculations is adjusted towards the adjoining class interval on either side and the size of the modified class interval is used while applying the formula. Half of the class interval 14-16 has been adjusted towards the class intervals 11-13 and 17-19 respectively. The modified class intervals are mentioned in terms of exact limits (modified size of class interval being 3 + 1.5 = 4.5).

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

L = 10.5

cfb = 8

N/2 = 17

fm = 9

$$Median = 10.5 + \frac{17 - 8}{9} \times 4.5$$

= 10.5 + 4.5 = 15 (Answer)

(***Source:*** *The special cases for calculation Median has been taken from, ES-333, IGNOU, 2010)*

14.6.3 Use of Median

Median is used in the following situations :

- If few scores are not known.
- When the point dividing the distribution into two equal parts is needed.
- When a distribution is markedly skewed.

14.6.4 Limitations of Median

The major limitations of Median are as follows:

- Individual scores are not considered in calculating the Median. Even in the absence of an individual score, Median can be calculated.
- Median is not an accurate measure of central tendency.
- It can not be used as the centre of gravity of the distribution.
- It can not be used for inferential statistic 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.

5. Define Median.

....

....

....

....

6. Median is related to which scales of measurement?

....

....

....

....

14.7 THE MODE

Mode as a measure of Central Tendency is used when the data are in nominal scale. The data obtained in nominal scale is of classificatory type and mostly qualitative in nature. We can classify the data, make categories in different heads and obtain frequencies. Mode is simply defined as the 'most frequently occurred item in the series'. For example, let 30 students are in a group, and in a test of Mathematics, the score and frequencies of the students are as follows:

Score	Frequencies
30	1
35	2
37	3
40	4
46	2
53	7
59	3
62	3
67	2
70	1
72	2
	N = 30

From the above distribution, among the 30 students, we find that seven students have secured the score 53. As per the definition of Mode, the score 53 which occurred most frequently in the distribution, is the mode. From the above table, you might have observed that, when we compute Mode of a distribution, we simply ignore the other scores. It implies that Mode is not an accurate or reliable measures of central tendency as because each and every score values are not considered or contributed for computing Mode. Again, there is also the possibility of more than one Mode in a distribution. If frequency of two different scores is highest and same in the group, there will be the situation of two Modes in that distribution. In such case, we can say that the distribution is bi-modal or in case three Modes are there in the group, we can say it is as tri-modal distribution. It is possible that a distribution may have more that one Mode but there exist only one Median and Mode in a distribution. Now, let us discuss, how to compute Mode.

14.7.1 Calculation of Mode for Ungrouped Data

Calculation of Mode in an ungrouped data is called as empirical or crude Mode. As discussed above, Mode is calculated in an ungrouped data by tallying the frequencies of the scores or items. In a simple ungrouped set of scores, the mode is the score which occurs most frequently. For example, if the scores in a group of 15 students are 30, 32, 41, 45, 40, 32, 40, 45, 49, 57, 59, 60, 59, 45 and 65. The Mode of the distribution will be 45, which occurs most frequently i.e. thrice in the group. Accordingly, Mode can be calculated in ungrouped data.

There is the possibility of more than one mode in a distribution. For example, if the individual scores in a distribution is 20, 23, 15, 33, 41, 54, 38 33, 23, 11, 10, 23, 51, 41, 50, 33, 45, 42, 15 and 38. In this distribution, the frequency of the scores 23 and 33 is 3 each and that is also highest. So in this distribution there are two modes i.e. 23 and 33. This is called as a bi-modal class. Accordingly tri-modal class is also possible if there are three modes in a distribution.

14.7.2 Calculation of Mode for Grouped Data

Calculation of Mode in group data is called as the true mode. It can be calculated by employing direct as well as indirect methods. In indirect method, mode is calculated by using the following formula:

Mode in Indirect Method:

Mode = (3 × Median) – (2 × Mean)

i.e., for calculating Mode, there is need to calculate Median and Mean first and than to calculate Mode. Let us take an example and calculate Mode in indirect method.

Example: Calculate Mode of the following distribution.

Class Interval	90-99	80-89	70-79	60-69	50-59	40-49	30-39	20-29	10-19
Frequencies	2	3	4	7	10	6	5	2	1

Solution:

Class Interval (C.I.)	Mid-point	Frequencies (f)	Cumulative Frequency (cf)	Deviation (d)	f d
90-99	94.5	2	40	+ 4	8
80-89	84.5	3	38	+ 3	9
70-79	74.5	4	35	+ 2	8
60-69	64.5	7	31	+ 1	7
50-59 (Median Class)	**54.5 (A.M.)**	**10**	**24**	**0**	**0**
40-49	44.5	6	14	- 1	- 6
30-39	34.5	5	8	- 2	- 10
20-29	24.5	2	3	- 3	- 6
10-19	14.5	1	1	- 4	- 4
		N = 40			$\sum fd = 6$

Median in Short Method:

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

L = 49.5

cfb = 14

N/2 = 20

fm = 10

c.i. = 10

$$Median = 49.5 + \frac{20 - 14}{10} \times 10$$

= 49.5 + 6 = 55.5

Mean in Assumed Mean Method:

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

A.M. = 54.5

N = 40

c.i. = 10

$\sum$fd = 6

$$Mean\ (M) = 54.5 + \frac{6}{40} \times 10$$

= 54.5 + 1.5

= 56

Mode = (3 × Median) – (2 × Mean)

= (3 × 55.5) – (2 × 56)

= 166.5 – 112

= 54.5 (Answer)

Mode in Direct Method:

In a grouped frequency distribution, it is unrealistic to calculate the Mode by looking only 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. f_1

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

Where, L = Lower Limit of the Modal Class

c.i. = Size of the class interval

f_1 = Frequency of the modal class

f_0 = Frequency of the class preceding of the modal class

f_2 = 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	90-99	80-89	70-79	60-69	50-59	40-49	30-39	20-29	10-19
Frequencies	2	3	4	7	10	6	5	2	1

Solution:

Class Interval (C.I.)	Frequencies (f)
90-99	2
80-89	3
70-79	4
60-69	**7 (f_2)**
50-59 (Modal Class)	**10 (f_1)**
40-49	**6 (f_0)**
30-39	5
20-29	2
10-19	1
	N = 40

Step 1 : Identification of the Modal Class [The class having the highest frequency, i.e. 50-59]

Step 2 : Lower Limit of the Modal Class 50-59 is 49.5

Step 3 : Calculation of 'f_1' (is the frequency of the Modal Class, i.e. 10)

Step 4 : Calculation of 'f_0' (is the frequency of just preceding the Modal Class, i.e. 6)

Step 5 : Calculation of 'f_2' (is the frequency of just succeeding the Modal Class, i.e. 7)

Step 6 : Size of the Class is 10.

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

$$= 49.5 + \frac{10-6}{2\times10-6-7} \times 10$$

$$= 49.5 + \frac{4}{7} \times 10$$

$$= 49.5 + 5.71$$

$= 55.21$ (Answer)

Now you can compare both the values of Mode, calculated by indirect method and direct method.

14.7.3 Use of Mode

Mode is used in the following situations:

- Mode is used, when it requires to find the most frequently occurred item as the measure of central tendency.
- When a quick and approximate measures of central tendency is required.
- When data is incomplete and skewed.

14.7.4 Limitations of Mode

The following are some of the limitations of Mode:

- Mode can only be used as a rough estimate of measures of central tendency, it can never be an accurate measure of central tendency.
- There may be possibility of more that one mode in a distribution, whereas, it is not possible for the case of Mean and Median.
- In case, all observations/scores in a group are different or there is no repetition of scores, it is difficult to get the Mode as all scores can be represented as Mode.
- Mode is only a crude measure which can be of value when a quick and rough estimate of central tendency is required.

Activity 2

Discuss with an example the possibility of more than one Modes in a distribution.

....

....

....

....

....

....

....

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. Mode is related to which scales of measurement?

....

....

....

14.8 COMPARISON OF MEAN, MEDIAN AND MODE

You have already discussed the concept, processes of calculation, uses and the limitations of Mean, Median and Mode in the previous sections. In this section, we will compare the three.

Table 14.2: 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 measure of central tendency.	Median is a approximate measure of central tendency.	Mode is a rough estimate of measure 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.

Activity 3

Critically analyse the use of Mean, Median and Mode with suitable examples.

....

....

....

....

14.9 SELECTION OF AN APPROPRIATE MEASURE OF CENTRAL TENDENCY

Selection of an appropriate average requires experience and insight to examine the nature of data and the purpose in hand. A careful study of the various sections of this chapter would guide you to select appropriate statistics to calculate the measures of central tendency. The scale of measurement on which data are available also plays an important role in selecting an appropriate average. By converting data from one scale to an other, different measures of central tendency can be used. The purpose of using the average should be kept in mind while selecting a measure of central tendency. For the research purpose, mean is popularly used as measure of central tendency as it the most accurate measure of central tendency.

14.10 LET US SUM UP

In this Unit, you studied about various methods of calculating measures of central tendency. In a lay person language, when we talk about average score it implies that to add all the scores secured by the students in a group and divide it by the total number of students. But when we elaborately study various methods of calculating the measures of central tendency, we find that for calculating Mean, Median and Mode, the formula are different even techniques are also different to calculate the measures of central tendency for ungrouped and grouped data. You have also studied, how the nature of the data determines the use of statistics and the scale of measurement that it deals with. It is therefore, one has to take a conscious decision about the use of various measures of central tendency and also their applicability.

14.11 REFERENCES AND SUGGESTED READINGS

IGNOU (2010). Measures of Central Tendency (Unit 13; Block 4), Educational Evaluation (ES-333, B.Ed.), New Delhi: IGNOU.

Garrett, H.E. (1973). Statistics in Psychology and Education, Vakils, Feffer and Simons Pvt. Ltd., Bombay.

Guilford, J.P. (1965). Fundamental Statistics in Psychology and Education, McGraw Hill Book Company, New York.

Nayak, B.K. and Rath, R.K. (2010). Measurement, Evaluation, Statistics and Guidance Services in Education, New Delhi: Axis Publications.

14.12 ANSWERS TO CHECK YOUR PROGRESS

1. (c) Equal interval scale
2. (a) Symmetrical
3. (b) Each score is given weightage to calculate Mean
4. (d) All the above
5. Median is the mid-point of the distribution, below and above which 50 percent of the cases lie.
6. Ordinal scale
7. Mode is defined as the most common item in the series.
8. Nominal scale

UNIT 15 MEASURES OF DISPERSION

Structure

15.1 INTRODUCTION

In the previous units, you must have studied about tabulation of graphical representation of data and measures of central tendency. The measures of central tendency is after all a single numerical value and may fail to reveal the data entirely. Thus, the next step to measures of central tendency is to know about measures of dispersion. In this Unit, you will understand the concept of dispersion or variability among the data. Variability may be understood simply in the following lines.

"Mean as a measure of central tendency describes only one of the important characteristics of a distribution of scores but, it is equally important to know how compactly the scores are distributed from this point of location or conversely how far they are scattered away from it, the latter explains the concept of variability."

For example, the mean of the following two sets of scores are equal i.e. 10, however the spread of scores are different in two groups, even the range of highest and lowest scores is also different.

Group I : 8, 12, 11, 12, 10, 8, 9, 11, 12, 10, 8, 10, 9, 10, 12, 8, 10, 9, 10 and 11.

Group II : 15, 2, 8, 12, 4, 17, 20, 6, 2, 18, 16, 0, 3, 9, 6, 10, 15, 17, 9 and 11.

In the above example, the two groups are not comparable in terms of homogeneity, even if the mean scores of two groups are the same. The dispersion or the variability of the scores of the groups are different.

In the present Unit, we will discuss measures of dispersion and measures of position, its calculation, interpretation and use in classroom situation. Mastery on such concepts will make you able to understand the spread and variability of abilities of the students in your class individually as well as in a group.

15.2 OBJECTIVES

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

- understand the measures of position;
- differentiate between percentile and percentile rank;
- calculate the percentile and percentile rank;
- interpret the obtained values of percentile and percentile rank;
- understand the concept of dispersion;
- state the importance of the measures of dispersion;
- define and calculate different measures of dispersion viz. – range, quartile deviation, mean deviation, and standard deviation; and
- use appropriate measures of dispersion according to the need of classroom situations.

15.3 MEASURES OF POSITION

The purpose of Statistics is to help you understand the data. Therefore, until we know the position of a particular score in any group, that score remain meaningless for us, e.g., if Rajesh got 85 marks in Physics out of 100, the score informs you of the achievement of Rajesh, but if you want to judge the ability of Rajesh in Physics, then you need to know his relative position in the group; may be after securing 85 marks he obtained last position in the group, all other members of the group might have secured more than 85 marks or else Rajesh might be standing in the middle or may be at the top position. Therefore, to know the ability of Rajesh with respect to his class, it is necessary to know that among all scores, what is the position of 85. For example, if percentage of students who scored less than 85 marks is 20%, then it is clear that Rajesh is one among the below 20% achievers. Those scores which possess specific position in the group are known as measures of position.

15.4 PERCENTILE AND PERCENTILE RANK

There are several measures of position e.g. Percentile, Decile, Quartile, etc., but in this Unit, you will study the commonly used measures of position i.e. percentile and percentile rank.

15.4.1 Concept of Percentile and Percentile Rank

In measures of central tendency you have studied about median. Median 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.

Similarly, we may define percentile and percentile rank in the following words:

Percentile	The Kth percentile of a given score in any distribution is the point on the score scale below which 'K' percent of the scores fall.
Percentile Rank	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.

15.4.2 Calculation of Percentile and Percentile Rank

i. Percentile

In order to calculate the values of percentile, one needs to locate the points on the scale of measurement upto which the given percent of cases lie. For calculating percentile, the formula which we use to calculate median are used.

The formula is:

$$P_p = l + \left(\frac{pN - Cf_B}{f_p} \right) \times i \quad \text{........ (1)}$$

Where,

P = Percentile

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

l = Exact lower limit of the class interval upon which Pp lies

pN = Part of N

Cf_B = Cumulative Frequency below l

Example: Calculate P_{10}, P_{20}, P_{30}, P_{40}, P_{50}, P_{60}, P_{70}, P_{80}, P_{90}, for the following data:

Table: 15.1

Scores	*f*	*c.f.*
10 – 19	2	2
20 – 29	4	6
30 – 39	5	11
40 – 49	10	21
50 – 59	35	56
60 – 69	20	76
70 – 79	13	89
80 – 89	8	97
90 – 99	3	100
	N = 100	

Solution:

$$P_{10} = 10\% \text{ of } 100 = 10 \rightarrow = 29.5 + \left(\frac{10-6}{5}\right) \times 10 = 37.5$$

$$P_{20} = 20\% \text{ of } 100 = 20 \rightarrow = 39.5 + \left(\frac{20-11}{10}\right) \times 10 = 48.5$$

$$P_{30} = 30\% \text{ of } 100 = 30 \rightarrow = 49.5 + \left(\frac{30-21}{35}\right) \times 10 = 52.07$$

$$P_{40} = 40\% \text{ of } 100 = 40 \rightarrow = 49.5 + \left(\frac{40-21}{35}\right) \times 10 = 54.92$$

$$P_{50} = 50\% \text{ of } 100 = 50 \rightarrow = 49.5 + \left(\frac{50-21}{35}\right) \times 10 = 57.78$$

$$P_{60} = 60\% \text{ of } 100 = 60 \rightarrow = 59.5 + \left(\frac{60-56}{20}\right) \times 10 = 61.5$$

$$P_{70} = 70\% \text{ of } 100 = 70 \rightarrow = 59.5 + \left(\frac{70-66}{20}\right) \times 10 = 66.5$$

$$P_{80} = 80\% \text{ of } 100 = 80 \rightarrow = 69.5 + \left(\frac{80-76}{13}\right) \times 10 = 72.57$$

$$P_{90} = 90\% \text{ of } 100 = 90 \rightarrow = 79.5 + \left(\frac{90-89}{8}\right) \times 10 = 80.75$$

Thus, the value of P_{10}, P_{20}, P_{30}, P_{40}, P_{50}, P_{60}, P_{70}, P_{80}, P_{90}, is 37.5, 48.5, 52.07, 54.92, 57.78, 61.5, 66.5, 72.57, 80.75 respectively.

ii. Percentile Rank (PR)

You have seen, how percentiles e.g., P_{10} to P_{90} have been calculated directly from any given frequency distribution and based on the above calculation we may say that, "percentiles are points in a continuous distribution below which lie given percentages of N". Now you can calculate the problem of finding an individual's percentile rank (PR). The following formula may be used for calculating the PR :

$$PR = \frac{100}{N} + \left(Cf_B + \frac{X - l}{i} f\right) \qquad \text{........ (2)}$$

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

i = Class Interval

N = Total frequency

Example

Calculate the PR for the score 35 and 55 for the distribution of scores given in Table '15.1'.

(i) for X = 35;

X = 35, Cf_B = 6, l = 29.5, f = 5, i = 10, N = 100

$$PR = \frac{l00}{100} + \left(6 + \frac{(35 - 29.5)}{10} \times 5\right)$$

= 8.75

∴ PR of 35 is 8.75

(ii) for X = 55;

X = 55, Cf_B = 21, l = 49.5, f = 35, i = 10, N = 100

$$PR = \frac{l00}{100} + \left(21 + \frac{(55 - 49.5)}{10} \times 35\right)$$

= 40.25

∴ PR of 55 is 40.25

15.4.3 Interpretation of Percentile

Percentile is a set of measure used to indicate the relative position of a single item of individual in context with the group to which the item of individual belongs. In other words, it may be said that it is used to indicate the relative position of a given score among other scores. As you have already studied that percentile refers to a point in a distribution of scores or values below which a given percentage of cases lie, therefore 85% of the observations are below the 85th Percentile, which may be denoted as P_{85}.

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 Percentile Rank and not on the basis of scores they have secured. For any standardized tests, percentile norms are given with the test and, therefore, with the help of norms one may interpret the results properly.

15.4.4 Limitations

In spite of having certain important and relevant characteristic of knowing the ability of an individual in any group based on their achievement scores, the percentile has some limitations which you need to understand, viz., the mastery of an individual may not be judged by the percentile, because the same person in a poor group may perform better and may secure good rank, while the person in a better group may perform poor and may secure poor rank.

Check Your Progress 1

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

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

1) (i) Define Percentile and Percentile Rank.

....

....

....

(ii) Calculate Percentile Rank of 34 for the following scores given below:

Scores Interval	70–79	60–69	50–59	40–49	30–39	20–29	10–19	0–9
f	4	3	5	6	3	5	2	2

....

....

....

15.5 MEASURES OF DISPERSION

You have already studied that the purpose of statistics is to describe the characteristics of any variable with reference to any group. Measures of central tendency provide some meaningful information on any data but it is not at all sufficient to make a holistic and comprehensive view e.g. if there are three groups having eight students in each who secured the following marks on any achievement test:

A:	20	20	20	20	20	20	20	20
B:	10	14	18	20	10	52	20	16
C:	20	20	20	40	60	00	00	00

It is evident from the data that 'mean score' for all the three groups are the same, but whether the groups are also the same? Think! Equal mean score i.e. 20 for all the three data sets informs you that all the members of a group are varying around '20', but when you see the group scores individually, you find a significant variation in scores, and this variation is because of spreadness or so to say dispersion. Therefore, in order to make any concrete decision for the group, you need to know the dispersion that exists among the data. This dispersion may be known by applying the any of the following measures:

i. Range

ii. Quartile Deviation

iii. Mean Deviation

iv. Standard Deviation

15.5.1 Range

Range may be defined simply as that interval between the lowest and the highest scores. It is very common and general measure of spread and is computed when we need to know at a glance comparison of two or more groups for variability. Since, it is based on two extreme values and tells nothing about the variation of the intermediate values, it is not an authentic measure of dispersion. Range may be computed by the following formula:

(Range = Highest Score – Lowest Score) .... (3)

It is deceptive and not authentic. For example, in any class of 40 students, 1 student got 20 marks, while all the rest 39 students scored between 70 to 80 marks out of 100. Range informs you that there exist a range of 60 marks, while the majority secured between 70 to 80 and the more appropriate and near variation is of 100 marks only. Therefore, it may be used as a quick and at a glance measure of dispersion, but you cannot be dependent on "Range" in order to know true dispersion.

15.5.2 Quartile Deviation and its Interpretation

The second measure of measures of dispersion is Quartile Deviation. It is also known as 'semi-interquartile range'. As you know the interval between highest and lowest score 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:

$$\text{QD} \quad \text{or} \quad Q = \frac{Q_3 - Q_1}{2} \qquad \text{........ (4)}$$

Since you are already aware on the concept of percentile, therefore, you may simply infer the formula in the following manner:

$$\text{QD} \quad \text{or} \quad Q = \frac{P_{75} - P_{25}}{2} \qquad \text{........ (5)}$$

In order to calculate Q, it is clear that we must first compute the 75^{th} and 25^{th} Percentile and, therefore, the formula (1) previous discussed for calculating percentile may be used in the following manner:

$$Q_1 = l + \left(\frac{\frac{N}{4} - Cf_B}{f_q} \right) \times i \qquad \text{........ (6)}$$

and

$$Q_3 = l + \left(\frac{\frac{3N}{4} - Cf_B}{f_q} \right) \times i \qquad \text{........ (7)}$$

Note: You must remember that formula (1), used for Percentile has been taken from the formula which you used in calculating the median.

Example: Calculate the Quartile Deviation for the data given in Table 15.1.

Solution: Using the data given in Table 15.1 and Formula (6) and (7):

(i) $Q_1 = P_{25} = 49.5 + \left(\frac{25 - 21}{35} \right) \times 10 = 50.64$

(ii) $Q_3 = P_{75} = 59.5 + \left(\frac{75 - 56}{20} \right) \times 10 = 69$

Therefore, $\text{QD} = \frac{Q_3 - Q_1}{2} = \frac{69 - 50.64}{2} = 9.18$

Q = 9.18

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 – Median = Median – Q_1

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

I. Q_3 – Median > Median – Q_1 (Positive Skewed Curve)

II. Q_3 – Median < Median – Q_1 (Negative Skewed Curve)

15.5.3 Mean Deviation and its Interpretation

The mean deviation (MD), also known as average deviation (AD), is the mean of all deviations of all individual and separate scores in a distribution taken 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):

$$MD = \frac{\sum |x|}{N}$$ (For ungrouped data) .... (8)

$$MD = \frac{\sum |fx|}{N}$$ (For grouped data) .... (9)

Where, ($x = X - M$)

Example: Calculate the mean deviation for the distribution given in Table 15.1 of this unit:

Scores	***X* (Mid-point)**	***f***	***fX***	$\|x\| = \|X - M\|$ (M=58.7)*	$\|fx\|$
10 – 19	14.5	2	29	44.2	88.4
20 – 29	24.5	4	98	34.2	136.8
30 – 39	34.5	5	172.5	24.2	121
40 – 49	44.5	10	445	14.2	142
50 – 59	54.5	35	1907.5	4.2	147
60 – 69	64.5	20	1290	5.8	116
70 – 79	74.5	13	968.5	15.8	205.4
80 – 89	84.5	8	676	25.8	206.4
90 – 99	94.5	3	283.5	35.8	107.4
		N=100	$\sum fX = 5870$		$\sum \|fx\| =$ 1270.40

* Calculation of Mean Score you have already learned in Unit-14 of this Block.

$$\text{MD} = \frac{\sum |fx|}{N} = \frac{1270.40}{100} = 12.704$$

Therefore, **MD = 12.704**

Interpretation of Mean Deviation

Unlike 'Range' and 'Quartile Deviation', the mean deviation is the simplest measure of dispersion that takes into account all the scores of distribution. In spite of having this feature, mean deviation is rarely used in modern statistics, however, you will find it in the older literature. Because this method, it ignores the importance of plus and minus sign. Therefore, the method doesn't fulfill the assumptions of algebraic properties and hence it is not possible to use this method in higher statistics.

Check Your Progress 2

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

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

2. (i) Define Average Deviation.

(iii) Calculate Quartile Deviation for the data given below:

C.I	190-199	180-189	170-179	160–169	150–159	140–149	130–139	120–129	110–119
f	6	14	22	36	42	15	8	6	4

15.5.4 Standard Deviation

Standard deviation (SD) is commonly and frequently used measure of dispersion, because SD is the most consistent and stable index of variability and is usually used in experimental research. Deviations of all scores taken from their mean and then square root of their average, is known as, Standard Deviation. The standard deviation differs from mean deviation in many respects. In calculating the mean deviation, we ignore signs and treat all deviations as positive, whereas in standard deviation we avoid such complexities of signs by squaring the individual deviations. Also, unlike mean deviation, in the standard deviation, the squared deviations used in the process are always taken from the mean only, never from the median or mode. The commonly acknowledged symbol for the standard deviation is the Greek letter 'σ' (known as Sigma).

In conclusion, it may be said that 'The sum of the squared deviations from the mean, divided by N, is known as the variance and the square root of the variance is known as standard deviation'.

Following are the formula, which you may use according to your convenience and the nature of the data:

a) $$\sigma = \sqrt{\frac{\sum x^2}{N}} \qquad \text{........ (10)}$$

where (x = X – M) (For ungrouped data)

b) $$\sigma = \sqrt{\frac{\sum fx^2}{N}}$$ .... (11)

where (x = X – M) (For grouped data)

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

c) $$\sigma = \sqrt{\frac{\sum fx^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$ .... (12)

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

d) $$\sigma = i\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$$ .... (13)

(**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
- i: class-interval
- N: Total frequency

Now, we will learn the computation of standard deviation using formula 10, 11, 12 and 13.

Example: (For ungrouped data)

(1) Calculate the Standard Dispersion for the following data:

7, 8, 11, 14, 15

Solution:

X	***x = X – M***	***x²***
7	7 – 11 = – 4	16
8	8 – 11 = – 3	9
11	11 – 11 = 0	0
14	14 – 11 = 3	9
15	15 – 11 = 4	16
$\sum X = 55$	$\sum X = 0$	$\sum x^2 = 50$

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

$$= \sqrt{\frac{50}{5}} = \sqrt{10}$$

$$\sigma = 3.16$$

Example : (For grouped data)

2. Calculate the standard deviation for the distribution given in Table 15.1 of this Unit.

Example: Calculate the mean deviation for the distribution given in Table 15.1 of this unit:

Scores	***f***	***X* (*Mid-point*)**	***fX***	$f.X^2$	$x = X - M$	$x^2 = (X-M)^2$	$fx^2 = (X-M)^2$
10 – 19	2	14.5	29	420.5	44.2	1953.64	3907.28
20 – 29	4	24.5	98	2401	34.2	1169.64	4678.56
30 – 39	5	34.5	172.5	5951.25	24.2	585.64	2928.2
40 – 49	10	44.5	445	19802.5	14.2	201.64	2016.4
50 – 59	35	54.5	1907.5	103958.75	4.2	17.64	617.4
60 – 69	20	64.5	1290	83205	5.8	33.64	672.8
70 – 79	13	74.5	968.5	72153.25	15.8	249.64	3245.32
80 – 89	8	84.5	676	57122	25.8	665.64	5325.12
90 – 99	3	94.5	283.5	23790.75	35.8	1281.64	3844.92
	N = 100		$\sum fX =$ 5870	$\sum fX^2 =$ 371805			$\sum fx^2 =$ 27236

Calculation of Standard Deviation using different formulas.

A. Formula 11.

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

$$= \sqrt{\frac{27236}{100}}$$

$$\sigma = 16.50$$

B. Formula 12.

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

$$= \sqrt{\frac{371805}{100} - \left(\frac{5870}{100}\right)^2}$$

$$= \sqrt{3718.05 - 3445.69}$$

$$\sigma = 16.50$$

Example: (Using Short Method)

Calculate the Standard Deviation for ‘Table 15.1’ given in this Unit, using short method.

15.5.5 Sheppard's Correction

When σ is computed from a frequency distribution, the scores in each interval are represented by the midpoint of that interval. Fair possibility of equal distribution and symmetry cannot be ensured every time and, therefore, frequencies tend to lie below the midpoint more often than above in the intervals above the mean of the distribution whereas in intervals below mean, the scores tend to lie above the midpoints. These errors may be taken care of if mean is calculated from all the intervals. But the 'grouping error' may inflate σ and situation may become worst, if the intervals are wide and N is small. In order to prevent the error and adjust for grouping 'Sheppard's Correction' is frequently used. The formula for Sheppard's correction is as follows (the value 12 used in this formulla is a constant):

$$\sigma = \sqrt{\sigma^2 - \frac{i^2}{12}} \quad \text{.... (14)}$$

For example, if you apply this formula for the aforesaid result, it will be:

$$\sigma = \sqrt{(16.50)^2 - \frac{(10)^2}{12}}$$

$$\sigma = \sqrt{272.25 - 8.53}$$

$$\sigma = 16.24$$

∴ Value of SD after applying correction formula is 16.24.

15.5.6 Interpretation of Standard Deviation

Standard deviation is commonly and widely used statistics in data analysis and research. Like mean deviation, it utilizes all the values of the distribution. This is called as accurate measure of dispersion because as it gives equal weightage to each and every score in the series. Standard deviation may be seen as spine of statistics, it is the master measure of dispersion and amenable to algebraic operations and also we use it in corelational studies and in further statistical analyses like ANOVA and ANCOVA. The standard deviation is less affected by sampling errors than Q and MD.

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.

3) (i) Define the relation between 'Variance' and Standard Deviation'.

....

....

....

(ii) Calculate Standard deviation for the following set of scores:

40, 38, 42, 60, 72, 54

....

....

....

15.6 USE OF MEASURES OF POSITION AND DISPERSION

Needless to mention the important and significant role of 'assessment' in progress and advancement of student as an individual and classroom and institutions as a group, also needless to highlight the role of statistics in the process of assessment of a student and a class. In classroom situation the need of assessment has been highlighted by almost all the committees and commissions set up in independent India. It is a commonly acknowledged fact that we need to use some basic statistics in classroom situation frequently to know the relative and comparative progress of a student as well as a class over a period of time. However, majority of the time, it has been observed that the teacher uses the simplest measures of central tendency to highlight students' achievement and tries to avoid further analysis which can give an overall and clear-cut picture of the progress of an individual student in the class and of the class as a whole.

The following points may be kept in mind:

- In a classroom situation a teacher must calculate the achievement of a student using suitable measures of central tendency but the teacher must put the individual scores in the form of percentiles and percentiles ranks to know the achievement along with the student's relative position in the class.
- There are various measures of dispersions, Range, Quartile deviation, mean deviation and standard deviation and often the question arises which one is to be used in the classroom situation. Needless to mention the strength of standard deviation, however, other measures are also to be used, keeping in view the need of assessment. A brief description is given here to understand the use of various measures of dispersion.

Range:

(i) When you need quick, crude and at a glance measure.

(ii) When the scores spread widely.

(iii) When you need information from extremes only.

Quartile Deviation:

(i) When median has been used as a measure of central tendency.

(ii) When score is lying both sides of central score are of prime importance.

(iii) When extreme scores are affecting the Standard Deviation unexpectedly.

Mean Deviation

In general, it is avoided because mean deviation ignores the importance of sign and, therefore, it lacks algebraic properties.

Standard Deviation

Strength and effect of SD is well known to us, because it satisfies the assumptions of algebraic operations SD may be used:

(i) When you need to know the most reliable measure of variability.

(ii) When mean has been used as measure of Central tendency.

(iii) When you want to consider all the scores according to their size.

(iv) When you want to use higher statistics, viz. Correlation, ANOVA, ANCOVA, etc.

These measures of dispersion informs a teacher of the variability in class and hence it helps the teacher to prepare personalized instruction to address the need of deprived students. In order to be more authentic and comprehensive three important concepts viz. standard scores, the coefficient of variation and variance are being discussed here as follows:

Standard Scores

The raw scores of any text are no more than simple arithmetic scores having no meaning in itself until and unless it is compared in some frame of reference. This frame of reference we get after calculating mean and SD of the whole group. Therefore, in order to give meaning to raw scores, it needs to be changed in some standard score. Standard scores carry their meaning in themselves and comparisons between different individuals are possible with them. The commonly known standard scores are : Z=score and T-Score.

Z-Score

Z-Score is the oldest and commonly known standard scores. Z-Scores are the scores taken from the deviation of each score from mean and divided by SD. Therefore, it may be expressed as:

$$Z = \frac{X - M}{\sigma}$$

Where, X – Raw Score, M – Mean, σ – SD.

However, Z-Score provides sufficient information but keeping in view, few limitations of getting Z-Score positive or negative and its expression upto two point of decimal some people try to avoid it.

T-Score

In order to overcome the difficulties faced in Z-Score experts suggest T-Score. T-Score is basically a linear transformation of Z-Score. The word 'T' here used to pay respect to 'Thorndike'. You may calculate T-Score as follow:

$$T = 50 + 10\ Z$$

Coefficient of Variation

It is very useful in comparing the standard deviation of many groups exposed to similar test i.e. when many groups are given similar test and their means are almost equal, than with the help of CV the deviation between groups may be calculated. CV may be calculated as:

$$CV = \frac{\sigma}{M} \times 100$$

Variance

The square of SD is known as variance.

$$V = \sigma^2 = \frac{\sum (X - M)^2}{N} = \frac{\sum x^2}{N}$$

Keeping in view its suitability and fitness on algebraic properties it is mostly used in higher statistics.

15.7 LET US SUM UP

- Percentile is nothing but a sort of measure used to indicate the relative position of an individual with reference to the group to which he/she belongs.

- Commonly used measures of dispersion are Range, Quartile Deviation, Mean Deviation and Standard Deviation.
- Range is used when scores are spread widely and need quick, crude and at a glance measure.
- Quartile deviation is used when median has been used as a measure of central tendency.
- Standard Deviation is the most consistent and stable index of variability. SD is used when mean has been used as measure of central tendency.
- Raw scores are always compared with some frame of reference or standard scores. The commonly known standard scores are Z-score and T-score.

15.8 REFERENCES AND SUGGESTED READINGS

Blommers, P. and Lindquist, E.F. (1959). Elementary Statistical Methods. Houghton 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.

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.

Walker, H.M. and Lev, J. (1965). Elementary Statistical Methods, Oxford & IBH Publishing Co., Calcutta.

Wine, R.L. (1976). Beginning Statistics, Winthrop Publishers Inc., Massachusetts.

15.9 ANSWERS TO CHECK YOUR PROGRESS

1. (i) A percentile is a point on the score scale below which a given percent of the cases lie. Whereas Percentile Rank is defined as the number representing the percentage of the total number of cases lying below the given score.

   (ii) $N = 30,\ X = 34,\ L = 29.5,\ cf_b = 9,\ i = 10,\ f = 3$

   So, $PR = 34.5$

2. (i) Average Deviation is measured by the mean deviation of all separate scores in the series taken from their mean.

   (ii) $Q_1 = P_{25} = 150.62$

   $Q_3 = P_{75} = 171.54$

   So, $Q.D. = \frac{Q_3 - Q_1}{2} = 10.46$

3. (i) The square root of variance is known as Standard Deviation.

   (ii) $\Sigma x = 306,\ N = 6,\ M = 51,\ \Sigma x^2 = 902$

   Hence, $SD = 12.26$

UNIT 16 CORRELATION : IMPORTANCE AND INTERPRETATION

Structure

16.1 INTRODUCTION

In Units 13 & 14, we have discussed those statistical measures that we use for a single variable i.e. the distribution relating to one quantitative variable. Now, we shall study the problem of describing the degree of simultaneous variation of two or more variables. The data in which we secure measures of one variable for each individual is called a univariate distribution. If we have pairs of measures on two variables of each individual, the joint presentation of the two sets of scores is called a bivariate distribution.

You may want to study the possibility of a relationship between two variables and see what kind of relationship exists. For example if you want to study the relationship between height and weight - whether the change in one will bring a change in other or not. Or if you want to find the relationship between hours of study and achievement, sex and enrolment etc., you can do so by finding correlation 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. According to Tuttle, "*Correlation is an analysis of the co-variation between two or more variables*". With the change in one variable, the other related variable changes is known as correlation. You know that as the child grows, the height as well as weight increases. Similarly, if a person has more height than the other, the person is likely to have more weight than the latter and we can say height and weight are positively correlated. You will study about correlation in greater details in this unit.

16.2 OBJECTIVES

After reading this Unit, you will be able to:

- define correlation;
- define co-efficient of correlation;
- recognize various types of correlation;
- calculate the co-efficient of correlation according to nature of scores and their distribution;
- interpret the results obtained i.e. interpret the Co-efficient of correlation;
- take necessary precautions in interpreting the co-efficient of correlation; and
- discuss the importance of correlation and its coefficient.

16.3 THE CONCEPT OF CORRELATION

Let us take an example of the scores of 5 students in mathematics and physics. What pattern do you find in the data? You may notice that in general those students who score well in mathematics also get high scores in physics. Those who are average in mathematics get just average scores in physics and those who are poor in mathematics get low scores in physics. In short, there is a tendency for students to score at par on both variables. Performance on the two variables is related; in other words the two variables are related, hence co-vary.

If the change in one variable appears to be accompanied by a change in the other variable, the two variables are said to be co-related and this inter-variation is called correlation.

You must have understood from the above discussion that correlation is a statistical technique that helps us know whether there is any relationship or not between any two pairs of variables and how strong is this relationship. This relationship is perfect or not can be determined by further analysis of observations. This is a measure of simultaneous variation of variables described by an integer. This can also be called the *measure of relationship*. If you take measure of every subject (individual) of your sample on two different variables e.g. if you take the marks of every student in Urdu and Mathematics in a Class VI of your school and find out the relationship between these marks of Urdu and Mathematics, relationship which is found out is called correlation. Such a data where we study two variables at a time is called bi-variate data. In this kind of data scores of one individual in one subject could be paired with the scores of the same individual in another subject. In this case scores of Urdu can be paired with scores of Mathematics. This relationship is important as the scores of the individual may change with the scores of other or the relationship could be due to some common factor between them.

You will find that when one of these variables increases or decreases on observation, the other paired variable is also changed proportionately. However, it may not be taken as change of one variable with the manipulation of other variable. When you are using correlation, you should not assume that a change in one variable causes a change in the other. It is not a cause and effect relationship. Even a high degree of correlation does not necessarily mean that a relationship of cause and effect exists between the variables. Thus the *correlation is simply a degree of relationship between the variable of a bi-variate data*. This correlation can tell you just how much of the variation in one variable is related to the other paired or correlated variable. Like all statistical techniques correlation is only appropriate for certain kind of data. Correlation works only with the quantifiable data in which numbers are meaningful, usually in quantities of some sort. It cannot be used for purely non-quantifiable data like categorical data such as gender, socio economic status, goodness, etc. Correlation is just a co-variation and does not manifest any kind of causation of functional relationship.

16.4 CO-EFFICIENT OF CORRELATION

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 changes with the variations in the other.

This coefficient of correlation is determined to find the relationships because it is simple to understand and convenient to express. It is a constant.

Symbol of co-efficient of correlation

The coefficient of correlation is always symbolized either by r or r (Rho). The notion 'r' is known as product moment correlation co-efficient or Karl Pearson's Coefficient of Correlation. The symbol 'r' (Rho) is known as Rank Difference 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.

16.4.1 Maximum Range of Values of Co-Efficient of Correlation

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

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 is Co-efficient of Correlation?

   ....

   ....

   ....

3. What can be the possible range of Co-efficient of correlation?

   ....

   ....

   ....

16.5 TYPES OF CORRELATION

Correlation can be classified in several different ways. Three most common ways of classification can be:

(i) Positive and negative Correlation

(ii) Simple, partial and multiple Correlation

(iii) Linear and non-linear Correlation

Before we come to the classified types of correlation, many times you will come across two more terms about correlation i.e. (i) no correlation and (ii) perfect correlation. We will study about them first, before we come to types of correlation.

(i) ***No Correlation-*** when there is no relationship observed or established between two variables i.e. the change in one variable cannot establish the change of other variable in a definite pattern. It may first time increase, second time decrease and at third time may not change. In such cases, we can say no definite change or Zero correlation. Most of the times one variable changes and other remains constant. Sometimes, it may even be a chance that there is a relationship.

(ii) ***Perfect Correlation –*** When change in one variable brings similar change in the other variable under paired variate, it can be said to have perfect correlation, that is, it has correlation value = one. In such cases the variable changes in a constant proportion. Specifically, when the coefficient correlation is ‘+1’, it is called as perfect positive correlation and when it is ‘-1’, it is called as perfect negative correlation.

Now let us come to classified correlation.

16.5.1 Positive and Negative Correlation

As the name clearly suggests they are directional associations or correlations which means the two variables of a bi-variate data vary in the same direction or the different direction. You will understand it better with the following explanations about them one by one.

(i) Positive Correlation**:** If both the paired variables under study vary in the same direction with any 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 direction of change. For example, height is positively correlated to weight. Similarly, height and weight are positively correlated to age. In education, the motivation or intelligence may be positively correlated to achievement. Other examples could be reading or study habits and achievement, attendance and achievement, etc. This positive correlation can further be explained with the following example.

Example:

In a class the marks of ten students were studied in Urdu and History to know whether the marks in Urdu have some relationship with the marks obtained in History or to know the correlation between marks of two subjects. We got the marks of ten students as under:

Student	Marks in Urdu	Marks in History
1	25	35
2	20	30
3	35	45
4	28	39
5	45	50
6	36	47
7	50	65
8	18	25
9	12	20
10	40	48

Now you can see that the marks of Urdu and History achieved by all the ten students are moving in the same direction i.e. increasing/decreasing in the same order. Thus, we say achievement of students in Urdu and History is positively correlated.

(ii) Negative Correlation: If the two paired variables under study vary in opposite direction with any change, then 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 have opposite direction of change. For example, the supply and the prices of commodities. You know, if the commodity is available in abundance, the prices fall and if it is in shortage, the prices will rise. The same could be the case with absenteeism and achievement. More the students absent themselves from the school, the lesser will be the achievement and vice versa. These are common examples which must have clarified the concept of negative correlation. You could give some more examples of negative correlation at your end. Negative correlation is not very common in our studies but it may appear, needing explanation in many of your studies.

16.5.2 Simple, Multiple and Partial Correlation

This classification is based on the number of variables being studied.

(i) ***Simple Correlation:*** In simple correlation only two variables are studied. For example, in most of bi-variate studies like the correlation between height and weight, correlation between the hours of study and achievement, correlation between the method of teaching and achievement of students etc., we have to find simple correlation.

(ii) ***Multiple Correlation:*** In multiple correlation three or more variables are studied simultaneously. For example, when you want to study method of teaching, sex, social-economic status and achievement and find the correlations among them simultaneously, you have to see the effect of interactions among the variables along with correlation and these are studied through multiple correlations.

(iii) ***Partial Correlation:*** Partial correlation is also used in study of correlation of more than two variables, but in this case it is supposed that we want to study only two variables which are said to be influencing each other, the effect of other influencing variables being kept constant. For example, if we want to study the effect of hours of study and achievement but the third variable intelligence is all the time influencing, then its effect has to be kept constant or partialled out.

16.5.3 Linear and Curvilinear Correlation

(i) ***Linear Correlation:*** Most of the correlations you will deal with in your research studies will be linear correlations and in common sense correlation is generally understood by the term simple linear correlation only. If the amount of change in one variable tends to bear a constant ratio to the amount of change in the other related variable, then the correlation is said to be linear correlation.

Example:

The scores obtained by students of Class VIII on two variables say intelligence and achievement are as under:

Intelligence (IQ)	*Achievement*
50	25
60	30
70	35
80	40
90	45

Now it is clear from the above data that the ratio of change between one variable to the other is the same. If such a data of the variable quantities is plotted on a graph, you will get a straight line.

(ii) Curvilinear Correlation (Non – Linear Correlation): Now it would be clear that if the amount of change in one variable does not bear a constant ratio with the amount of change in the other related variable, the correlation is said to be curvy – linear Correlation.

Example:

The following data in hours of study and the achievement scores of five students are given below :

Hours of Study	Achievement Score
1	60
2	65
3	72
4	78
5	83

We will find a non-linear relationship between the two variables. If we plot a graph of such a data, we would get a curve.

Normally, it is assumed that a linear relationship exists between the two variables under study because the techniques of correlation for curvi-linear correlation are complicated and do not bring much difference.

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. Suggest examples of the following:

a) Perfect positive correlation.

....

....

....

b) Zero correlation.

....

....

....

c) Curvilinear correlation.

....

....

....

16.6 METHODS OF COMPUTING CO-EFFICIENT OF CORRELATION (UNGROUPED DATA)

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

1. Rank Difference Co-efficient of Correlation or Spearman's Rank Order Co-efficient of Correlation.

2. Pearson's Product Moment Co-efficient of Correlation.

3. Pearson's Product Moment Co-efficient of Correlation from Scatter Diagram.

16.6.1 Spearman's Rank Difference Coefficient of Correlation (ρ)

Rank Difference Coefficient of Correlation is used only when the data is in 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 characters. This is the most widely used because of its easiness to compute. This is denoted by the symbol rho (ρ) and is computed as follows:

1. First of all the data being discrete is converted into Rank form. For ranking the data, we have to start ranking from the highest to the lowest. Highest score is give rank value as 1 and the next is given the rank 2 and so on. We begin with giving ranks to first set of scores followed by giving ranks to the next set of scores separately. Suppose there are two individuals getting the same score, then their positional ranks are added and an average rank is given to each individual. For example, if 6^{th} and 7^{th} highest ranked individuals are getting the same score, then their ranks are added and average is taken which is 6+7/2 = 6.5. So each of them is given a rank of 6.5. Similarly, if three individuals get the same score, then their positional ranks are added and average is taken. After completing the ranking of first set of data, the second set of data is also awarded ranks in the same manner. Although too many such ties will affect the size of correlation coefficient; but usually there are not so many ties to justify the formula that is available for correcting these ties.

2. After giving ranks to both the set of variables, find the difference between these ranks of two set of scores. This is calculated as an absolute score because the sign of these differences does not effect and is of no importance as the differences have to be squared in the next step.

3. Square each of these differences in the next column.

4. Add all these difference squares.

5. Find the Rank Order Coefficient of Correlation 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

Let us understand this with an example to illustrate it further.

Example:

Suppose you have administered a Personality test and a Mental Ability test to some students of your school and you are interested in finding the degree of association between personality and mental ability. Let us assume that you got the following scores on the two tests. These scores are taken for convenience only.

Scores of students on Personality Test and Mental Ability Test

S.No. of Student	Score on Personality	Score on Mental Ability
1	60	60
2	54	68
3	53	40
4	49	52
5	49	51
6	47	38
7	46	51
8	45	32
9	45	39
10	45	41
11	43	50
12	41	48
13	39	36
14	38	48
15	32	40
16	32	46
17	30	37

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

Table 16.1 : Table for Computation of Rank Order Coefficient of Correlation

Student	Personality Score	Mental Ability Score	Rank Personality (R_1)	Rank Mental Ability (R_2)	D = R_1-R_2	D^2
1	60	60	1	2	1	1
2	54	68	2	1	1	1
3	53	40	3	11.5	8.5	72.25
4	49	52	4.5	3	1.5	2.25
5	49	51	4.5	4.5	0	0
6	47	38	6	14	8	64
7	46	51	7	4.5	2.5	6.25
8	45	32	9	17	8	64
9	45	39	9	13	4	16
10	45	41	9	10	1	1
11	43	50	11	6	5	25
12	41	48	12	7.5	4.5	20.25
13	39	36	13	16	3	9
14	38	48	14	7.5	6.5	42.25
15	32	40	15.5	11.5	4	16
16	32	46	15.5	9	6.5	42.25
17	30	37	17	15	2	4
						ΣD^2 = 386.5

Now in this Table:

(i) First find the ranks of Personality scores. In column P these ranks have been calculated like 60 is the highest score which has been given rank 1. 54,53 being the next scores have been given ranks 2,3 respectively whereas, 42 is repeated two times. So on averaging 4 and 5, both of these are given the rank 4.5 each. Similarly, 32 is repeated two times and is given the rank of 15.5 each on averaging 15 and 16 ranks. So all the personality scores are given the ranks from 1 to 17.

(ii) Similarly, find the ranks of Mental Ability scores. In column Q, these ranks have been calculated like, score 68 is given rank one, 60 and 52 have been given ranks 2 and 3 respectively whereas, 51 appears two times and has been given the ranks of 4.5 each. Similarly, 48 appears two times and is given rank of 7.5 each being an average of 7 and 8. Similarly 40 is also repeated and is given the rank of 11.5 each, being an average of 11 and 12. Thus, scores are awarded the ranks from 1 to 17.

(iii) Now find the difference of Ranks of each of the score pairs of Personality and Mental ability in each row. Thus, the difference of ranks of each of these ranks is column R = Column Q – Column P without sign.

(iv) In next column find the squares of the difference of ranks, i.e. S column is the square of the Difference. Column S is (column R)^2. For example, in Table you have got D^2 = 1,1,72.25, 2.25 etc. for the first four students. Hence Column S = Column R^2.

(v) Apply the values in the formula to find rank order coefficient of correlation:

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

In this example

$\sum D^2$ = 386.50

N = 17

N^2 = 289

$$(\rho) = 1 - \frac{6 \times 386.5}{17\ (289 - 1)}$$

$$(\rho) = 1 - \frac{2319.0}{4896}$$

$$= 1 - 0.47 = 0.53$$

16.6.2 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.

$$r_{xy} = 1 - \frac{\sum xy}{nS_x S_y}$$

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

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

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

n = Total number of pairs of scores

S_x = Standard Deviation of first variable scores

S_y = Standard Deviation of second Variable Scores.

Example

Suppose you have the scores of ten students in Physics and Mathematics as follows and are interested in finding the coefficient of correlation between them.

Table 16.2 : Scores of students in Physics and Mathematics

Student	**Scores in Physics**	$X = X - \overline{X}$	X^2	**Scores in Maths**	$Y = Y - \overline{Y}$	Y^2
1	37	-2.1	4.41	75	-2.6	6.76
2	41	1.9	3.61	78	0.4	0.16
3	48	8.9	79.21	88	10.4	108.16
4	32	-7.1	50.41	80	2.4	5.76
5	36	-3.1	9.61	78	0.4	0.16
6	39	-0.1	0.01	71	-6.6	43.56
7	40	0.9	0.81	75	-2.6	6.76
8	45	5.9	34.81	83	5.4	29.16
9	39	-0.1	0.01	74	-3.6	12.96
10	34	-5.1	26.01	74	-3.6	12.96
	$\sum X=391$		$\sum X^2=208.9$	$\sum Y=776$		$\sum Y^2=226.4$

$$\overline{X} = \sum X / N = 39.1 \qquad \overline{Y} = \sum Y/N = 77.6$$

$$S_x = \sqrt{\frac{\Sigma X^2}{N}} = 4.57$$

$$S_y = \sqrt{\frac{\Sigma Y^2}{N}} = 4.76$$

In this Method,

i. Find means of both the scores i.e. scores of Physics and Mathematics as calculated above are 39.1 and 77.6

ii. Then find deviation of each score from means as (x) and (y).

iii. Then find deviation squares in each case as calculated below and sum them up. These values in this example are: $\sum X^2=208.9$ and $\sum X^2=226.4$

Table 16.3 : Table for Calculation of Product Moment Coefficient of correlation by Deviation Score Method

Student	Phy (x)	Maths (y)	x Deviation	y Deviation	xy	x^2	y^2
1	37	75	-2.1	–2.6	5.46	4.41	6.76
2	41	78	1.9	+0.4	0.76	3.61	0.16
3	48	88	8.9	+10.4	92.56	79.21	108.16
4	32	80	-7.1	+2.4	-17.04	50.41	5.76
5	36	78	-3.1	+0.4	-1.24	9.61	0.16
6	39	71	-0.1	–6.6	0.66	0.01	43.56
7	40	75	0.9	–2.6	-2.34	0.81	6.76
8	45	83	5.9	+5.4	31.86	34.81	29.16
9	39	74	-0.1	–3.6	0.36	0.01	12.96
10	34	74	-5.1	–3.6	18.36	26.01	12.96
	Σx = **391 M=39.1**	Σy = **776 M=77.6**		**Total**	**Σxy= 129.4**	**ΣX^2= 208.9**	**ΣY^2= 226.4**

iv. Then find standard deviation of both the group of scores i.e. scores of Physics and Scores of Mathematics as calculated above are 4.57 and 4.76. These are done after finding x^2 and y^2 and their sum.

v. Then find 'x × y' and their sum as done above, $\Sigma xy=129.4$.

vi. Then find coefficient of correlation.

$$r_{xy} = \frac{\Sigma xy}{nS_xS_y}$$

$$= 129.4 \ / \ (10\times4.57\times4.76)$$

$$= 129.4 \ / \ 217.53$$

$$= 0.59$$

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

In the raw score method if you substitute the values of deviations in raw score form and value of standard deviations in raw score form you will get the following formula for the calculation of coefficient of correlation from raw scores :

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

Let us take an example for 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

Table 16.4: Table for Computation of 'r' by Raw Score Method

S. No.	Score in language	Score in History (x)	X^2 (y)	Y^2	XY
1	20	12	400	144	240
2	18	16	324	256	288
3	16	10	256	100	160
4	15	14	225	196	210
5	14	12	196	144	168
6	12	10	144	100	120
7	12	9	144	81	108
8	10	2	100	4	20
9	8	7	64	49	56
10	5	8	25	64	40
	$\Sigma X=130$	$\Sigma Y=100$	$\Sigma X^2=1878$	$\Sigma Y^2=1138$	$\Sigma XY=1410$

(i) First find sum of all raw scores of language and History ΣX and ΣY Where, $\Sigma X = 130$ and $\Sigma Y = 100$ as above

(ii) Then find squares of the scores of language and History X^2 and Y^2 which are calculated in column no. 4 and 5.

(iii) Find sum of squares of raw scores of language and History as above

$\Sigma X^2 = 1878$ and $\Sigma Y^2 = 1138$

(iv) Then find multiple of raw scores of each pair (XY)

(v) Find sum of multiple of raw scores of each pair as above.

$\Sigma XY = 1410$

(vi) Now find Coefficient of Correlation by substituting each value.

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

$$r = \frac{10(1410) - (130)(100)}{\sqrt{\left[10(1878) - (130)^2\right]\left[10(1138) - (100)^2\right]}}$$

$$r = \frac{14100 - 13000}{\sqrt{(18780 - 16900)(11380 - 10000)}}$$

$$= \frac{1100}{\sqrt{(1880)(1380)}}$$

$= 1100 \;/\; 1610.7$

$= 0.68$

(c) Standard Score or Z – Score Method

In this method, we neither use raw scores nor their deviations from the mean. In stead we use the standard scores to find the coefficient of correlation. Thus in this case first we transform both the variables into their standard scores and then find coefficient of correlation by the following formula:

$$r = \frac{\Sigma\left(Z_x Z_y\right)}{N}$$

The following steps are employed in these calculations:

(i) Find Mean of both the scores sets (X and Y).

(ii) Find deviation of each score from their respective mean which will help you in finding standard deviation of both the groups of scores X and Y.

(iii) Now find Standard Deviation (S_x and S_y) of both the sets of scores.

(iv) Then transform both the group scores into their standard score (Z scores) by the formula which you know

$$Z_x = \frac{x}{S_x} \quad \text{and} \quad Z_y = \frac{y}{S_y}$$

(v) Then find multiple of both the Z scores (Z_x) (Z_y) and sum it up

(vi) Find correlation coefficient by substituting the above value in the formula

$$r = \frac{\Sigma\left(Z_x Z_y\right)}{N}$$

Let us illustrate it with the same example as in the previous case of scores of language and scores of history.

Example

Table 16.5 : Table for the calculation of Coefficient of correlation by Standard Score of Z – Score Method

S.No.	X	Y	X	Y	Z_x	Z_y	$Z_x Z_y$
1	20	12	7	2	1.61	0.54	0.8694
2	18	16	5	6	1.15	1.62	1.8637
3	16	10	3	0	0.69	0.00	0
4	15	14	2	4	0.46	1.08	0.4968
5	14	12	1	2	0.23	0.54	0.1242

6	12	10	–1	0	–0.23	0.00	0
7	12	9	–1	–1	–0.23	–0.27	0.0621
8	10	2	–3	–8	–0.69	–2.16	1.4904
9	8	7	–5	–3	–1.15	–0.81	0.9315
10	5	8	–8	–2	–1.84	–0.54	0.9936
	$\Sigma x=130$	$\Sigma y=100$	$S_x=4.34$				$\Sigma_{ZxZy}=6.8317$
	$\overline{X} = 13$	$\overline{Y} = 10$	$S_y=3.71$				

$$r = \frac{6.8317}{10}$$

$$= 0.68$$

The above calculation is similar to the one calculated by other method. It does not change for the same set of data. Thus we can say that the *coefficient of correlation calculated from deviation method, raw score method or the standard scores method will remain unchanged* because all these methods are derived from the common formula of deviation in product moment coefficient of correlation.

16.6.3 Pearson's Product Moment Co-efficient of Correlation (Scattergram)

Scattergram as you know is based on plotting the scores of two sets of paired data on a graph or diagram in a table. By doing so, it gives us a visual picture of the bi variate data formed by the pairs of related scores. For drawing scatter gram, the unit of each variable is decided on the basis of its range and accordingly, score is represented on a graph. The advantage of this method is that it gives an 'eye check' on the number of factors that may influence the value of coefficient, which you can use when you interpret your observations. It can also give you a rough check of your coefficient of correlation, if you have already calculated by any of above three techniques. For example, if you find that scatter plot or scatter diagram corresponds to some value r = +0.72, whereas your calculated value by other means comes out to be 0.42. Then you would be aware of the discrepancy and can begin a recheck of your calculations. In computer, it is very easy to plot scatter diagrams with the help of simple statistical softwares. You may be curious to know the construction of a scatter diagram. Let us understand the construction of a scatter diagram in the following simple steps:

(1) Make a table of your set of paired scores.

(2) Decide the class interval for each set of scores and make a frequency table for each set by making tallies. These class intervals may be between 10 and 20 but 10 to 15 are most convenient classes.

Now you have both the scores in class interval forms. It is not necessary to have exactly the same number of class interval for both the variables, nor is necessary to use the same class interval width in both variables. This selection of width and number of classes depends upon the kind of data or the scores therein.

(3) Next we tally the scores pair by pair so that two scores meet at a single point i.e. we enter our tally mark in the cell where these two class intervals meet at one single place. Then we take the next pair and again we mark a tally for the pair in the same way. Thus we make tally for each pair of scores.

(4) Inspect your own tally marks once again to see that all tallies are marked correctly.

(5) Now you add your tallies on either side of your table to get frequencies and total of each of these sums will be the same. This is at the bottom of the table for the sum of columns and on right hand side for the sum of rows frequencies. These are row A and column P in your table.

(6) Now, we write moments on both sides of these sum total columns and sum total rows. In these moments the highest is at the top of column and designated as y and is at the right side in a row which is designated as x. These are deviations from arbitrary reference points. So you get second column and row of the bottom as moments i.e. Y' and X'. These are row B and column Q of table

(7) In the next column and next row multiply each moment by their frequencies. So in next row you will get fx' and in next column on right you will get fy'. These are row C and column R of your table.

(8) Now in next column and next row multiply each of previous row and column with moment again. So in next row you will get $\Sigma fx'^2$ and in next column you will get fy'^2. These are row D and column S of your table.

(9) In next column you find sum of last two rows C and D i.e. find $\Sigma f\,x'$ and $\Sigma fx'^2$.

(10) Similarly, sum your column R and S i.e. find $\Sigma fx'$ and $\Sigma fx'^2$.

(11) Now take the product of the deviations from the two arbitrary reference points. Thus, it is product of moments as you have seen above that these deviations from arbitrary reference points are called moments. Here the arbitrary points are not the class interval but it is the mid point of that class interval. These products of moments are calculated by multiplying the y' with frequency of x' (total frequency).

(12) Now find the sum of these product moments i.e. x' y'.

(13) Find the coefficient of correlation by the following formula.

$$r = \frac{\Sigma x'y' - \left[(\Sigma fx')(\Sigma fy')/N\right]}{\sqrt{\left\{\Sigma fx'^2 - \left[(\Sigma fx')^2/N\right]\right\}\left\{\Sigma fy'^2 - \left[(\Sigma fy')^2/N\right]\right\}}}$$

Let us understand this scatter diagram with the help of an example.

Example

Suppose, you have scored 35 students of your class on the mental ability before teaching and examining them on their achievement after six months. You got the scores and you are interested in finding the correlation between their mental ability scores and the achievement scores with the help of scattergram. The scores of the students in Mental Ability Test and Achievement Test are as in table 16.6.

Table 16.6 : Scores of the students of a class on Mental Ability and Achievement

S.No. of the student	Mental Ability Score	Achievement Score
1.	80	61
2.	95	28
3.	94	74
4.	101	46
5.	105	44
6.	89	38
7.	106	72
8.	92	41
9.	105	49
10.	107	69
11.	111	82
12.	114	76
13.	83	39
14.	112	64
15.	91	77
16.	88	50
17.	105	55
18.	106	59
19.	105	86
20.	80	63
21.	85	31
22.	93	57
23.	85	70
24.	92	43
25.	90	70
26.	89	54
27.	85	51
28.	96	58
29.	85	63
30.	98	73
31.	101	71
32.	106	76
33.	112	76
34.	93	59
35	109	72

Table for Scattergram of Pearson's Product Moment Correlation

X-Axis – Mental Ability

Y-Axis – Achievement

		25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65-69	70-74	75-79	80-84	85-89	P	Q	R	S	T
															f	y'	$f\,y'$	$f\,y'^2$	$y'\,x'$
114-116												${}_{24}1^{24}$			1	6	6	36	24
111-113									${}_{5}1^{5}$			${}_{20}1^{20}$	${}_{25}1^{25}$		3	5	15	75	50
108-111											${}_{12}1^{12}$				1	4	4	16	12
105-107					${}_{-9}1^{-9}$	${}_{-6}1^{-6}$		${}_{0}2^{0}$		${}_{6}1^{6}$	${}_{9}1^{9}$	${}_{12}1^{12}$		${}_{18}1^{18}$	8	3	24	72	30
102-104															0	2	0	0	0
99-101						${}_{-2}1^{-2}$					${}_{3}1^{3}$				2	1	2	2	1
96-98								${}_{0}1^{0}$			${}_{0}1^{0}$				2	0	0	0	0
93-95		${}_{6}1^{6}$						${}_{0}2^{0}$			${}_{-3}1^{-3}$				4	–1	–4	4	3
90-92					${}_{12}2^{6}$						${}_{6}1^{6}$	${}_{-8}1^{-8}$			4	–2	–8	16	10
87-89				${}_{12}1^{12}$			${}_{6}2^{3}$								3	–3	–9	27	18
84-86			${}_{20}1^{20}$				${}_{4}1^{4}$		${}_{-4}1^{-4}$		${}_{-12}2^{-12}$				4	–4	–16	64	8
81-83				${}_{20}1^{20}$											1	–5	–5	25	20
78-80									${}_{-12}2^{-6}$						2	–6	–12	72	–12
A	f	1	1	2	3	2	3	5	4	1	7	4	1	1	N=35				
B	x'	–6	–5	–4	–3	–2	–1	0	1	2	3	4	5	6					
C	$f\,x'$	–6	–5	–8	–9	–4	–3	0	4	2	21	16	5	6	$\sum fx' = 33$				
D	$f\,x'^2$	36	25	32	27	8	3	0	4	4	63	64	25	36	$\sum x'^2 = 327$				
T	$x'y'$	6	20	32	3	–8	10	0	–11	6	15	48	25	18	$\sum x'y' = 164$				

Let us understand the calculations in this scattergram table through the steps we had studied.

1. Paired set of scores are there in the form of mental ability scores and achievement scores.

2. The classes of mental ability are formed from 78 to 116 with class intervals of 4 as shown on Y axis.

3. The classes of achievement scores are formed from 25 to 90 with class intervals of 5 each. They are shown on X-axis.

4. Now scores are tallied as per their value on both the axis as shown in between the Table. We enter first score pair (80, 61) in their respective column and row. Similarly next (95, 28) pair is placed as tally in the respective position and so on.

5. Recheck that they have marked tally in correct place.

6. Add the tally on either side in their respective columns and rows to get frequency as shown in row A and column P. These sums of all frequencies of A and P will be same i.e. total N. This will recheck the tallies.

7. Now mark moments as the deviation from the middle class interval. These are there in column Q and row B. The deviation of starting class is 'O' and above the class, deviations are +1, +2, +3 .... and below the class –1, –2, –3.... etc.

8. Now multiply each moment by their respective frequency on either side i.e. columns and rows to get f y' and f x' i.e. column R and row C in this table. It is product of

   P × Q = R and A × B = C

9. Now again multiply them by moment on each column and in each row to get S and D column and row respectively in this table. It is the product of R × Q = S column and

   B × C = D row. This is fx'^2.

10. Now sum all the rows and columns to get their sum (Σ) i.e. get $\Sigma f x'$, $\Sigma fx'^2$, $\Sigma y'$ and $\Sigma y'^2$.

11. Now find the product of the moments, i.e., the product of the x' and y' in T column.

12. Add the product moments which are there in column T i.e. you get $\Sigma x'y'$.

13. Now find the correlation 'r' by the following formula

$$r = \frac{\Sigma x'y' - \left[(\Sigma fx')(\Sigma fy')/N\right]}{\sqrt{\left\{\Sigma fx'^2 - \left[(\Sigma fx')^2/N\right]\right\}\left\{\Sigma fy'^2 - \left[(\Sigma fy')^2/N\right]\right\}}}$$

$$r = \frac{164 - (33 \times 6 / 35)}{\sqrt{327 - (33^2/35) \times 409 - 6^2/35)}}$$

$$= \frac{164 - 5.66}{\sqrt{(327 - 31.11) \times (409 - 1.03)}}$$

$$= \frac{158.34}{\sqrt{295.89 \times 407.97}}$$

$$= \frac{158.34}{347.44} = 0.46 \text{ (Approx)}$$

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. 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

   ....

   ....

   ....

6. Compute correlation of the marks of 10 students in Pearson's product moment deviation score method.

   Scores (Sc.) : 35, 37, 60, 40, 62. 55, 40, 43, 55, 50

   Scores (Soc. Sc.) : 33, 40, 61, 39, 52, 50, 36, 44, 50, 60

   ....

   ....

   ....

16.7 INTERPRETATION OF THE CO-EFFICIENT OF CORRELATION

Merely computation of correlation does not have any meaning until and unless we determine how large the coefficient need to be in order to be significant, and what does correlation tell us about the data? Coefficient of correlation is positive or negative depending upon the direction as explained earlier in positive and negative correlations. Thus coefficient of correlation will have values between +1 and –1. When the value of obtained correlation is closer to 1 or –1, then the variables are more closely related and if it is close to 0, it means that there is no relationship between these variables. The coefficient of correlation tells us the intensity of linear relationship in quantitative terms. In real life situations perfectly correlated variables are rare.

Full interpretation of correlation 'r' depends upon the circumstances, one of which is the size of the sample. All that can really be said is that when estimating the value of one variable from the value of other; the higher is the value of coefficient of correlation; the better will be the estimate. The closeness of relationship is not proportional to the value of "r". If the value of this "r" is 0.85, it does not necessarily indicate that a relationship is five times as close as one having "r" 0.17. We can only say that they have different kind of relationships. It will represent the proportion of common variation in two variables both in strength and magnitude.

Coefficients of correlation are not even additive. An average of correlation coefficients in a number of samples does not represent an average correlation in all those samples. Thus we should not add the coefficients of correlation to find correlation of two or more samples. But we should first convert them into additive measures and find correlations after that.

The coefficient of correlation is interpreted in verbal description. The rule of thumb for interpreting the size of a correlation coefficient is presented below:

Table 16.7: Correlation coefficient and its interpretation

Size of Correlation	Interpretation
±1	Perfect Positive/negative Correlation
± .90 to ± .99	Very High Positive/Negative Correlation
± .70 to ± .90	High Positive/Negative Correlation
± .50 to ± .70	Moderate Positive/Negative Correlation
± .30 to ± .50	Low Positive/Negative Correlation
± .10 to ± .30	Very low Positive/Negative Correlation
± .00 to ± .10	Markedly Low and Negligible Positive/Negative Correlation

16.8 MISINTERPRETATION OF THE COEFFICIENT OF CORRELATION

Sometimes, we misinterpret the value of coefficient of correlation and establish the cause and effect relationship, i.e. one variable causing the variation in the other variable. Actually we cannot interpret in this way unless we have sound logical base. Correlation coefficient gives us, a quantitative determination of the degree of relationship between two variables X and Y, not information as to the nature of association between the two variables.

- The degree of relationship or association is not ordinarily interpretable in direct proportions to the magnitude of the coefficient of correlations. For example, with 'r' = 0.25 we cannot say one variable has one fourth association with the other variable, or with "r" = 0.50 we can not interpret that one variable has half relationship with the other variable. This cannot be objectively considered. In general, change of 0.10 point in coefficient may have greater consequences when applied to coefficients having a high value than if the same is applied to lower values.

- The strength of relationship between two variables depends, among other things, on the nature of measurement of the two variables as well as on the kind of subjects being studied. Thus, it is not possible to speak of the correlation between two variables without taking these factors into consideration.

- Sometimes, you may find a high correlation between two variables, which may tempt you to think that they are substantially correlated. But in actual practice they may not have any relationship. You have to logically see whether such an association could exist or not. There must be some common cause for the association between the two. Mere high coefficient is insufficient to claim a relation between the two variables.
- Although a straight line graph between two variables is an indication of a straight correlation or high correlation between the two variables, but sometimes, it may not be so and in such cases it may be mere chance or it is misleading.
- The correlation coefficient is affected by the range of talent (variability) characterizing the measurements of the two variables. In general, the smaller is the range of talent in two variables; the lower will be coefficient of correlation, other things being equal. For example, in any school the correlation between the aptitude test score and marks obtained by the students may be +0.50. But if you find the coefficient of correlation of the students having high scores, the coefficient may be much lower.
- Coefficient of correlation like any other statistics is also dependent upon sampling fluctuations. It is also affected by sampling variations. Depending upon the characteristics of a particular sample, the obtained coefficient may be higher or lower than it will be in a different sample.
- Correlation works with quantifiable data in which the number is meaningful as we have already explained above. It cannot be used for purely categorical data such as gender, colours etc.
- Squaring coefficient of correlation makes it easier to understand and gives the percent of the variation in one variable that is related to the other variable and makes it easier to understand. If r = 0.5; then square becomes 0.25 and we can say there is 25% variation in one variable with respect to the other variable under study. This squared correlation coefficient (r^2) is the proportion of variance in y variable that can be accounted for by knowing X variable. One important property of variance is that it may be portioned into separate additive parts. Like total correlation of a sample variable X with variable Y where X is total human beings. So X if divided into Male and Female, the total "r" can also be divided into two parts r_1 and r_2 where $r = r_1 + r_2$.

16.9 FACTORS INFLUENCING THE SIZE OF THE CORRELATION COEFFICIENT

It will be helpful for you to know that the following factors influence the size of the coefficient of correlation and can some times lead to misinterpretation:

1. The size of "r" is very much dependent upon the variability of measured values in the correlated sample. The greater the variability, the higher will be the correlation, everything else being equal.
2. The size of "r" is altered, when an investigator selects an extreme group of subjects in order to compare these groups with respect to certain

behaviour. "r" obtained from the combined data of extreme groups would be larger than the "r" obtained from a random sample of the same group.

3. Addition or dropping the extreme cases from the group can lead to change in the size of "r". Addition of the extreme case may increase the size of correlation, while dropping the extreme cases will lower the value or "r".

16.10 IMPORTANCE AND USE OF CORRELATION IN EDUCATIONAL MEASUREMENT AND EVALUATION

Correlation is important in many areas of measurement and evaluation on education. It is of particular importance in the study of similarities and individual differences. With this common property of correlation in mind, correlation is used in the following contexts:

(i) For Determining Reliability:

Suppose there are two examiners examining the same class on some parameters or in some examination. The comparison of two Examiners or two observations of behaviour by the same examiner can be judged with the correlation. Correlation tells how reliable the inter observer rating of behaviour is. It tells us whether the two observers agree on the scores given or whether there is agreement between the two obtained values or scores.

It is not only the inter observer score but even the intra observations by the same examiner for two or more times can be easily compared with the help of correlation.

(ii) For Determining Validity of Scores in Predictions:

Like reliabilities, the correlation is useful in determining validity of two or more sets of scores. A high correlation is an indicator of high validity. For example, when you are interested in knowing to what extent the scores of students on aptitude test are related to their achievement or performances. If there is a high correlation between the two variables, then you can rely on any one of the tests which can be a valid measure of the variable.

(iii) Study of Individual Differences:

These relationships and the differences can be studied through the correlations of negative and positive kind. The concept of correlation is basic to the theory and practice of trait measurement. For example, the use of different types of tests like mental ability tests, aptitude tests, tests of reasoning, mathematical ability test, etc. are all used based on their property of correlation with achievement or performance, etc.

(iv) For predicting one or More Variables:

You might have estimated academic performance of the children of your school on the basis of your knowledge of their aptitude or some other test score. Even sometimes you might have predicted the scores of the students in their Board examination on the basis of their scores in previous examination or Pre-Board examination. If some one asks you the basis of your prediction, you may answer the query on the basis of your knowledge of correlation. Correlation is helpful for predicting or making estimation of one variable from our knowledge of the other variable.

(v) ***For Determining the Usefulness of Regression Line:***

Coefficient of correlation is very useful and the most frequently used measure for determining the usefulness of a regression line for the estimation purposes. It helps in proper estimations and the partial correlation of different variables that affect a particular variable.

(vi) ***Determining Prediction Validity of Tests & Measurement:***

Predictive validity is itself a correlation between a set of test scores or some other predictor with an external measure. This external measure is called the criterion. For example, to validate the intelligence tests you obtain a set of scores on a group of your school students and later find out the grade point averages (GPA) that these students get during their next examination. A correlation is then run between our two set of scores or measurements.

(vii) ***Qualifying the Relationship Between Variables:***

As correlation qualifies the relationship between the two variables, it also tells us whether the two variables are related to each other or not. It tells us about the kind of relationship, i.e., what kind of relationship exists between two variables.

(viii) ***Quantification of Degree of Relationship:***

The correlation not only qualifies for the relationship but also tells us about the quantity of relationship. The coefficient of correlation tells us the quantity of relationship. Square of coefficient of correlation (r^2) is the coefficient of determination which may even help in determining one variable with the knowledge of the other related variable. Hence, coefficient of correlation helps us in quantification of the degree of relationship between two related variables.

(ix) ***Finding Error in Prediction:***

Coefficient of correlation also gives a way of specifying the error involved in prediction of one variable from the other. The square of coefficient of correlation expresses the amount of variance commonly shared by the two variables, offering a means whereby the variability of human behaviour and performance can be explained.

16.11 LET US SUM UP

Measure of relationship of paired variables is quantified by a coefficient of correlation. The correlation coefficient is a value between – 1.0 to + 1.0. It should be noted that a high coefficient does not imply a cause-and-effect relationship, but merely quantifies a relationship that has been logically established.

The size of the correlation coefficient is affected by the homogeneity of the scores on the variables. If a relationship exists between two variables and that relationship is linear, then the scores are more heterogeneous and the greater the range of measurement, the greater is the absolute value of correlation.

The correlation coefficient is very useful in educational evaluation, standardizing tests and in making predictions.

16.12 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.

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.), Lowa State University Press, Ames, Lowa.

16.13 ANSWERS TO CHECK YOUR PROGRESS

1. Correlation is the association or relationship between 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 (a, b, and c)
5. Self exercise
6. Self exercise

UNIT 17 NATURE OF DISTRIBUTION AND ITS INTERPRETATION

Structure

17.1 INTRODUCTION

By now you are well aware of the distribution of scores. In the previous units, you have also learnt to illustrate the shape of a frequency distribution through different kinds of graphs like histogram and frequency polygons. In histogram or frequency polygon all the scores of a distribution have some place or space. The central tendency helps you describe the central value of this distribution and the measure of variability helps you indicate the extent of variation in the distribution. These distributions or graphs help you in getting information about the set of scores or position of scores in the set. These graphs are varied from distribution to distribution.

But many a times, we need to categorise the group of individuals on certain measure or test or the trait. We may like to know the procedures for describing individual's position within a group or a set of scores on that trait. It may be distribution of the students on the basis of ability, achievement, and intelligence through a set of frequency distribution.

For Example: Suppose you had administered an achievement test in your subject to ascertain the level of achievement of your students and a student had got some marks (score) on that test. What does this score mean? Any obtained score has a meaning only with respect to other scores of that distribution or the scores of some other distribution on similar criteria. If you see any distribution that is occurring or happening without any effect from outside, it will have a definite trend or shape in graphical term. If we take any variable or any event of the nature whether educational, psychological or otherwise, it is occurring on 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. This curve helps us understand the meaning of any score with respect to other scores or the position of an individual or even helps us in doing categorisation of a group.

The present unit presents the concept and use of Normal Distribution in relation to the educational variables through suitable illustrations and explanations.

17.2 OBJECTIVES

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

- explain the concept of Normal Distribution and Normal Probability Curve;
- recall the theoretical base of the Normal Probability Curve;
- explain the need of a Normal Distribution;
- write the properties of Normal Probability Curve;
- draw a curve of any kind of distribution;
- recognize the various divergence in the Normal Curve;
- recall the definitions of various divergence of Normal Probability Curve;
- justify the significance of Skewness and Kurtosis in the educational measurement and evaluation;
- interpret the Normal Curve obtained on the basis of large number of observations;
- discuss the importance of Normal Curve in educational measurement;
- recall the various applications of Normal Curve in educational measurement and evaluation;
- find the number of cases in any one sector of a normal distribution; and
- apply the knowledge of Normal Probability Curve in solving various practical problems related to educational measurement and evaluation.

17.3 NORMAL DISTRIBUTION/NORMAL PROBABILITY CURVE

A normal distribution is also called as a normal probability curve. Let us try to understand first the concept of a normal distribution.

17.3.1 The Concept of Normal Distribution

Although it is not possible to predict the exact nature of occurrence of any distribution, still mathematics and statistics, with the help of well established theories, have applied certain methods of calculating the numerical values and predict the kind of distribution that is expected in general. This is made possible with the help of theory of probability.

If we measure any variable for a large number of times, we will find that they observe a definite pattern or symmetry. On plotting a graph of these measurements we will find the distribution to be symmetrical. Such distributions which are perfectly symmetrical and follow the laws of nature are called Normal Distributions i.e. in which the left half is a mirror image of the right half, as well as, mean, median and (if the distribution is uni-modal) the mode have the same value. Such normal distributions are widely occurring in nature. However, equality of mean and median does not guarantee that the distribution is symmetrical, although it is not likely to depart very far from that condition. On the other hand, if mean and median are different or have different values, the distribution cannot be normal.

Normal distribution is important because many of the educational and psychological variables are distributed close to normality. Measures of intelligence, reading ability, introversion and extroversion, job satisfaction, other personality traits or memory are among some of the psychological variables distributed approximately normally. Although these distributions are only approximately normal, they are quite close to normality. Many kinds of statistical tests can be derived for normal distributions. These tests work very well even if the distribution is approximately normally distributed. If the normality of the distribution is known, it is easy to convert back and forth from raw scores to the distribution and distribution to the raw scores.

Carefully look at the following hypothetical frequency distribution, which a teacher has obtained after examining 150 students of class X on a mathematics achievement test (out of 150) (see Table 17.1).

Table 17.1: Frequency Distribution of the Mathematics Achievement Test Scores

Class Intervals	Tallies	Frequency
115–119	I	1
110–114	II	2
105–109	IIII	4
100–104	~~IIII~~ II	7
95–99	~~IIII~~ ~~IIII~~	10
90–94	~~IIII~~ ~~IIII~~ ~~IIII~~ I	16
85–89	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	20
80–84	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	30
75–79	~~IIII~~ ~~IIII~~ ~~IIII~~ ~~IIII~~	20
70–74	~~IIII~~ ~~IIII~~ ~~IIII~~ I	16

65–69	~~IIII~~ ~~IIII~~	10
60–64	~~IIII~~ II	7
55–59	IIII	4
50–54	II	2
45–49	I	1
	Total	**150**

Can you see some special trend in the frequencies shown in the column 3 of the above table? Probably Yes! The concentration of maximum frequency (*f*=30) is at the central value of distribution and frequencies gradually taper off symmetrically on both the sides of this value.

If we draw a frequency polygon with the help of the above distribution, we will have a curve as shown in the Fig. 17.1

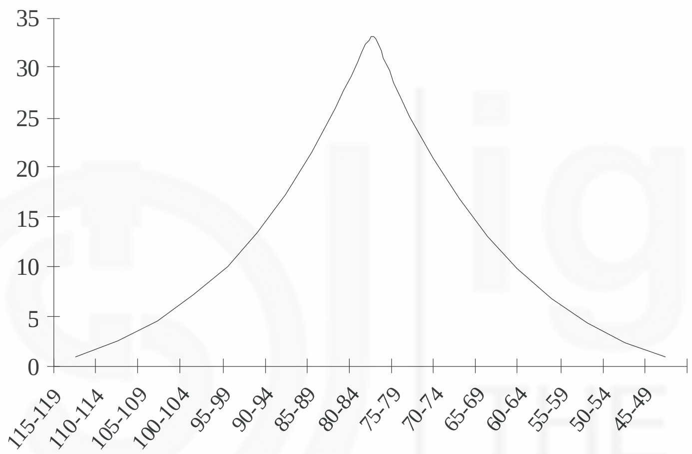


Figure 17.1: Frequency Polygon of the data given in Table 17.1

The shape of the curve in Fig. 17.1 is just 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 are approximately the same (M = Md = Mo = 82).

This 'Bell' shaped curve is technically known as Normal Probability Curve or simply Normal Curve and the corresponding frequency distribution of scores, having equal values of all three measures of central tendency, is known as Normal Distribution. This normal curve has great significance in cognitive and educational measurement. In measurement of behavioural aspects, the normal probability curve has often been used as reference curve.

17.3.2 The Normal Probability Curve: Its Theoretical Base

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. He developed its mathematical equation and graphical representation also. You must have studied Binomial Distribution in mathematics in your earlier classes at the school or college level. Normal distribution is a limiting form of the Binomial distribution in which neither of the two variables (chances) p and q is very small and number of trials is very large to make it random.

17.3.3 Properties of Normal Probability Curve

By now you must have understood the meaning of normal distribution from the above discussion that it is a peculiar distribution and has its own specific properties, which make it so important. It is more important for you to understand its important properties as they will be useful for you in times to come. In this section, we will be discussing these important properties or characteristics which make it so distinct. Some of the properties of a normal distribution are:

1. ***The Normal Curve is Symmetrical:*** The normal distribution curve 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. It is not only equal number of cases but even their distribution on either side i.e., below and above the mean value will be equal. The curve will be symmetrical around the vertical axis. If we draw a curve for a number of trials or various cases of a distribution means as the vertical axis, it will look alike on both sides of the axis.

It implies that the size, shape and slope of the curve on one side of the curve is identical to that of the other. In other words the left and right values to the middle central point are mirror images, as shown in Figure 17.2 below.

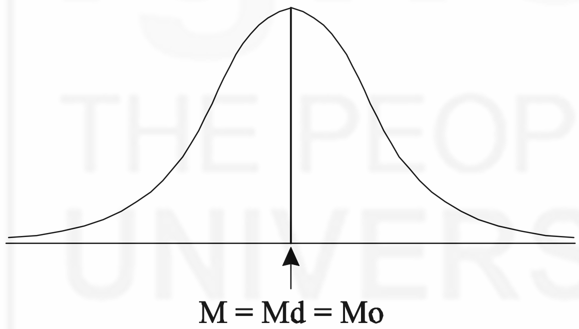


Figure 17.2: Normal Distribution Curve

2. When we draw a normal probability curve for a distribution, it is always a bell shaped curve irrespective of the nature of the variable for which it is being drawn. We may take any variable and see its occurrence in large number of cases; we will get the same kind of distribution. Hence the curve drawn for any variable will be a normal distribution curve or a bell shaped curve. e.g. if we take a coin which is symmetrical or unbiased and toss it a large number of times and we represent these observations on a graph; it shows a bell shaped curve. Similarly, if we throw an unbiased dice showing a three – sport and get a large number of observations. On plotting these scatter diagram points on a graph, we will get the same bell shaped curve because the distribution in this case also tends to be normal.

   Such a bell shaped curve is observed whenever we plot any normal distribution graphically for any variable. Even by seeing a graph of a distribution as bell shaped it could easily be said that the distribution is a normal distribution.

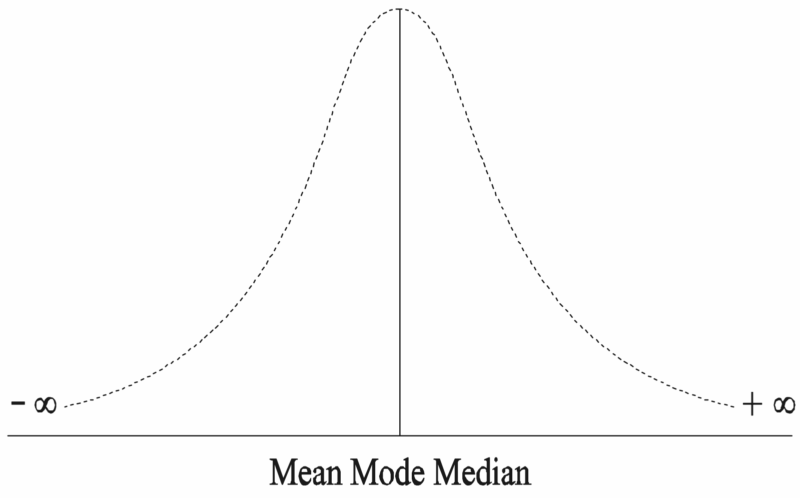


Figure 17.3: Normal Distribution Curve

3. Size, shape, and slope of the curve on one side of the curve is identical to the size, shape or the slope on the other side of the curve. This can be seen in the above said figure or any other diagram of a normal distribution curve.

4. The values of the mean, median and mode (in uni-modal distributions) computed for a normal distribution or those following the normal distribution curve always coincide at the same point and have the same value. The above diagram has shows it clearly.

   Mean = Mode = Median

5. The height of the normal curve is maximum at the mid point or the mean value. We can say that the height of the vertical axis drawn from the peak (called ordinate) is maximum at mean and in the unit normal curve, its value is equal to 0.3989.

6. 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

   $Q_3 -$ Median = Median $- Q_1$

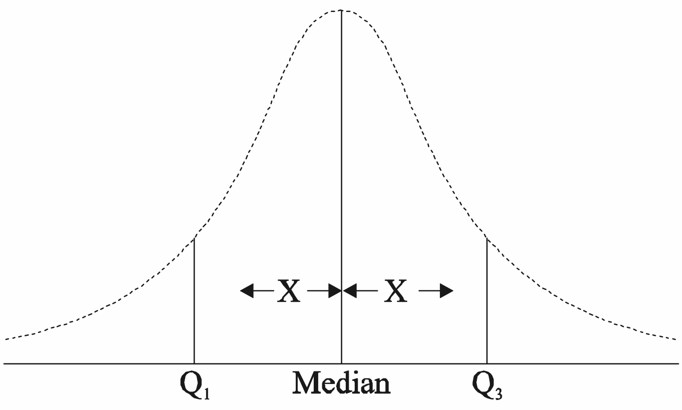


Figure 17.4: Quartiles showing in Normal Distribution Curve

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

7. 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 base line, no matter how far it is extended. Such a curve which is never ending is called *symptotic* curve and it extends from minus infinity to plus infinity. Normal distribution is continuous for all values of x between plus infinity to minus infinity so that each conceivable interval of real numbers has a probability other than zero. You can see the above diagram for your understanding. The curve is not touching the base line at the ends on left side or the right side. It continues to approach the base line but never touches it. Another point to note in these diagrams is that even no portion of the curve lies below the base line in a normal distribution curve.

8. Since there is only one point where the frequency is maximum in a normal distribution (i.e., Mean), so the normal distribution will always be a uni – modal distribution. Normal distribution has only one mode.

9. The mean deviation in a normal distribution, if calculated, is four fifth of its Standard deviation

   The mean deviation = 4/5 standard deviation = 4/5 σ

   If Standard deviation of a normal distribution is 5 then its mean deviation will be 4. However the distribution has to be perfectly normal.

10. Semi – inter quartile range in a normal distribution also has a fixed value which is also called the probable error and its value is 0.6745 of the standard deviation.

    Semi – inter quartile range = Probable Error = 0.6745 standard deviation = 0.6745 σ

11. If we want to find the points in the same normal curve at which the curve changes its path or direction, we will observe that these points are also symmetrical. They are at a distance of one standard deviation from the mean on either side. These points are called points of inflexion.

12. The various ordinates at different points or different distances (distance in terms of standard deviation) from the mean ordinate in a normal distribution stand in a fixed proportion to the height of the mean ordinate. Thus the height of the ordinate at one standard deviation distance on either side of the mean ordinate is 60.653% of the height of the mean ordinate and so on.

13. The data cluster around the mean: 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. Between mean and one Standard deviation or

       Between Mean to 1 S.D. = 34.13%

       Mean to 1σ = 34.13%

Mean to – 1 S.D. = 34.13%

Mean to – 1 σ = 34.13%

+ 1 σ to – 1 σ = 68.26%

ii. Between one Standard deviation to two standard deviation or

Between 1 S.D. to 2 S.D. = 13.59%

1 σ to 2 σ = 13.59%

Between – 1 S.D. to –2 S.D. = 13.59%

iii. Between mean and two Standard deviation or

Between Mean to 2 S.D. = 47.72%

Mean to 2 σ = 47.72%

Mean to –2 S.D. = 47.72%

Mean to –2 σ = 47.72%

+ 2 σ to –2 σ = 95.44%

iv. Between mean and 1.96 Standard Deviation or

Mean to 1.96 S.D. = 47.5%

Mean to 1.96 σ = 47.5%

Mean to –1.96 S.D.= 47.5%

Mean to –1.96 σ = 47.5%

+1.96 σ to –1.96 σ = 95%

v. Between mean and 2.58 Standard Deviation or

Mean to 2.58 S.D. = 49.5%

Mean to 2.58 σ = 49.5%

Mean to – 2.58 S.D. = 49.5%

Mean to – 2.58 σ = 49.5%

+ 2.58σ to –2.58σ = 99%

vi. Between mean and three Standard deviation or

Mean to 3 S.D. = 49.86%

Mean to 3 σ = 49.86%

Mean to –3 S.D. = 49.86%

Mean to –3 σ = 49.86%

+ 3 σ to –3 σ = 99.72%

v. Between mean and 2.58 Standard deviation or

Mean to 2.58 S.D. = 49.5%

Mean to 2.58 σ = 49.5%

Mean to –2.58 S.D.= 49.5%

Mean to –2.58σ = 49.5%

+ 2.58 σ to –2.58 σ = 99%

vi. Between mean and three Standard deviation or

Mean to 3 S.D. = 49.86%

Mean to 3 σ = 49.86%

Mean to –3 S.D.= 49.86%

Mean to –3 σ = 49.86%

+ 3 σ to –3 σ = 99.72%

vii. Between two and three Standard deviation or

2 S.D. to 3 S.D. = 2.15%

2 σ to 3 σ = 2.15%

–2 S.D. to – 3 S.D.= 2.15%

–2 σ to –3 σ = 2.15%

All these figures are approximated to two decimal places.

Let us try to understand this analysis with the help of normal distribution curve. In the following curve we have shown the areas under different conditions.

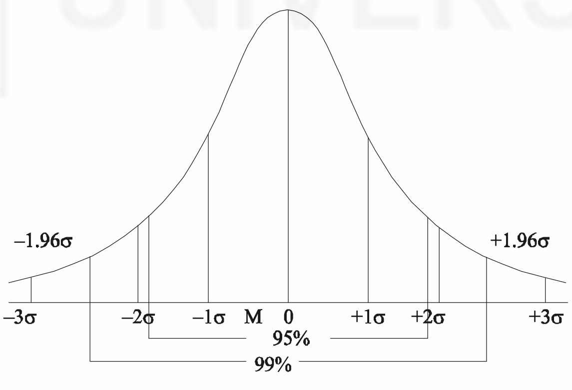


Figure 17.5 : Areas under Normal Probability Curve

The percentage of cases in a given normal curve are greatest around the centre i.e. around 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 ±1.96 standard deviation, they come out to be 95% and between ±2.58 standard deviation are 99% of the cases. These observations are very important in applying the significance limits.

14. The total area under the normal curve in terms of probability is taken as 1 (or 100%).

15. The probability that a normal random variable X equals particular value is 0

16. It links frequency distribution to probability distribution.

17. Normal distribution has the same shape as Standard Normal Distribution.

18. Normal distribution is actually a family of distributions and has two important parameters as mean and standard deviation. These two parameters determine the shape of the distribution. If we have different means and different standard deviations then yield will be different density curves. We will have different normal distributions in this case. All these distributions have the same general shape. They may differ in their spread but they bear similar bell shaped curve.

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 a Normal Probability Curve.

....

....

....

2. Mention five properties of Normal Curve.

....

....

....

3. Why is Normal Distribution so important in education?

....

....

....

4. In a normal distribution, what percentage of frequencies is :

(a) between -1σ to $+1\sigma$

(b) between -2σ to $+2\sigma$

(c) between -3σ to $+3\sigma$

....

....

....

5. Practically, why are the two ends of normal curve considered closed at the points ± 3 σ of the base line?

   ....

   ....

   ....

6. Fill in the blanks

   i. The shape of a normal distribution curve is …………….

   ii. In a standard normal distribution, mean = ……………. and standard deviation = …………….

   iii. Most of the statistical tests are ……………. distributed.

   iv. Percentage of cases between Mean and 1 σ = …………….

   v. Percentage of cases between Mean and 2 σ = …………….

   vi. 95% cases lie between Mean ± ……………. σ

   vii. 99% cases lie between Mean ± ……………. σ

17.3.4 Interpretation of Normal Probability Curve

Normal Curve has great significance in the cognitive measurement and educational evaluation. It gives important information about the trait being measured. If the frequency polygon of observations of measurements of a certain trait is a normal curve, it indicates that:

i. The measured trait is normally distributed in the Universe;

ii. Most of the cases are average in the measured trait and their percentage in the total population is about 68.26%;

iii. Approximately 15.87% of (50–34.13%) cases are high in the trait measured;

iv. Similarly, 15.87% cases approximately are low in the trait measured;

v. The test which is used to measure the trait is good;

vi. The test has good discrimination power as it differentiates between poor, average and high ability group individuals; and

vii. The items of the test used are fairly distributed in term of difficulty level.

17.3.5 Importance of Normal Probability Curve

The Normal Distribution is by far the most used distribution for drawing inferences from statistical data because of the following reasons:

1. Number of evidences are accumulated to show that normal distribution provides a good fit or describe the frequencies of occurrence of many variable and facts in (i) *biological statistics* e.g. sex ratio in births in a country over a number of years, (ii) *the anthropometrical data* e.g. height, weight, (iii*) wages and output of large numbers of workers* in the same occupation under comparable conditions, (iv) *psychological measurements* e.g. intelligence, reaction time, adjustment, anxiety and (v) *errors of observations* in Physics, Chemistry and other Physical Sciences.

2. The Normal distribution is of great value in educational evaluation and educational research, when we make use of cognitive measurements. It may be noted that normal distribution is not an actual distribution of scores on any test of ability or academic achievement, but is instead, a mathematical model. The distribution of test scores approach the theoretical normal distribution as a limit, but the fit is rarely ideal and perfect.

17.3.6 Applications of Normal Probability Curve

Some of applications of normal curve in the field of educational measurement and evaluation are:

(i) to determine the percentage of cases (in a normal distribution) within given limits or scores;

(ii) to determine the percentage of cases that are above or below a given score or reference point;

(iii) to determine the limits of scores which include a given percentage of cases;

(iv) hypothesis testing;

(v) determining standard error of mean;

(vi) to determine the percentile rank of a student in his own group;

(vii) to find out the percentile value of a student's percentile rank;

(viii) to compare the two distributions in terms of overlapping;

(ix) to determine the relative difficulty or test items; and

(x) dividing a group into sub-groups according to certain ability and assigning the grades.

17.3.7 Table of Areas Under the Normal Probability Curve

In order to use above applications of normal curve in educational measurement and evaluation, we need to know the Table of areas under the normal curve.

The Table 17.6 gives the fractional parts of the total area under the normal curve found between the mean and ordinates erected at various σ (sigma) distances from the mean.

The normal probability curve table is generally limited to the areas under unit normal curve with N = 1, σ = 1. When the values of N and σ are different from these, the measurements or scores should be converted into sigma scores (also referred to as standard scores or Z scores). The process is as follows:

$$Z=\frac{X-M}{\sigma} \quad = \quad Z=\frac{x}{\sigma}$$

Where:

Z = Standard Score

X = Raw Score

M = Mean of X Scores

σ = Standard Deviation of X Scores

The table of areas of normal probability curve are then referred to find out the proportion of area between the mean and the z value.

Though the total area under the N.P.C. is 1, but for convenience, the total area under the curve is taken to be 10,000 because of the greater ease with which fractional parts of the total area may then be calculated.

The first column of the Table 17.6, I.E., x/σ gives distance in tenths ofσ measured off on the base line for the normal curve from the mean as origin. In the row, the x/ σ distance is given to the second place of the decimal.

To find the number of cases in the normal distribution between the mean, and the ordinate erected at a distance of 1 ? unit from the mean, we go down the x/ σcolumn until 1.0 is reached and in the next column under .00 we take the entry opposite 1.0, namely 3413. This figure means that 3413 cases in 10,000; or 34.13 per cent of the entire area of the curve lie between the mean and 1σ. Similarly, if we have to find the percentage of the distribution between the mean and 1.56 σ, say, we go down the x/σ column to, then across horizontally to the column headed by .06, and note the entry 44.06. This is the percentage of the total area that lies between the mean and 1.56σ.

Table 17.2 : *Fractional parts of the total area (taken as 10,000) under the normal probability curve, corresponding to distances on the baseline between the mean and successive points laid off from the mean in units of standard deviation*

Example : Between the mean and a point 1.38σ (x/ σ = 1.38) are found 41.62% of the entire area under the curve:

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0.7	2580	2611	2642	2673	2704	2734	276	2794	2823	2852
0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3708	3729	2749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015

1.3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4383	4394	4406	4418	4429	4441
1.6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2.9	4981	4982	4982	4988	4984	4984	4985	4985	4986	4986
3.0	4986.5	4986.9	4987.4	4987.8	4988.2	4988.6	4988.9	4989.3	4989.7	4990.0
3.1	4990.3	4990.6	4991.0	4991.3	4991.6	4991.8	4992.1	4992.4	4992.6	4992.9
3.2		4993.129								
3.3		4995.166								
3.4		4996.631								
3.5		4997.674								
3.6		4998.409								
3.7		4998.922								
3.8		4999.277								
3.9		4999.519								
4.0		4999.683								
4.5		4999.966								
5.0		4999.997133								

We have so far considered only 'σ' distances measured in the positive direction from the mean. For this we have taken into account only the right half of the normal curve. Since the curve is symmetrical about the mean, the entries in Table 17.2 apply to distances measured in the negative direction (to the left) as well as to those measured in the positive direction. If we have to find the percentage of the distribution between mean and -1.28σ, for instance, we take entry 3997 in the column .08, opposite 1.2 in the x/σ column. This entry means that 39.97 percent of the cases in the normal distribution fall between the mean and -1.28σ.

For practical purposes we take the curve to end at points -3σ and $+3\sigma$ distance from the mean as the normal curve does not actually meet the base line. Table of area under normal probability curve shows that 4986.5 cases lie between mean and ordinate at $+3\sigma$. Thus 99.73 percent of the entire distribution, would lie within the limits -3σ and $+3\sigma$. The rest 0.27 percent of the distribution beyond $\pm3\sigma$ is considered too small or negligible except where N is very large.

17.3.8 Points to be Kept in Mind while Consulting Table of Area under Normal Probability Curve

The following points are to be kept in mind to avoid errors, while consulting the NPC. Table.

1. Every given score or observation must be converted into standard measure i.e. Z score, by using the following formula:

$$Z = \frac{X - M}{\sigma}$$

2. The mean of the curve is always the reference point, and all the values of areas are given in terms of distances from mean which is zero

3. The area in terms of proportion can be converted into percentage, and

4. While consulting the table, absolute values of z should be taken. However, a negative value of Z shows that the scores and the area lie below the mean and this fact should be kept in mind while doing further calculation on the area. A positive value of z shows that the score lies above the mean i.e. right side.

17.3.9 Problems Related to Application of the Normal Probability Curve

(a) To determine the percentage of cases in a Normal Distribution within given limits or scores

Example 1

In a class mean score in an Urdu test is 26.1 and the standard deviation is 6.45. How many cases lie between mean and score 35. Now let us find Z score

We know Z score = $\dfrac{\text{Mean deviation}}{\text{Standard deviation}}$

$$Z = \frac{\text{Score} - \text{Mean}}{\sigma}$$

$$Z = \frac{X - \overline{X}}{\sigma}$$

In this $X = 35$ $\overline{X} = 26.1$ and $\sigma = 6.45$

Thus,

$$Z = \frac{35 - 26.1}{6.45}$$

$$\frac{8}{6.45}$$

$$= 1.38$$

Now from the table of standard score, we can find area. Since 35 is 1.38 standard deviation on right side of the curve, the area between mean and 1.38 σ from table comes to be 41.62%. Hence 41.62% students get marks between mean and 35 marks.

Example 2

If you want to find the percentage of students getting marks between Mean and 20 then

$$Z = \frac{20 - 26.1}{6.45}$$

$$= \frac{-6.1}{6.45}$$

$$= -\ 0.946$$

Thus score 20 in a normal probability curve will lie 0.946σ on left side of mean as it is negative .The percentage of cases between mean and 0.946 σ from table is 32.76%. Hence 32.76% students score marks between 20 and 26.1.

Example 3

If we want to know the students getting marks between 20 and 35 then we can add the two percentages

Percentage of students scoring marks between 20–35 = 41.62 + 32.76 = 74.38%.

(b) To determine the percentile rank of a student in his own group

The percentile rank is defined as the percentage of scores below a given score.

Example 4

The raw score of a student of class X on an achievement test is 60. The mean of the whole class is 50 with standard deviation 5. Find the percentile rank of the student.

$$Z = \frac{X - M}{\sigma}$$

$$Z = \frac{60-50}{5} = \frac{10}{5}$$

$$Z = +2.00\ \sigma$$

According to the table of area under N.P.C., the area of the curve that lie between M and +2 σ is 47.72%.

The total percentage of cases below the score 60 is 50 + 47.22 = 97.72% or 98%

Thus, the percentile rank of a student who secured 60 marks in an achievement test in the class is 98.

(c) To determine the percentile value(score) of a student whose percentile rank is known

Example 5

In a class Rohit's percentile rank in the mathematics class is 75. The mean of the class in mathematics is 60 with standard deviation 10. Find out Rohit's marks in Mathematics achievement test. According to definition of percentile rank the position of Rohit on the N.P.C. scale is 25% scores above the Mean.

According to the N.P.C. Table the σ score of 25% cases from the Mean is +.67 σ.

Using the formula

$$Z = \frac{X-M}{\sigma}$$

$$\text{or} \quad +.67 = \frac{X-60}{10}$$

$$\text{or} \quad X - 60 = 10 \times .67$$

$$\text{or} \quad X = 60 + 6.7$$

$$\text{or} \quad X = 66.7 \text{ (Say 67)}$$

Rohit's marks in mathematics are 67.

(d) Dividing a group into sub-groups according to the level of ability

Example 6

Given a group of 500 college students who have been administered a general mental ability test. The teacher wishes to classify the group in five categories and assign them grades A, B, C, D, E according to ability. Assuming that general mental ability is normally distributed in the population; calculate the number of students that can be placed in groups A, B, C, D and E.

We know that total area of the Normal Curve extends from – 3σ to +3 σ that is over a range of 6 σ.

Dividing this range by 5, we get the σ distance of each category = 6 σ /5 = 1.2 σ. Thus, each category is spread over a distance of 1.2σ. The category

C will lie in the middle. Half of its area will be below the mean, and the other half above the mean.

The σ distance of each category is shown in the figure.

According to N.P.C. table the total percentage of cases from mean to .6 σ is 22.57.

The total cases in between –.6 σ to +.6 σ are 22.57 + 22.57 = 45.14%

Hence, in category C, the total percentage of students is = 45.14.

Similarly, according to N.P.C. Table the total percentage of cases from Mean to 1.8 σ is 46.41.

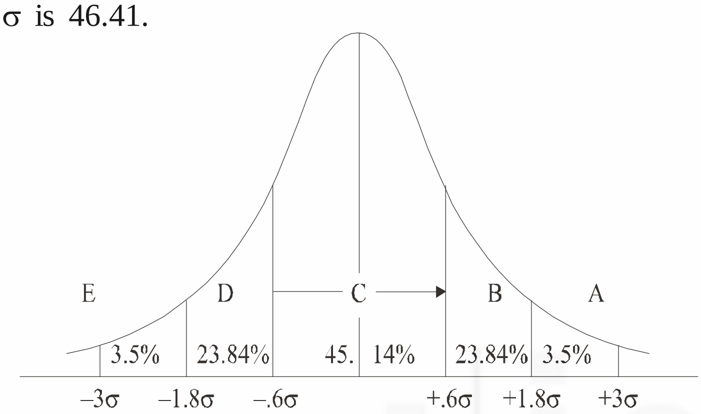


Figure 17.6: Percentages of Areas under Normal Probability Curve

The total percentage of cases in category B is 46.41 – 22.57 = 23.84%

In category A the total percentage of the cases will be 50 – 46.41 = 3.59%

Similarly, in category D and E the total percentages of the students will be 23.84% and 3.59% respectively. Thus :

In category A the total percentage of students is 3.59%

In category B the total percentage of students is 23.84%

In category C the total percentage of students is 45.14%

In category D the total percentage of students is 23.84%

In category E the total percentage of student is 3.59%

Exact numbers out of 500 are:

Category ‘A’ = 3.59 × 5 = 17.95 = 18 (Approx.)

Category ‘B’ = 23.84 × 5 = 119.2 = 119 (Approx.)

Category ‘C’ = 45.14 × 5 = 225.7 = 226 (Approx.)

Category ‘D’ = 119 (Approx.)

Category ‘E’ = 18 (Approx.)

17.4 DIVERGENCE FROM NORMALITY

The important property of the normal curve is its symmetry about the mean and bell shaped curve. However, there may be distributions which may be different, then normal may be asymmetrical. For studying this it becomes essential to understand the nature of asymmetry. The asymmetry of the distribution is studied though the measure of skewness. In such cases the measures of central tendency and the dispersion are not sufficient to describe a distribution completely. In a normal distribution also there is a divergence in shape and size or form. The peak may also differ in different distributions and even in a number of normal distributions. It is also possible to have frequency distributions which differ widely in their nature and compositions but they have same measures of central tendency and dispersion. In uni-variate data, we need some more measures to supplement the measures of central tendency and dispersion so as to make it more complete in its description. Two such measures are:

(i) Skewness,

(ii) Kurtosis.

(i) Skewness

The skewness will help you in understanding the nature and kind of asymmetry in a distribution. It will be helping you in understanding the distributions which are not normal although they may be symmetrical. In histogram you got a general idea of the shape of a distribution but what about the amount and direction of asymmetry. How much was the deviation from symmetry and to which side is the distribution departing from a normal distribution can be understood by studying the skewness of the distribution.

The skewness can be defined as a degree of asymmetry of a distribution around its mean. It is due to the lack of symmetry. Skewness quantifies how symmetrical the distribution is. A symmetrical distribution has a skewness value of '0'. Two distributions have the same means as 20 and the same standard deviation of 5. But the two distributions may not be alike in nature. One may be symmetrical distribution but the other may not be symmetrical distribution. The second may be asymmetrical or skewed. Measure of skewness will help you to distinguish between different types of distributions. In such cases, the median falls at a point other than the mean.

There are two types of skewness which appear in the Normal Curve.

(a) Negative Skewness

(b) Positive Skewness

(a) Negative Skewness

Distribution is said to be skewed negatively or to the left, when scores are massed at the high end of the scale, i.e. the right side of the curve, and are spread out gradually towards the low end i.e. the left side of the curve. In a negatively skewed distribution the value of median will be higher than that of the value of the mean. Can you think why would it be so?

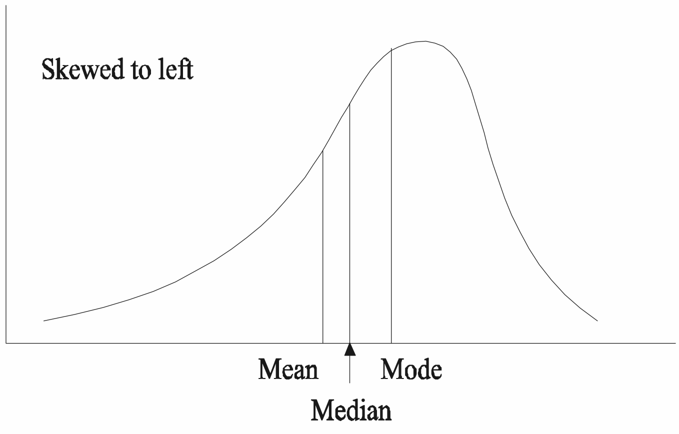


Figure 17.7: Negative Skewness

(b) Positive Skewness

Distributions are skewed positively or to the right, when scores are massed at the low, i.e. the left end of the scale, and are spread out gradually towards the high or right end as shown in the figure below.

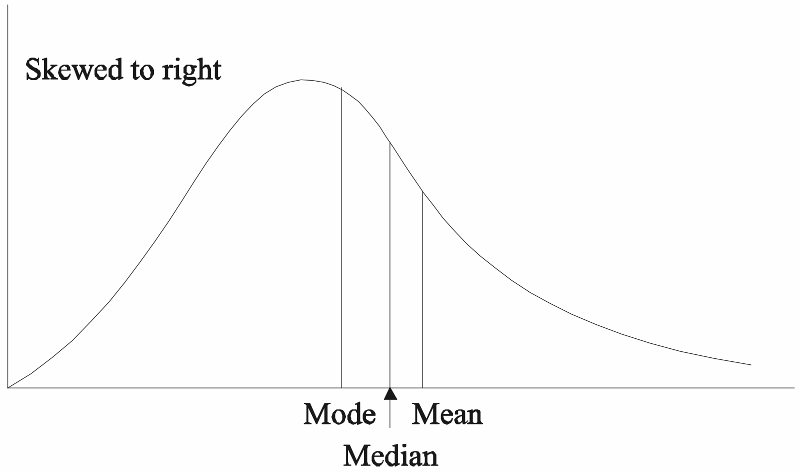


Figure 17.8: Positive Skewness

ii) Kurtosis

Another parameter of divergence is the vertical divergence in symmetry. Tallness or sharpness of the peak of a distribution comes under this category of divergence. Tallness and shapes of the central peak can be measured through kurtosis. Kurtosis will help you to get an idea of the height of the peak, sharpness of the peak or the broadness of the central peak of a distribution. Even if the distribution happens to be normal or symmetrical; the height, breadth or the sharpness of the central peak in such a distribution may vary. If we try to measure the height, breadth and the sharpness of the central peak with respect to the rest of the data, a new term or number comes and this number is known as kurtosis. Kurtosis is actually the measure of the shape of distribution.

There are two types of kurtosis which appear in the Normal Curve.

(a) Lepto kurtosis

(b) Platy kurtosis

(c) Mesokurtic

(a) Leptokurtosis

If you have a normal curve which is made up of steel wire and you push both the ends of the wire curve together, the curve become more peaked i.e. its top becomes narrower than the normal curve and scatterness in the scores or area of the curve shrink towards the centre.

Thus in a Leptokurtic distribution, the frequency is more peaked at the centre than in the normal distribution curve.

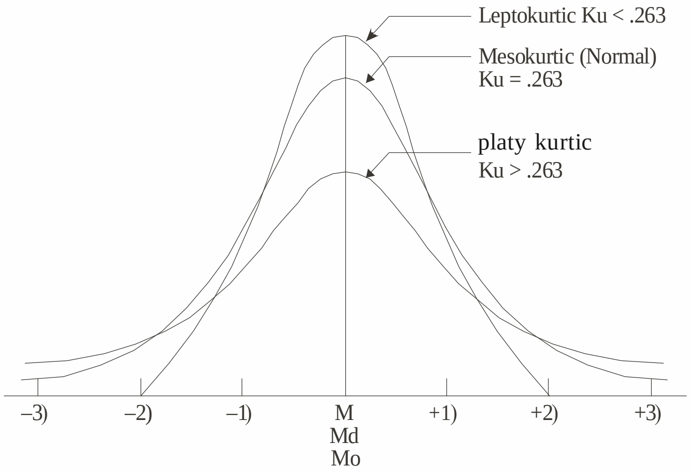


Figure 17.9: Kurtosis in the Normal Curve

(b) Platy kurtosis

If we put a heavy pressure on the top normal curve made from the steel wire. The top of the curve would become more flat than that of the normal.

Thus a distribution of flatter peak than of the normal distribution is known as platykurtic distribution.

When the distribution and related curve is normal, the value of kurtosis is .263 (Ku = .263). If the value of the Ku is greater than .263, the distribution and related curve obtained will be Platykurtic. When the value of Ku is less than .263, the distribution and related curve obtained will be Leptokurtic.

(C) Mesokurtic

Mesokurtic is called as the normal probability curve. Which is moderately peaked and not highly flattened. This curve is also not skewed to any direction. It follows all the characteristics of a normal distribution.

17.4.1 Factors Causing Divergence from Normal Probability Curve

The reasons why distributions exhibit skewness and kurtosis are numerous and often complex, but a careful analysis of the data can often throw some light on the symmetry. Some of the common causes are:

1. Selection of the Sample

Selection of the subjects (individuals) can produce skewness and kurtosis in the distribution. If the sample size is small or sample is biased one, skewness is possible in the distribution of scores obtained on the basis of selected sample or group of individuals.

The scores made by small and homogeneous group are likely to yield narrow and leptokurtic distribution. Scores from small and highly heterogeneous group yield platykurtic distribution.

2. Unsuitable or Poorly Made Tests

If the measuring tool of test is inappropriate for the group on which it has been administered, or poorly made, the asymmetry is likely to occur in the distribution of scores. If a test is too easy, scores will pile up at the high end of the scale, whereas when the test is too difficult, scores will pile up at the low end of the scale.

3. The Trait being measured is Non-Normal

Skewness or Kurtosis will appear when there is a real lack of normality in the trait being measured. e.g. interests or attitudes.

4. Errors in the Construction and Administration of Tests

A poorly constructed test may cause asymmetry in the distribution of the scores. Similarly, while administrating the test, unclear instructions, error in timings, errors in the scoring practice and lack of motivation to complete the test may cause skewness in the distribution.

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.

7. Define the following:

(a) Skewness

....

....

....

(b) Negative and Positive Skewness

....

....

....

(c) Kurtosis

....

....

....

(d) Platykurtosis

....

....

....

(e) Leptokurtosis

....

....

....

(8) In case of normal distribution what should be the value of kurtosis?

....

....

....

....

(9) What is the significance of the knowledge of skewness and kurtosis for a school teacher?

....

....

....

(10) Fill in the blanks.

1. Skewness is the degree of .... of a distribution.
2. Negatively skewed distribution has .... in a distribution curve.
3. Positively skewed distribution has .... in a distribution curve.

17.5 LET US SUM UP

The normal distribution helps us to study variables used in behavioural research because they tend to be normally distributed.

Normal curve is very helpful in educational evaluation and measurement. It provides relative positioning of the individual in a group. It can also be used as a scale of measurement in behavioural sciences.

The normal distribution is a significant tool in the hands of teacher, through which one can decide the nature of the distribution of the scores obtained on the basis of measured variable. He/she can judge the difficulty level of the test items in the question paper and finally he may know about his/her class, whether it is homogeneous to the ability measured or it is heterogeneous one. The normal distribution also helps any individual to classify the distribution in different categories as per needed criteria.

17.6 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.

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.), Lowa State University Press, Ames, Lowa.

17.7 ANSWERS TO CHECK YOUR PROGRESS

1. Normal Probability Curve is a bell shaped curve obtained for a distribution having maximum frequency near the central value of distribution and the frequency gradually tapers off symmetrically on both the sides.

2. The Normal Curve is symmetrical about the ordinate at the central point of the curve.

   - It is unimodal; the mode is always at the central point of the curve.
   - It is asymptotic to the x-axis.
   - The points of influx occur at ±1.
   - The area of the curve between the points of influxation is fixed.

3. Normal Distribution is important in education for use in biological statistics, psychological measurement, errors in observations, ability testing, estimation of scores, categorisation, etc.

4. (a) Between –1 σ and + 1 σ, there are 68.26% of the frequencies

   (b) Between –2 σ and + 2 σ, there are 95.44% of the frequencies

   (c) Between –3 σ and + 3 σ, there are 99.73% of the frequencies

5. The two ends of the normal probability curve are considered closed at the points ± 3 σ, as almost all the cases (99.73% of the cases, to be exact) lie between these two points and there is a rare probability of a case going beyond these two limits.

6. i. Bell shaped

   ii. Mean = 0 and Standard Deviation = One

   iii. Normally

iv. 34.13 percent

v. 47.72 percent

vi. 2 σ

vii. 2.58 σ

7. (a) A distribution is said to be skewed, if the point of centre of gravity is located on one side of the distribution i.e. away from the centre of the scale of measurement.

   (b) A distribution is said to be negatively skewed if the scores are concentrated at the higher end of the measurement scale and it is said to be positively skewed it the scores are concentrated at the lower end of the measurement scale.

   (c) Kurtosis refers to the divergence in the height of the curve or the peakedness of the curve.

   (d) A distribution of flatter peak than the normal one is known as platy – kurtosis distribution.

   (e) A distribution which is more peaked than the normal one is known as lepto-Kurtosis distribution.

8. In case of normal distribution the value of Kurtosis is 0.263.

9. If the distribution of scores obtained by a school teacher is not normal, he/she will try to find out the reasons of skewness and kurtosis of the distribution. One of the reasons may be that the group of individuals are different from the normal one, or it may be due to the nature of the trait itself. Teacher should know that ultimately the goal of education is to have the distribution of scores of individuals as negatively skewed.

10. i. Asymmetry

    ii. Skewness on left

    iii. Skewness on right