PUNJABI UNIVERSITY PATIALA

M.A. (EDUCATION) PART-I
SEMESTER-I

PAPER-III
METHODOLOGY OF EDUCATIONAL RESEARCH-I

UNIT NO - II

Department of Distance Education
Punjabi University, Patiala

Part-B

Lesson No.

2.1 : Quantitative measurement and level of measurement, frequency distribution, graphical representation of data

2.2 : Measures of Central Tendency

2.3 : Measures of Variability

2.4 : Normal Probability Curve

Note : Students can download the syllabus from department's website www.pbidde.org

Lesson No. 2.1

Quantitative Measurement and Graphical Representation of Data

Structure of the Lesson :

2.1.1 Objectives :

After going through this lesson learners will be able to:

i) Know about levels of measurement
ii) Define frequency distribution
iii) Describe the need of graphical representation of data
iv) Explain the different methods of graphical representation of data
v) Draw the different types of graphical representation of data

2.1.2 Quantitative Measurement and Level of Measurements

Quantitative Measurement: Characteristic of an individual can be measured either by quantitative or qualitative. Characteristic which measures quantitatively involves a numerical measurement. Thus quantitative measurements are those which can be put into numbers or involve the collection of numbers.

Level of measurement of a variable is helps us to determine what kinds of statistical tools may be used to describe the variable, and decide what statistical analysis is appropriate on the values that were assigned to the variable. In other words, the level of measurement determines the types of descriptive and inferential statistics that may be applied to the variable.

There are typically **four levels of measurement** that are:

i) Nominal
ii) Ordinal
iii) Interval
iv) Ratio

2.1.2.1 Nominal Variables

- A type of categorical data in which objects fall into ***unordered*** categories i.e These are Categorical variables with no inherent order or ranking
- There is no preference of one category over other
- Nominal variables may be coded with numbers.

Examples

i) Gender (Male and Female)
ii) Hair Color (Black, Brown, Gray, etc.)
iii) Marital Status (Married, Single, Divorced, etc.)

2.1.2.2 Ordinal Variables

- A type of categorical data in which ***order*** is important i.e Variables with an inherent rank or order
- One can arrange the categories in ascending or descending order of magnitude
- The numerical distance between any categories in unknown
- Can be compared for greater or less, but not *how much* greater or less.

Examples

i) Grades (A, B, C, D. E)

ii) Evaluations
  - High, Medium, Low
  - Likert Scales
  - 5 pt. (Strongly Agree, Agree, neither Agree nor Disagree, Disagree, Strongly Disagree) and
  - 7 pt. liberalism scale (Strongly Liberal, Liberal, Weakly Liberal, Moderate, Weakly Conservative, Conservative, Strongly Conservative)

2.1.2.3 Interval Variables

- Values of the variable are ordered as in Ordinal
- Addition and subtraction may be valid, but not multiplication and division are meaningful operations
- Zero does not mean the absence of Character.

Variables or measurements where the difference between values are measured by a fixed scale.

Examples

i) Temperatures on the Fahrenheit scale

ii) Education (in years)

2.1.2.4 Ratio Variables

i) Variables with all properties of Interval plus an absolute, non-arbitrary zero point

ii) All mathematical operations are valid i.e. Addition, Subtraction, Multiplication, and Division are all meaningful operations

iii) This is the highest rating scale.

Examples: Height, Weight, Age etc.

Check your progress

Q1. What do you understand by level of measurement?

Q2. What are the four levels of measurement?

2.1.3 Frequency Distribution

A frequency distribution is a tabular summary of data showing the frequency (or number) of various outcomes in a sample. Each entry in the table contains the frequency of the occurrences of values within a particular group (or interval). The main objective of a frequency distribution is to provide insights about the data that cannot be quickly obtained by looking only at the original data. Thus, with the help of frequency distribution the information about the original data can be interpreted more easily.

Example

Frequency Distribution

Rating	Frequency
Poor	2
Below Average	3
Average	5
Above Average	9
Excellent	1
Total =	20

2.1.4 REPRESENTATION OF DATA

The transformation of data through graphs, diagrams, maps and charts is called representation of data. The pictorial representation of data is a better tool for conveying details of the data to the reader in a simpler and summarized form. Thus, after placing the frequency distribution into tabular form, the next stage comes of presenting it graphically. Graphs have a better sense of appeal than tables and as such many conclusions regarding the shape and pattern of the distribution are easier to be drawn with graphs.

2.1.4.1 Need of representing data graphically

The diagrammatic or graphically study of the data is necessary because:

- If the figures in the tabular form are large in number or in size then their study needs much time.
- The graphic method of the representation of data enhances our understanding.
- It makes the comparisons easy.
- Such methods create an imprint on mind for a longer time.
- It is a time consuming task to draw inferences about whatever is being presented in non–graphical form.
- It presents characteristics in a simplified way.
- It makes easy to understand the patterns of population growth, distribution and the density such as sex ratio, age–sex composition, occupational structure, etc.

2.1.5 Methods of graphical presentation of data

- Histogram
- Frequency polygon
- Cumulative frequency curve or Ogive

2.1.5.1 Histogram

A histogram is a bar graph that shows how frequently data occur with certain ranges or intervals. The height of each bar gives the frequency in the respective interval.

A **histogram** is a most common device for representation of a frequency distribution of data. This type of graphical representation is more suitable for frequency distributions with continuous classes. In this type of distribution the upper limit of a class is the lower limit of the following class. Here the class frequencies are represented by areas of the vertical rectangles, having their bases on horizontal line with centers at the mid-values and width are equal to the size of the class intervals.

Features of a Histogram:

- The height of the column shows the frequency for a specific range/interval of values.
- Columns are usually of equal width, however it is as good for equal interval as for unequal.
- The values represented by each column must be mutually exclusive and exhaustive. Therefore, there are no spaces between columnsand each observation can only ever belong in one column.

- It is important that there is no ambiguity in the labelling of the intervals on the x-axis for continuous or grouped data.

Drawing the Histogram

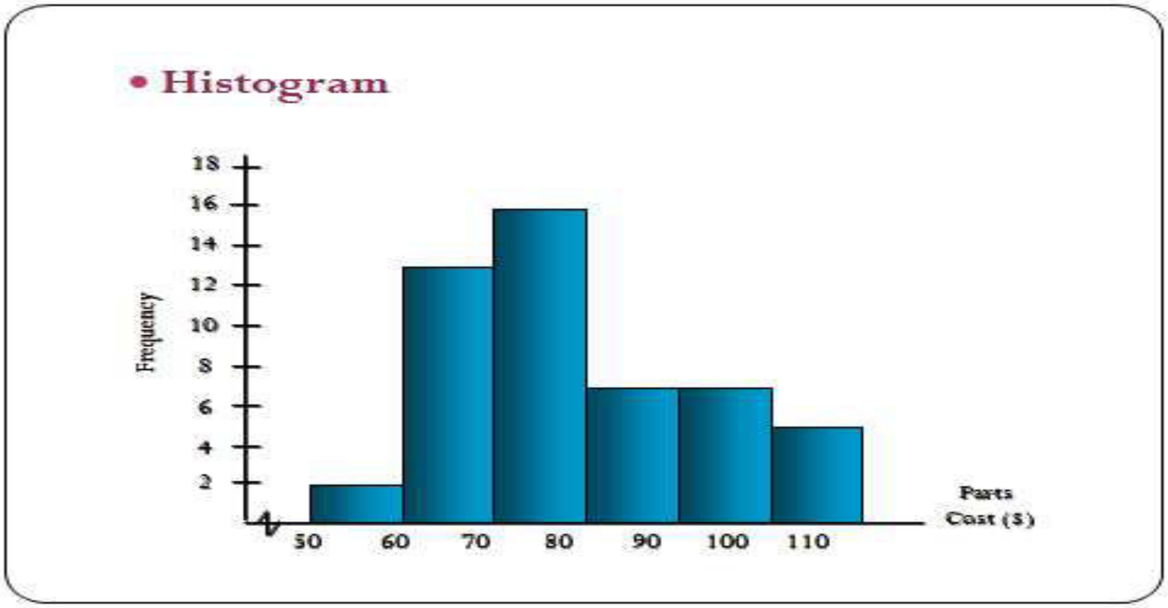
After organizing data into frequency distribution (with classes), we are ready to draw our histogram.

1. Draw a horizontal line on which we denote our classes.
2. Place evenly spaced marks along this line that correspond to the classes.
3. Label the marks so that the scale is clear and give a name to the horizontal axis.
4. Draw a vertical line just to the left of the lowest class.
5. Choose a scale for the vertical axis that will accommodate the class with the highest frequency.
6. Label the marks so that the scale is clear and give a name to the vertical axis.
7. Construct bars for each class. The height of each bar should correspond to the frequency of the class at the base of the bar.

For example:

Let us draw a Histogram for the following Frequency Distribution

Cost ($)	Frequency
50-60	2
60-70	13
70-80	16
80-90	7
90-100	7
100-110	5
Total =	**50**



2.1.5.2 Frequency Polygons

A frequency polygon is a graph that displays the data by using lines that obtained by connecting points plotted for the frequencies at the midpoints of the classes. Frequency polygon are quite similar to that of the corresponding histogram but instead of rectangles drawn on class intervals, here vertical lines are erected at the class mid-points with heights in proportional to class frequencies.

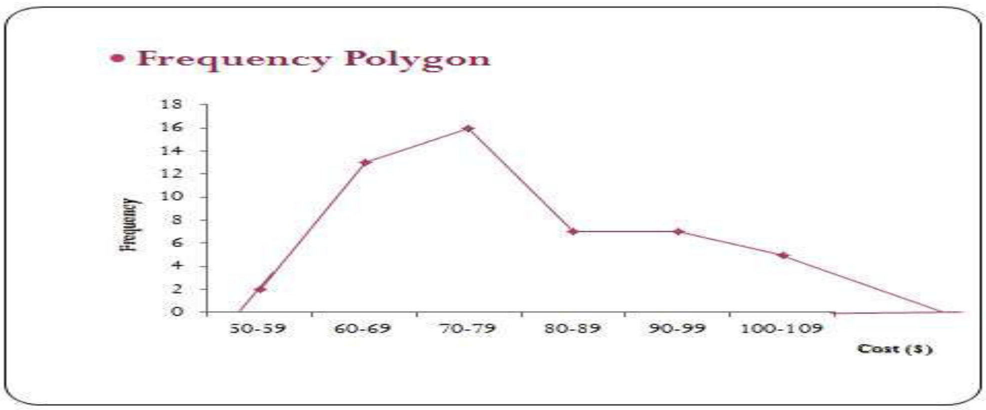
Steps for constructing a Frequency Polygon:

1. Draw and label the X(horizontal) and the Y (vertical) axes.
2. Mark class interval along the X-axis.
3. Represent the frequencies on the Y axis and the midpoints on the X axis.
4. Draw points representing frequencies against mid points of the respective intervals.
5. Connect adjacent points with line segments. Draw a line back to the X axis at the beginning and the end of the graph at the same distance that the previous and the next midpoints would be located.

Note: For the frequency polygon we need the frequencies and the midpoints.
Example:

Let us draw afrequency polygon for the following Frequency Distribution

Cost ($)		Frequency
50-59		2
60-69		13
70-79		16
80-89		7
90-99		7
100-109		5
Total	=	**50**



2.1.5.3 Cumulative Frequency Curve or Ogive

An ogive is a graph where the cumulative frequencies are used and not the class frequencies. Cumulative frequencies are obtained by adding frequencies either from the top to bottom or from bottom to top. The obtained cumulative frequencies are plotted along the y-axes against the upper or the lower class limits as the case may be. If the different points so plotted are joined through straight lines then the graph (smooth curve) so obtained is called cumulative frequencies curve or Ogive. The ogive curve may be traced either on less than basis (if the cumulative frequencies taken from top to bottom) or more than bases (if the cumulative frequencies taken from bottom to top)

Steps for constructing an ogive:

1. Draw and label the X(horizontal) and the *Y* (vertical) axes.
2. Represent the cumulative frequencies on the *Y* axis and the class boundaries on the *X* axis.
3. Plot the cumulative frequency at each upper class boundary with the height being the corresponding cumulative frequency.
4. Connect the points with segments. Connect the first point on the left with the *X* axis at the level of the lowest lower class boundary.

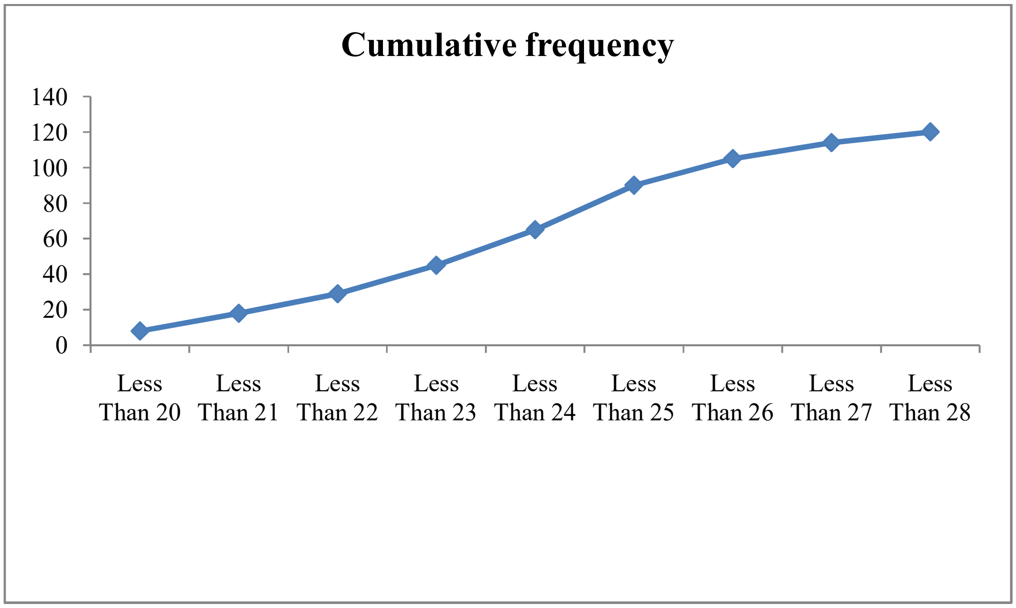
Note: For the ogive we need the class boundaries and the cumulative frequencies

Example: Draw a cumulative frequency graph of the following data:

Wages (Rs)	No. of workers (Frequency)	Cumulative frequency
Less Than 20	8	8
Less Than 21	10	18
Less Than 22	11	29
Less Than 23	26	45
Less Than 24	20	65
Less Than 25	25	90
Less Than 26	15	105
Less Than 27	9	114
Less Than 28	6	120

Solution:

Plotting the upper class limits such as 20,21,22…..,etc., on the x-axis and Cumulative frequencies on Y-axix and then joining the plotted points, the required cumulative graph or less than ogive is obtained as



2.1.6 Summary

The theory of measurement has been developed in combining with the concepts of numbers and units of measurement. Statisticians classify measurements according to levels. Each level corresponds to how this measurement can be treated mathematically. There are typically four levels of measurement that are-Nominal, Ordinal, Interval and Ratio.After presenting the data into tabular form, data may be displayed in pictorial form by using a graph. A well-constructed graphical presentation is the easiest way to depict a given set of data. There are different methods of graphical presentation of data like Histogram, Frequency polygon and cumulative frequency curve or Ogive.

2.1.7 Suggested Questions

Q1 Construct a Histogram for the following distribution.

Class Interval (C.I.)	Frequency (f)
51 – 55	1
46 – 50	1
41 – 45	2
36 – 40	5
31 – 35	7
26 – 30	6
21 – 25	5
16 – 20	3
11 – 15	1

Q2. Construct a Frequency Polygon for the following distribution.

Class Interval (C.I.)	Frequency (f)
60 – 64	1
55 – 59	1
50 – 54	3
45 – 49	4
40 – 44	6
35 – 39	7
30 – 34	13
25 – 29	6
20 – 24	8
15 – 19	2

Q3. Draw a cumulative frequency curve of the following data.

Class Interval (C.I.)	Frequency (f)
90 – 94	1
85 – 89	2
80 – 84	4
75 – 79	5
70 – 74	8
65 – 69	10
60 – 64	6
55 – 59	4
50 – 54	4
45 – 49	2
40 – 44	3
35 - 39	1

Q4. Draw a cumulative frequency curve of the following data.

Class Interval (C.I.)	Frequency (f)
100- 109	1
90 – 99	2
80 – 89	4
70 – 79	6
60 – 69	5
50 – 59	10
40 – 49	8
30 – 39	5
20 – 29	4
10 – 19	3
0 – 9	1

2.1.8 Suggested Readings

Aggarwal, B.L. : Basic Statistics, New age Publication

Garret, H.E. : Statistics in Psychology and Education, Bombay: Vakils, Feffer and Simons Ltd.

Gupta, O.P : Mathematical Statistics, Kedarnath, Ramnath & Co.ss

M.A.(EDUCATION)PART-I
(SEMESTER-I)

PAPER - III
METHODOLOGY OF
EDUCATIONAL RESEARCH-I

Lesson No. 2.2 **Author : Dr. (Mrs.) S.K. Bawa**

MEASURES OF CENTRAL TENDENCY

Structure of the Lesson :

2.2.1 OBJECTIVES :

After going through this lesson, you should be able to:

(i) Explain the meaning of measures of central tendency.

(ii) calculate mean from ungrouped and grouped data and interpret it.

(iii) compute median from ungrouped and grouped data and interpret it.

(iv) enlist situations which call for use of median.

(v) compute mode from ungrouped and grouped data and interpret it;

(vi) Compare measures of central tendency.

2.2.2 INTRODUCTION :

Data are gathered on various attributes in order to ascertain the performance of the individual or a group of individuals. This helps in better understanding of the individuals or the group. The obtained data are classifiable on the scales of measurement e.g. nominal, ordinal, interval and ratio-ratio scale. It is also possible to convert these from one to other scale. A measure of central tendency represents average and gives concise description of the performance of the group as a whole. So as to allow comparability of groups in terms of typical performance. Measures of central tendency are thus used to interpret the nature of scores obtained by the group in general.

Measures of central tendency are most common with the statisticians because these measures help them to reduce the complexity of data and make it comparable. We cannot remember the whole set of data and the analysis of such data is impossible. So in order to reduce the complexity and make the data comparable we resort to averaging. This average must be representative of the whole data. Different statisticians from time to time give definitions of an average. According to J.P. Guilford, "An average is a central value of a group of observations or individuals." Clark says, "Average is an attempt to find one single figure to describe whole of figure".

2.2.3 CHARACTERISTIC OF AVERAGE:

According to Yule and Bowley following are characteristics of central tendency.

1. **It should be rigidly defined :** We have said that average should be representative of the distribution of the series. It is essential that average should be rigidly defined because it will make the result a definite one

and will leave no scope for the entry of deliberate basis of the statistician. The results will also be representative.

2. **It should be calculated with reasonable ease and rapidity :** Average should be such that it can be easily and rapidly calculated by men of mediocre intelligence. It should not involve highly complicated and advanced mathematical formulas and principles for the calculation, because this will limit the use of average and it will not become popular.

3. **It should be suitable for further mathematical treatment :** In other words, the average should possess some important and interesting mathematical properties so that its use in further statistical theory is enhanced. If an average is not amenable to further algebraic manipulation, then obviously its use will be very much limited for further applications in statistical theory.

4. **It should be based on all the observations made :** This characteristic will make the average as a representative one of the whole series. Even a single change in the group will lead to change in the average calculated and this change will make that changed figure represent itself through new average.

5. **It should not be affected much by extreme observations :** By extreme observation we mean very small or very large observations. Thus a few very small or very large observations should not unduly affect the value of a good average.

6. **If there is a special type, it should show it :** Average should be such that it represents the series accurately. Sometimes series are of a special type or they contain some particular feature, the average must represent them accurately.

7. **It should be least affected by fluctuations of sampling :** Average must not be affected by sampling fluctuations. If a sample is taken from a universe, the average of the sample should not be significantly different from the average of the universe. If it differs, it will mean that our sample average is not truly representative.

2.2.4 MEAN :

The mean is the average value of the total scores in a distribution. To simplify, the mean is a sum of the separate scores divided by their number. For example, five students of a class have secured 25,10,16,24 and 30 marks respectively in a test. The average score of the class in the test will be

$$\frac{25+10+16+24+30}{5} = \frac{105}{5} = 21$$

Therefore, we can say that mean is the sum of all scores in a distribution dividied by the number of scores in it.

2.2.4.1 Calculation of Mean from Ungrouped Data : The mean for ungrouped data can be calculated by using any of the following two methods:

(i) Direct method; (ii) Short-cut method

(i) Direct Method : When the number of scores is not too large, we calculate the mean simply by adding the scores and dividing them by the number of scores.

Example : Calculate the mean of the following scores :

12, 14, 18, 24, 20, 15, 13, 22, 26, 30

$\Sigma X=12+14+18+24+20+15+13+22+26+30 = 194$

(Here Σ is a Geek letter called (Capital) sigma. It stands for sum (total) of scores and X stands for raw scores. The mean of a population is symbolized by μ, number of elements is N but the mean of a sample is denoted by M and number of elements by n.

So, N = 10

$$\text{Mean (M)} = \frac{\Sigma x}{N} = \frac{194}{10} = 19.4$$

(ii) Short-cut Method : By choosing an assumed mean and calculating deviations of the given varieties or observation from it makes the calculation of mean simpler. Thus average is usually chosen to be a neat round number in the middle of the range of the observations, so that deviations can be easily obtained by subtraction. Then, a formula, based on deviations from assumed mean, for calculating arithmetic mean becomes:

$$\text{Mean} = A.M + \frac{\Sigma d}{N} \qquad \ldots(03)$$

Where A.M. = Assumed Mean.

Σ = Sum total of

d = The deviation of each value of the variable from the assumed Mean

N = Total Nos.

Example : A random sample of 10 boys had the following intelligence quotients (I.Q's)

70, 120, 110, 101, 88, 83, 95, 98, 105, 100

Find the mean I.Q.

Solution : In this example, the ungrouped data ranges from 70 to 120. Therefore, 95 a neat round value in the middle of 70 and 120 may be taken as assumed mean, i.e. A.M. = 95 Deviations and sum of deviations needed in formula may be calculated in a table given below:

Table : 2.1 Computation of Mean (Short-cut Method)

Sr. No.	**I.Q (X)**	**Deviations from A.M. = 95 D=(X-A.M.)**
1	70	-25
2	120	25
3	110	15
4	101	6
5	88	-7
6	83	-12
7	95	0
8	98	3
9	105	10
10	100	5
		$\Sigma d = 20$

Using formula the arithmetic mean will be

$$\text{Mean} = \text{A.M} + \frac{\Sigma d}{N} \qquad (\text{A.M.} = 95,\ \Sigma d = 20,\ N=10)$$

$$= 95 + \frac{20}{10} = 95 + 2 = 97$$

2.2.4.2 Calculation of Mean from Grouped Data :

Calculation of mean for a frequency distribution can also be done with the two procedures namely (i) Direct method/Long method (ii) Short method. When measures are too many, it is time consuming to add all the scores and then divide the total scores with N to find the mean. A short method is therefore used by grouping the data into a frequency table.

Example : Calculate the mean from the following grouped data :

Table 2.2.2

Class interval	**frequency (f)**	**(x)**	**fx**	
35-39	5	3	15	36
30-34	6	2	12	
25-29	9	1	9	
20-24	11	0	0	
15-19	8	-1	-8	-42
10-14	7	-2	-14	
5-9	4	-3	-12	
0-4	2	-4	-8	
	N=52		Σfx = -6	

$$\bar{x} = A.M. + \frac{\Sigma fx}{N} \times i, \quad AM = \left(\frac{20+24}{2}\right) = 22$$

$$\bar{x} = 22 + \frac{-6}{52} \times 5 = 22 - .57$$

Mean = 21.43

Where AM stands for assumed mean which is the mid value of the class interval having the maximum frequency.

Σ = sum

f = frequency

x = deviation

N = number of frequencies

i = class interval

2.2.4.3 Steps to Calculate Mean :

1. Prepare a frequency table if data are not given in the tabular form and calculate the frequencies. Make sure that lower scores are at the bottom.
2. Calculate the assumed mean by taking the middle point of the class-interval having the highest frequency. In this example, A.M. of CI 20-24 is 22.0
3. In the third column, the deviations from the assumed mean are to be entered.

   Against CI 20-24 the deviation will be Zero because A.M. = 22.0 and 22.0 does not deviate from 22.0. Therefore, write 0 against the interval. As we move towards the class intervals of increasing scores the deviations from the true mean are 5.0, 10.0, 15.0 and 20.0 and as we move towards decreasing class-intervals, the deviations are – 5.0, -10.0, -15.0, -20.0 and –24.0 (all from A.M. = 22.0). We see that as we go up, deviations from mean go up by 1,2,3 and so on and as we come down it goes down by –1, -2, -3, -4 and so on. Thus, write 1,2,3,4 and so on towards the class intervals of increasing scores and –1, -2, -3, -4 and so on towards the class intervals of decreasing scores.
4. The fourth column fx is the product of column 2 and 3 and e.g.

   5 x 3 = 15, 6x2 = 12 and so on
5. Now add all the positive values and the negative values separately, which are + 36 and –42 in the above example and add these two values, this will give Σfx which comes out to be –6 in the example.
6. Put all the values in the formula to calculate mean by short cut method i.e.

   $$x = A.M. + \frac{\Sigma fx}{n} \times i$$, where AM = 22.0, Σfx = -6

   i = 5 and N = 52, the mean calculated is 21.43.

2.2.4.4 PROPERTIES OF MEAN :

The following are the main properties of mean :

1. The sum of the deviation from the mean of the scores is always zero. Mean is the central value of a distribution of scores.

2. The sum of the squares of deviations from the mean is always less than the sum of the squares of deviations from any other assumed mean.
3. The mean is central value of a set of data. Both sides from the mean, the deviations are equal.
4. The mean is sensitive measure of central tendency. If we add, substract, multiply or divide the set of scores by a constant, the mean value may change.
5. The mean value is a measure of sample or group. It is the most accurate measure of central value of the scores. The group performance is best indicated by this value when the distribution indicated by this value when the distribution is normal. It is used for interval and ratio variables.

2.2.4.5 MERITS OF MEAN :

(i) It is clearly defined.

(ii) It is easy to calculate and understand.

(iii) It is based on all observations.

(iv) It is suitable for further mathematical treatment.

(v) It is least affected by fluctuations in sampling.

2.2.4.6 DEMERITS OF MEAN :

I. Every single score affects the mean.

II. It cannot be located graphically or determined by casual inspection.

III. Qualitative characteristics cannot be dealt with mean.

IV. Mean cannot be calculated if even single observation is missing.

SELF EVALUATION QUESTIONS:

1. Define mean and write its merits:

____

____

____

2. Calculate mean for the following frequency distribution :

C.I =	50-54,	55-59,	60-64,	65-69,	70-74,	75-79,	80-84.
Frequency =	5,	8,	17,	26,	13,	12,	9.

____

____

____

2.2.5 MEDIAN :

Median may be defined as the size of that item which falls just in the middle of a series or data arranged either in the ascending order or the descending order of their magnitude. Median is the size of that item which has as many items proceeding it as succeeding it. It lies in the center of a series or data and divides the series or data into two parts containing an equal number of items:

2.2.5.1 Calculation of Median from Ungrouped Data :

(i) Individual Series : The formula of calculating median is

e.g. Find out the median value of the following data :

3,9, 2, 8, 7, 4, 1, 3, 5, 4, 6, 2, 9, 7, 8, 4, 6

Arranging 1,2,2,3,3,4,4,4,5,6,6,7,7,8,8,9,9,

$$M = \text{Size of } \frac{N+1}{2} \text{ th item}$$

$$= \text{Size of } \frac{17+1}{2} \text{ th item} = 9\text{th item i.e.} = 5$$

Thus value of Median = 5

2.2.5.2 Calculation of Median from Grouped Data : When the data are available in grouped form i.e. in the form of frequency distribution, median can be calculated as follows :

Table 2.2.3

Class interval (ci)	Frequency (f)	Cumulative Frequency (cf)
35-40	2	40

30-34	5	38
25-29	7	33
20-24	10	26
15-19	9	16
10-14	4	7
5-9	0	3
0-4	3	3
	N = 40	

$$\text{Median} = l.l + \frac{N/2 - cf}{Fm} \times I$$

Where *l.l* is the lower limit of the class interval in which median lies.

N/2 = half the number of frequencies.

Fm = Frequency of the interval in which the median lies.

cf = cumulative frequency that lies below of class interval which contains the median. In above data, N/2 = 40 / 2=20.

Lower limit = 19.5

Fm = 10

cf = 16

i = 5

$$\text{Median} = 19.5 + \frac{40/2-16}{10} \times 5$$

= 19.5 + 2.0

Median = 21.5

2.2.5.3 Steps to calculate median :

I. Calculate cumulative frequency in column 3 of the table by adding the frequencies from below and writing them against each class interval. For example, the frequency in the last interval is 3 Add three to the above frequency 4 and write 4+3 = 7 against third class interval from below. Similarly, add 9 to 7 and write 16 against fourth class interval from below till all the cumulative frequencies are calculated.

II. Divide N by 2. here, 40/2 = 20.

III. We have to reach a point where 20 frequencies are summed up we find against 15-19 CI, there are 16 frequencies summed up. To make 20, we need 4 more frequencies. We take them from this next CI which has 10 frequencies.

IV. Class interval is 5 in the given example. Put all the values in the formula and calculate the median.

2.2.5.4 Merits of Median :

(i) It is rigidly defined.

(ii) It is easy to understand and calculate

(iii) It is not affected by extreme observations.

(iv) It can be located by simple inspection and can be computed graphically.

(v) It can be used for dealing with qualitative characteristics also. We can find average intelligence, average academic achievement etc., among a group of people.

2.2.5.5 Demerits of Median :

(i) It is not based on every item of the distribution

(ii) It is not suitable for further mathematical treatment

(iii) It is relatively less table than mean particularly for small sample because it is affected more by the fluctuations of sampling.

2.2.6 MODE : Mode is the value which occurs most frequently in a set of observations, for example, in the following set of scores : 13,15,15,18,14,12,12,12,16,19 only the score 12 occurs thrice. Therefore, we can say the mode in the above set of scores is 12. Important features of mode are it is the size of that item which has the maximum frequency. It is effected by frequencies of the neighbouring items.

2.2.6.1 Calculation of Mode : In an ungrouped data, mode can easily be located by counting the value which occurs most times in the distribution as in the above said case. But mode in grouped data can be calculated by using the following two formulas.

(1) Mode = 3 Median - 2 Mean.

(2) $$\text{Mode} = l.l. + \frac{f_0 - f_1}{2f_0 - f_1 - f_2} \times i$$

Where $l.l$ = lower limit of the class interval in which mode lies.

f_o = frequency against the modal CI

f_1 = frequency against the class interval preceding the model CI

f_2 = frequency against the class interval succeeding the model CI

i = class interval.

Table 2.2.4

Class Interval (ci)	frequency (f)
70-79	1
60-69	8
50-59	10
40-49	5
30-39	6
20-29	4
10-19	3
0 - 9	3

Using the above formula

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

$$49.5 + \frac{10 - 5}{2 \times 10 - 5 - 8} \times 10$$

$$= 49.5 + 7.1$$

Mode = 56.6

2.2.6.2 Step to Calculate Mode :

(i) Locate where maximum frequencies lie. This class interval will be called modal class. In the example, since maximum frequencies are 10 in class interval, 50-59, lower limit of which is 49.5.

(ii) The frequencies in modal (i) are denoted by to which is 10, the frequencies above the modal class interval are 8, which are denoted by f_2 and frequencies below the modal class interval are 5, denoted by f_1.

(iii) Put all the values in the above formula and calculate mode. There second method of finding the mode uses this following formula is Mode = 3 Mdn. - 2 mean.

This formula is the most correct result-giving formula. As such it is widely used. It makes use of mean and median.

2.2.6.3 Merits of Mode :

1. It is easy to understand and calculate.
2. It can be located graphically or by mere inspection.
3. It is not affected by extreme observation.

2.2.6.4 Demerits of Mode :

1. It is not rigidly defined.
2. It is not based on all observation of the distribution.
3. It is not suitable for a further mathematical treatment.
4. Mode is affected to a greater extent by fluctuations of sampling.

2.2.7 COMPARISON OF MEAN, MEDIAN AND MODE

Mean	Median	Mode
1. Mean is interval variable.	1. It is ordinal variable.	1. It is normal variable.
2. It is the central location of scores	2. It is the central location of the frequency.	2. The score occurs most frequently.
3. It is a sensitive measure of central tendency.	3. It is less sensitive measure of central tendency.	3. It is crude measure of central tendency.
4. It indicates the skewness of the distribution.	4. It is used for the skewness of distribution.	4. It indicates the modality of the distribution.
5. It is used or normal distribution accurately.	5. It is used for the skewed distribution.	5. It is used for uni, bi and multi model nature.

6. It has theoretical value rather than practical. 7. It is used for comparing two or more groups for variance.	6. It has theoretical value rather than practical. 7. It is used when extreme scores fluctuate.	6. It has high practical value. 7. It is quick and approximate measure of central tendency.

2.2.8 SUMMARY

Measure of central tendency are most common with the statisticians because these measures help them to reduce the complexity of data and make it comparable. We cannot remember the whole set of data and the analysis of such data is impossible. So in order to reduce the complexity and make the data comparable, we resort to averaging. Average must be representative the whole data, it should be based on all the observations, suitable for further mathematical treatment, calculated with reasonable ease and rapidity, rigidly defined and least affected by fluctuations of sampling. Mean, median and mode are measure of central tendency.

2.2.9 SELF EVALUATION :

1. Define median and mode along with their merits and demerits.
2. Calculate Mean, Median and Mode of the following grouped data:

	class interval	frequency
(a)	90-100	1
	80-90	3
	70-80	5
	60-70	10
	50-60	25
	40-50	9
	30-40	7
	20-30	7
	10-20	03
		70
(b)	40-44	1
	35-39	2
	30-34	8

	25-29	10
	20-24	19
	15-19	11
	10-14	6
	5 - 9	3
	0 - 4	01
		61
(c)	90-99	5
	80-89	10
	70-79	15
	60-69	20
	50-59	45
	40-49	15
	30-39	8
	20-29	7
	10-19	04
		124

2.2.10 SUGGESTED READINGS :

Garrett, H.E. *Statistics in Psychology and Education.*

Guilford, J.P. *Fundamental Statistics in Psychology and Education*

Sharma, T.R. *First Course in Statistics*

M.A.(EDUCATION)PART- I
(Semester-I)

PAPER - III
METHODOLOGY OF
EDUCATIONAL RESEARCH-I

Lesson No. 2.3 **Author : Dr. Y.P. Aggarwal**

Educational Statistics : Range, quartile and Standard Deviation

Structure of the Lesson :

2.3.1 OBJECTIVES :

This lesson is intended to realize the following objectives:

(i) to make the students aware about dispersion.

(ii) to enable the student to understand different types of dispersions and their merits and demerits.

(iii) to make the student understand the calculations of Range at Mean Deviation and Variance.

(iv) to enclose the students to understand and calculate standard deviation from grouped & ungrouped data.

2.3.2 INTRODUCTION :

The average gives an idea about the central values of the series but it does not give any idea how the varieties are clustered around or scattered away from the point of central tendency. The average value cannot give a complete idea about the homogeneity or heterogeneity of the series. The meaning of dispersion is Scatteredness". Thus the measurement of the scatter of the size of the items of a series about the average is said to be a measure of variation or dispersion. The following measure of dispersion are generally in use :

(i) The Range

(ii) The Mean Deviation or Average Deviation (A.D.)

(iii) Variance

(iv) Standard Deviation

(v) Semi-interquartile Range

2.3.3 THE RANGE :

The Range is the distance between the highest score and lowest score in the distribution.

Range = The highest score - The lowest score.

The Range in the three sets of scores given in Table-1 can be calculated as follows :

Table 2.3.1

Set x	Highest Score	Smallest Score	Range (Highest Score-Lowest Score)
1	10	10	10-10 = 0
2	13	7	13-7 = 6
3	19	1	19-1 = 18

Interpretation : Set 1: All scores are covered within a score distance of Zero units.

Set 2: All Scores are covered within a score distance of six units.

Set 3: The student may interpret this result himself.

Technically, the range should be defined as the difference between the upper real limit of the largest score minus the lower real limit of the smallest score. Since the range is at best approximate index of variability, it does not seem appropriate to insist upon this level of accuracy. As evident from formula x, range takes into account the extremes of the scores only and ignores others. Hence, it suffers from the following limitations.

2.3.3.1 Limitations of Range :

1. It is unrealiable when N is small or when there are gaps in the frequency distribution.
2. A change in the value of either the highest score or the lowest score leads to change in its value.
3. It does not consider the value of the scores between the highest and the smallest scores and does not reflect the change if made in them.
4. Further statistical analysis are difficult to make.

2.3.3.2 Uses of Range : However, range can be used in the following situations :

1. When a quick and crude estimate of variability is desired.
2. When data are too scant or scattered and more precise measure of variability is not warranted.
3. When the knowledge of only the extreme scores or of total spread is required.

4. When the phenomenon is prone to wide fluctuations such as range of fluctuating temperature of a patient, the daily fluctuating values of stock and the annual range of a temperature values for a geographical region.

5. When case of computation is an important consideration.

2.3.4 THE MEAN DEVIATION :

The mean deviation is the average distance between the mean and the score in the distribution. It is the arithmetic mean of all the deviation taken from the mean of the distribution. Scores larger than the mean will have positive or plus signs, and those smaller than the mean, negative or minus signs. Scores which coincide with the mean will have zero deviation. Algebraically, deviation can be defined x = X - M (deviation of a score from the mean).

Where X = original score: and M = arithmetic mean.

The sum of the deviations (with algebraic signs) from the arithmetic mean is always zero. Hence, their average will also be zero and thus, useless for measuring and describing dispersion. Hence, statisticians decided to disregard the algebraic signs and the direction of deviation. Only the sizes of the deviations are taken into account. The formula for the calculations of average deviations is :--

$$AD = \frac{\Sigma|x|}{N} \quad \text{(the average deviation)}$$

2.3.5 COMPUTATION OF MEAN DEVIATIONS

2.3.5.1 Ungrouped Data : Where Σ= sum of ; |x| = absolute value of deviation; N= Total number of scores of observations. Two vertical lines (called bars) to right and left of x indicate that signs + or - have been ignored.

Table-2.3.2 Calculation of Average Deviation (AD) from Set 2 of Table 1

Persons	**Scores X**	**Deviation with Sign X = X-M**	**Deviation without sign X**
1.	13	13-10 = +3	3
2.	12	12-10 = +2	2
3.	11	11-10 = +1	1
4.	10	10-10 = 0	0

5	9	9-10 = - 1	1
6.	8	8-10 = -2	2
7.	7	7-10 = -3	3
	70	zero	12

$$M = \frac{70}{7} = 10$$

$$\sum |x| = 12$$

N = 7

$$AD = \frac{12}{7} = 1.71$$

Interpretation : the result shows that the score deviation on the average is 1.71 points from mean

2.3.5.2 Grouped Data : In this calculation, in column 1 we write C.I in column 2 we write the corresponding frequencies, in column 3, we write the mid point of the C.I's in column 4 we write the product frequencies and mid points denoted by fx, in column 5, we write the absolute deviation of mid-points of C.I from the mean which is devoted by (d) and in column 6, we write the product of absolute deviations and frequencies, devoted by |fd|. As shown in the following table.

Table 2.3.3 Computation of Mean Deviation

		Mid-points		**Absolute deviation**	
C.I.	f	x	f.x	|d|	|fd|
45-49	2	47	94	15	30
40-44	3	42	126	10	30
35-39	5	37	185	5	25
30-34	9	32	288	0	0
25-29	6	27	162	5	30
20-24	4	22	88	10	40
15-19	1	17	17	15	15
	N=30		Σ fx=960		Σ|fd|=170

$$\text{Mean} = \frac{\Sigma fx}{N} = \frac{960}{30} = 32$$

$$\text{Mean Deviation} = \frac{\Sigma\ |fd|}{N} = \frac{170}{30} = 5.667$$

2.3.5.3 Merits of Mean Deviation :

1. Mean deviation is the simplest measure of dispersion that takes into account all the values in a given distribution.
2. It is easily comprehensible even by a person not well versed in statistics.
3. It is not very much affected by the value of extreme items.
4. It is the average of the deviations of individual scores from the mean.

2.3.5.4 Limitations :

1. Mean deviation ignores the algebraic signs of the deviations and as such it is not capable of further mathematical treatment. So it is used only as a descriptive measure of variability.
2. In fact, M.D. is not in common use. It is rarely used in modern statistics and generally dispersion is studied by standard deviation.

SELF-EVALUATION QUESTIONS :

1. Define range and its limitation.

   ____

   ____

2. Find the mean deviation :

   52, 50, 56, 68, 65, 62, 57, 70

   ____

   ____

3. Find Mean Deviation :

x = 5	10	15	20	25	30
f = 2	7	10	15	11	5

   ____

   ____

2.3.6 THE VARIANCE AND STANDARD DEVIATION

A more stable index that reflects the degree of variability in a group of scores is the Variance and Standard Deviation. In the previous section, it was shown that in any frequency distribution, the mean deviation from the mean must be zero. Hence, the device to get around the difficulty is to take the square of each deviation from the mean, and then to find the average of these squared deviation:

$$\text{Variance } (\sigma^2) = \frac{\Sigma (X - M)^2}{N} = \frac{\Sigma x^2}{N} \qquad (4)$$

Here the symbols are Σ = Summation; X = Any raw score;

Mx = Mean of X scores, N = number of cases, x^2 = Square of deviations from the mean:

σ = Greek letter (small sigma)

In a grouped distribution, for each interval, deviation of the midpoint from the mean in squared and multiplied by the frequency for that interval. When this has been done for each interval, the average of these products is the variance. The formula thus becomes:

$$\text{Variance } (\sigma^2) = \frac{\Sigma f x^2}{N} \qquad (5)$$

Where f stands for the frequency in each class interval; other symbols, as above, Standard Deviation is derived from the variance by taking the square root of the latter. The formula for the calculation of the standard deviation is thus as follows :

$$\text{SD or } = \sigma = \sqrt{\text{Variance}} = \sqrt{\frac{\Sigma x^2}{N}} \qquad (6)$$

σ is pronounced as sigma and is used to denote (SD).

Although variance is an adequate way of describing the degree of variability in a distribution, yet it has one drawback. The variance is a quantity in squared units of measurements. For example, if measurements are taken in inches, then the mean is some number of inches, and a deviation from the mean is a difference in inches. However, the square of deviation is a square inch unit, and thus, the variance, being a mean squared of deviation, must also be in square inches, thus the problem of obtaining an index of variability on original units, arises. This has been taken care of by further calculating the square on the linear measure. It gives us the root mean squared deviation or the standard deviation. Hence, Standard Deviation is the square root of the mean squared deviation and is an index of variability in the original units.

In the calculation of SD, deviations are always taken from the mean, never from the median or mode. The value of S.D. is always positive.

Standard Deviation has been termed so because it provides a standard unit for measuring distance of various scores from their mean.

2.3.7 METHOD OF CALCULATING VARIANCE AND STANDARD DEVIATION

2.3.7.1 For ungrouped Data : The conceptual definition of the variance and the SD are based on the formula which incorporates the deviation score method as shown in formula 1.3 in this lesson. However, to avoid inconvenience of working with fractional values, and when a calculating machine is available, the following formulas which are mathematically equivalent to them are also available. In this solved example, the use of both types of formulas has been made.

Table - 2.3.4 : Calculation of SD from ungrouped Scores.

Score X	Deviation from Mean (X-M)	Squared deviation x^2	Score X	Squared Score X^2
10	2	4	10	100
7	-1	1	7	49
9	+1	1	9	81
6	-2	4	6	36
8	0	0	8	64
$\Sigma x = 40$		$\Sigma x^2 = 10$	$\Sigma x = 40$	$\Sigma x^2 = 330$

$M = 40/5 = 8$

$N = 5$

$$S.D = \sqrt{\Sigma \frac{x^2}{N}}$$

$$S.D = \sqrt{\frac{10}{5}} = \sqrt{2}$$

Substituting the values

$$\sigma = \sqrt{2} = 1.414$$

Formula :

$$S.D = \sqrt{\frac{\Sigma X^2}{N} - \left[\frac{\Sigma X^2}{N}\right]}$$

$$\sigma = \sqrt{\frac{330}{5} - \left[\frac{40}{5}\right]^2}$$

$$= \sqrt{\frac{300}{5} - (8)^2}$$

$$= \sqrt{\frac{300}{5} - 64}$$

$$= \sqrt{\frac{300 - 320}{5}}$$

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

$$= \sqrt{2}$$

$\sigma = 2 = 1.414$

2.3.7.2 Steps in the computation of SD :

(a) Deviation Score Method :

(i) Calculate mean

(ii) Calculate Deviation of each score from the mean

(iii) Square each deviation.

(iv) Sum up the Squared deviation to obtain σ.

(v) Substitute the values in the formula and solve.

(b) Raw Score Method :

(i) List up the scores X.

(ii) Sum up the Scores to obtain ΣX.

(iii) Square each score.

(iv) Square up the squared score

(v) Substitute the values in the formula and solve.

Interpretation : The scores on the average vary 1.414 units from their mean.

2.3.7.3 Calculation of SD from Grouped Data : When scores are arranged in the form of score bands or class intervals, and frequencies are shown against each class interval, calculation of SD can be undertaken by using a long method. Both of these methods are demonstrated below with the help of solved examples.

2.3.7.4 Steps in the Computation of SD by Long and Short Method

Long Method :

Table - 2.3.5 : Calculation of SD by Long Method

(Using Real Mean)

(1)	(2)	(3)	(4)	(5)	(6)
Class Interval	Mid Point X	Frequency (f)	Deviation of X from Mean (x)	fx	fx^2
45-49	47	2	24.6(47.0-22.4)	49.2	1210.32
40-44	42	3	19.6 (42.0-22.4)	58.8	1152.48

35-39	37	2	14.6(37.0-22.4)	29.2	426.32
30-34	32	6	9.6	57.6	552.96
25-29	27	8	4.6	36.8	169.28
20-24	22	8	-.40	-3.2	1.28
15-19	17	7	-5.40	-37.8	204.12
10-14	12	5	-10.40	-52.00	541.84
5-9	7	9	-15.40	-138.6	2134.44
Mean =	22.40	N=50			Σfx^2=6393.04

$$= \sqrt{\Sigma fx^2 / N} = \sqrt{\frac{6393.04}{50}} = \sqrt{127.86} = 11.30$$

Steps in the Computation of SD by Long Method :

(i) Write down class intervals and frequencies as shown in Cols. (1 and 3)

(ii) Find out the mid-point of each class interval

$$\text{Midpoint} = \frac{\text{Upper Limit} + \text{Lower Limit}}{2}$$

(iii) Calculate Mean by using any method described in a previous chapter

(iv) Obtain deviation from Mean, (Mid point-Mean) as in Col (4).

(v) Multiply f and x, Col (3) x Col (4) to obtain (fx).

(vi) Multiply fx and x i.e. col (4) and (5) to obtain fx^2 as in col. (6)

(vii) Sum up col (6) to obtain Σfx^2.

(viii) Substitute the values in the formula and solve.

By Short Method : When large values of scores are involved, it is better to use the short method for calculation of SD. In this method like the calculation of Mean by the Assumed Mean Method deviations are taken from the assumed mean. The detailed procedure is given below :

Table 2.3.6

Calculation of SD by Short Method (Deviations taken from Assumed Mean)

(1)	(2)	(3)	(4)	(5)	(6)
Class Interval	Mid Point X	Frequency (f)	Deviation of X from AM in units of c.i (x)	(fx)	(fx^2)
45-49	47	2	5	10	50
40-44	42	3	4	12	48
35-39	37	2	3	6 +48	18
30-34	32	6	2	12	24
25-29	27	8	1	8	8
20-24	22	8	0	0	0
15-19	17	7	-1	-7	7
10-14	12	5	-2	-10 - 44	20
5-9	7	9	-3	-27	81
		N = 50		$\Sigma fx=4$	$\Sigma fx^1 = 256$

$$\sigma = i\left(\sqrt{\frac{\Sigma x^2}{N} - c^2}\right)$$

In which, i stands for the size of class interval, and c for correction.

Hero : I = 5, fx^2 = 256; N= 50; and C^2 =

$$\sigma = 5\left(\sqrt{\frac{256}{50} - \frac{4}{50}}\right)$$

$$\sigma = 5\left(\sqrt{5.12 - .0064}\right)$$

$\sigma = 5(\sqrt{5.1136})$

σ = 2.26 x 5

σ = 11.31

Computational Steps :

(i) Arrange the scores and fs in columns f (3).

(ii) Find out the mid-point of all the intervals and write in Col. (2).

(iii) Take a mid-point as an assumed mean. This point should be close to the middle of the distribution and as far as possible should have the largest f.

(iv) Deviations (x) are taken from the assumed mean (here 22) in units of class interval.

It can be a mechanical process. Assign 0 to the class interval in which the assumed mean lies go on assigning + 1, +2, +3, etc., to the class intervals above the mean and -1, -2, -3, etc., to those below the mean (Col.)

(v) Multiply cols. (3) and (4) to obtain fx and sum up to obtain Σfx.

(vi) Multiply Col (4) and (2) to obtain fx^2 and sum up to obtain Σfx^2.

Find out the value of c which is $\frac{\Sigma fx}{N}$

(vii) Substitute these values in formula and solve.

Note: When sample SD is to be used as an estimate of the population SD, the denominator of the formulas will be N-1, instead of N, as this form is considered as an unbiased estimate.

2.3.7.5 Properties and Uses of Variance and SD as Measures of Variability:

(i) To variance is proportional to the average squared deviation of each score from every other score. Hence, it reflects the variability of the score;

(ii) Since all deviations are squared, the variance will always be positive. The SD is the positive square-root of variance and hence will always be positive.

(iii) If there is no variability among the scores as all the scores in the distribution will be identifical, the values of SD will be Zero. As variability in the scores increases the variability also increases.

(iv) Variance and SD have more sensitivity to variability in a group of scores and are less variable in themselves.

(v) The variance and SD are frequently used in other statistical analysis and manipulation and hence are more important than the other measures of variability.

(vi) The variance can be partitioned into different parts attributed to different courses and hence finds its use in analysis of multivariate factorial designs.

(vii) Variance and SD should be used when extreme deviations are likely to exercise a proportionally greater effect upon the variability.

SELF-EVALUATION QUESTIONS :

1. Define the Standard Deviation.

   ____

   ____

   ____

2. Compute of S.D. (Direct method or Shortcut method)

   x = 12, 15, 10, 8, 11, 13, 18, 10, 14, 9

   ____

   ____

   ____

3. Find the S.D. for following frequency :

class interval :	10-19,	20-29,	30-39,	40-49,	50-59,	60-69,	70-79
Frequency :	2	3	5	10	12	6	4

   ____

   ____

   ____

2.3.8 THE SEMI-INTER QUARTILE RANGE OR Q :

The semi inter-quartile range which is also known as quartile deviation can be defined as half of the difference between the 75th percentile and the 25th percentile. Hence, it is one half the scale distance between the 75th and the 25th percentiles is in a frequency distribution : The 25 percentiles in Q, or the first quartile on the score scale. The 25th percentile is Q^2 or third quartile on the second scale. Hence, the formula for the calculation of Q is.

$$Q = \frac{Q_3 - Q_1}{2} \text{ or } \frac{P_{75} - P_{25}}{2}$$

Hence to find Q, it is essential to calculate the values of Q_3 (P_{75}) and Q_1 (P_{25}). Their calculation follows the same procedure as the calculation of median, as explained in previous chapter. These formulae are:

$$Q1\ (P_{25}) = l.l. + \frac{(N/4 - Cumf)}{fq}\ xi$$

$$Q3\ (P_{25}) = l.l. + \frac{(3N/4 - Cumf)}{fq}\ xi$$

In which

l.l = the exact lower limit of interval.

i = the size of the class interval.

Cum f = Cumulative f below the interval which contains the quartile.

Fq = that f on the interval which contains the quartile.

Table 2.3.7 Calculate Q_1, Q_3 and Quartile Deviation

(1) Class interval (C.I) (x)	(2) Frequency (f)	(3) Cumulative (Cum f)
45-49	2	50
40-44	3	48
35-39	2	45
30-34	6	43 Q_3 lies in this C.I

25-29	8	37
15-19	7	29
10-14	5	14 Q_1 lies in this C.I
5-9	9	7
	N=50	

Calculation of $Q_1 = \frac{N}{4} = \frac{50}{4} = 12.5$

Here, *l.l.* 9.5; cum f = Fq = 9 = f = 5, I = 5

Substituting these values in formula (13) we have

$$Q_1 = 9.5 + \frac{(50/4-9)}{5} \times 5 = 9.5 + 3.5 = 13.00$$

Calculation of $Q_3 = \frac{3N}{4} = \frac{3(50)}{4} = 12.5x3 = 37.5$

$$\text{Here, } Q_1 = 29.50 + \frac{(3 \times 50/\ 4 - 37)}{6} \times 5$$

$$= 29.50 + .42 = 29.92$$

Calculation of Q

Substituting the values of Q_1 and Q_3 in formula (12), we have

$$Q = \frac{29.92 - 13.00}{2} = \frac{16.92}{2} = 8.46$$

2.3.8.1 Properties and Uses of Q :

(i) In a distribution which is symmetrical around the mean when it is normal - Q marks off the 25% cases just above, and the 25% of the cases just below the median.

(ii) It is a measure of the variability of the middle 50% cases and ignores the 25% cases in each of the two tails.

(iii) It should be used when a measure of dispersion of the concentration of 50% of the cases round the median is required.

(iv) It should be used when the measure of central tendency is to be used as the median.

(v) It is a better measure when scores are scattered or extreme score which would influence the SD disproportionately.

(vi) Q is known as probable Error (PE) in a normal distribution.

2.3.8.2 Limitations of Q : The quartile deviation should be used along median, as both the measures are based on the same assumption. The quartile deviation does not indicate variation of total distribution. It represents the variability of middle 50 percent distribution. It is not a measure of variation of total distribution. It does not consider extreme 25 per cent top and 25 per cent bottom scores.

2.3.9 SUMMARY :

Some important and more popular measure of variability of dispersion have been presented in this lesson. The lesson show that "spread' or "Scatter" of the separate score around their central tendency : one of the these should be reported along with the relevant measures of central tendency of provide a better description of the distribution. The meaning of dispersion is scatteredness. Main dispersion are the Range, mean deviation, variance, standard deviation and semi-interquartile range. Range is the distance between the highest score and lowest score in the distribution. The mean deviation is the average distance between the mean and the score in the distribution. It is easily comprehensible much affected and simplest measure of dispersion.

2.3.10 SUGGESTED QUESTIONS :

1. Calculate Variance and SD from the following distribution of scores (use long method).

Class Interval	f
190-199	2
180-189	4
170-179	15

160-169	11 (Answer : Var. -458,75 SD =21.42).
150-159	38
140-149	8
130-139	6
120-129	2
110-119	7
100-109	7
	N=100

2. Calculate Variance and SD from the distribution in Q No. 1 above, by using short method and compare the results.

2. Calculate AD and Q from the distribution in No. 1 above (Answer AD = 15.92; Q = 11.305)

3. Define the following as precisely as you can SD = AD = Variance; Q : and Range.

4. Describe the essential properties and uses of Variance and Range.

2.3.11 SUGGESTED BOOKS

Edward, Allen I — *Statistical Method of Behavioural Sciences*

Garret, H.E. — *Statistics in Psychology and Education*

Guilford,J.P. — *Fundamental Statistics in Psychology and Education*

L. William, Hays — *Statistics for the Social Science.*

LESSON NO. 2.4 **AUTHOR : DR. S.S. AGGARWAL**

NORMAL PROBABILITY CURVE AND ITS IMPLICATIONS

Structure of the Lesson

2.4.1 Objectives :

After reading this chapter students will be able to learn the :

(i) Concept of Probability and meaning of normal distribution.
(ii) Properties of Normal Probability Curve.
(iii) Various applications of Normal Probability Curve.
(iv) Applications of knowledge of Normal Probability Curve in solving various practical problems.

2.4.2 Introduction :

We have already learnt, how calculated the measures of central tendency to describe the central value of the frequency distribution. We have also found the measures of variability to indicate its variations. All these descriptions have gone a long way in providing information about a set of scores. Sometime we need procedures for describing and individual's, position in a group or the cutting points to categories the group according to the level of ability or the nature of test paper which a teacher has used to assess the learning outcomes of the students.

In our daily life Probability or chance is very commonly used terms sometimes we use to say probably it may rain tommorrow. Probably may come for taking his class today. All these terms probability & possibility the conveys the same meaning. But in statistics probability has certain special connotation unlike in Layman's view. The theory of probability has been developed in 17th century. It has got its origin from games, tossing coins etc. In 1954 Antoine Gornbad had taken an initiation and an interest for this area. After him many authors in statistics had tried to remodel to idea given by the former. The probability has become one of the basic tools of statistics same time statistical analysis becomes paralysel without the theorem of probability.

2.4.3 Normal Probability Curve :

In statistics, normal distribution is considered as one of the most important distributions, particularly in the area of test construction. Let us consider the following smoothed polygon (fig. 1) and the histogram (fig. 2).

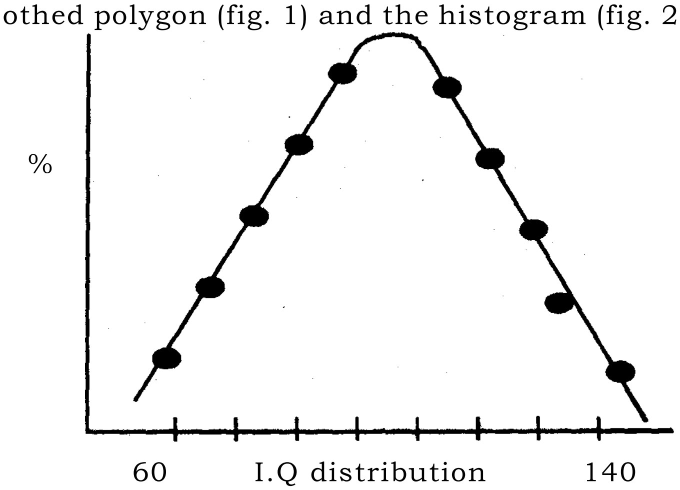


Fig. 1

The best fitting normal curve

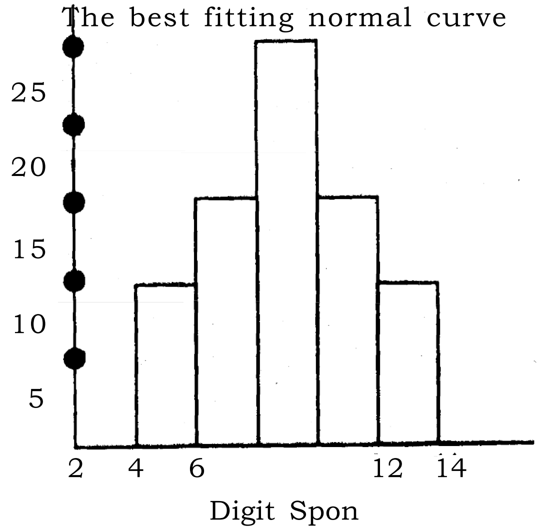
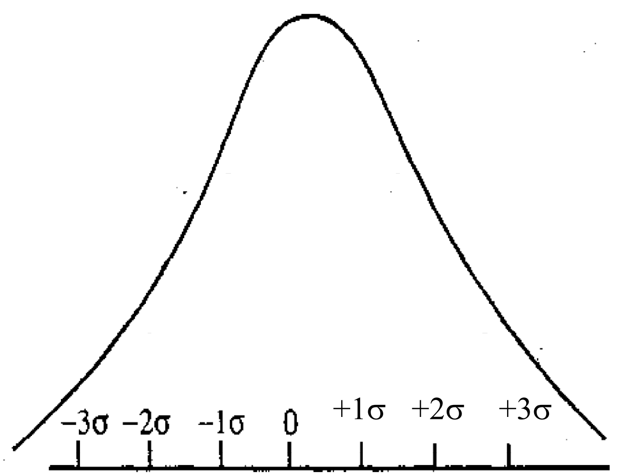


Fig. 2

Memory spanfor digits

These two graphs have the same general form; the measures are concentrated closely around the centre and taper off from the central high point to left and right. There are relatively few measures at the 'low score' end of the scale, and increasing number up to a maximum at the middle of position and a progressive falling-off towards the 'high score' of the scale. If we divide the area under each curve (the area between the curve and the X-axis) by a line drawn perpendicularly through the central high point to the baseline, the two parts thus formed will be similar in shape and very nearly equal in area. Each figure exhibits almost perfect bilateral symmetry. This perfectly symmetrical curve of frequency distribution to which the first two graphs (fig. 1 and fig. 2) approximate is shown in figure 3.



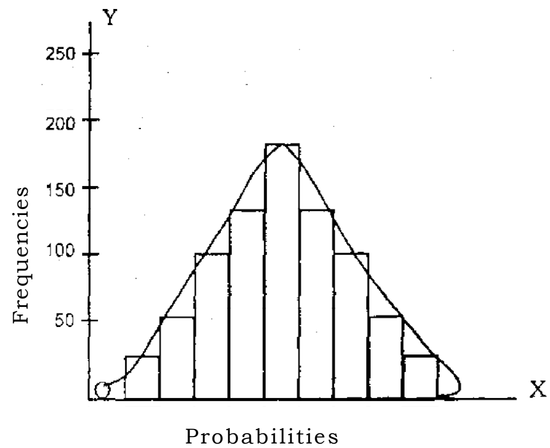
Normal Probability curve Figure . 3

This bell-shaped figure is called the normal probability curve (NPC) or simply normal curve, and is of great value in mental measurement. The curve is also called the Chance curve, Curve of Error, Probability curve or Gaussian Curve after the name of its discoverer Gauss. It has been seen that measurements of many natural phenomena and of many mental and social traits such as weight, height, wages and output of workers, intelligence, perception span, reaction time, educational test scores in spelling, reading, arithmetic etc. tend to be distributed symmetrically around their means in proportion which approximate those of the normal probability distribution. It has also been seen that many distributions are similar to those obtained by tossing coins or throwing dice because the former, like the later, are actual probability distributions.

If we toss ten coins simultaneously, the expression can be written $(H+T)^{10}$ where H stands for the probability of a head, T for the probability of a tail and 10 for the number of coins tossed $(H+T)^{10} = H^{10} + 10H^{9}T + 45H^{8}T^{2} + 120H^{7}T^{3} + 210H^{6}T^{4}$

$+ 252H^5T^5 + 210H^4T^6 + 120H^3T^7 + 45H^2T^8 + 10HT^9 + T^{10}$

If this expansion is graphically represented through a histogram and a polygon, we will have the following figure (Fig. 4).



Probabilities Curve obtained from expansion of $(H+T)^{10}$ (Fig.4)

The eleven terms of expansions have been laid off at equal distances along the X-axis and chance of occurrence of each combination of H's and T's are plotted as frequencies or the Y-axis. The result is a symmetrical frequency polygon with the greatest concentration in the centre and the score falling away by corresponding decrements above and below the central high point.

Now when n in the expression $(H+T)^n$ becomes infinite, the polygon would exhibit a perfect occurrence like that of this curve in figure 3. This ideal polygon represents the frequency of occurrence of various combinations of a very large number of equal, similar and independent factors (coins) when the probability of the appearance (H) or non-appearance (T) of each factor is the same. The normal distribution is not an actual distribution of test scores, but is instead a mathematical model.

2.4.3.1 Assumptions of Normal Distribution :

The basic assumption of normal distribution is that the variables of behavioural sciences, bio-sciences of a large sample or representative sample are commonly normally distributed.

1. The area covered by normal distribution is assumed an unit.
2. It assumes, mean is zero and standard deviation is one. The mean is three times of SD of normal curve.
3. The range of the curve is 6σ i.e. from -3σ to $+3\sigma$.
4. It indicates standard scores are in the form of σ scores on X-axis.

2.4.3.2 Properties of Normal Distribution :

1. The equation of the N. P. C. is

$$Y = \frac{1}{\sigma\sqrt{2\pi}} x \frac{-x^2}{2\sigma^2}$$

Where x = scores (expressed as deviations from the mean) i.e. x= (x-mean) laid along the X-axis.

y = the height of the curve above the X-axis.

σ = S.D. of the distribution.

∏ = 22/7 = 3.1416 (the ratio of the circumference of a circle to its diameter)

e = 2.7183 (base of Napierian system of logarithms)

2. The curve is symmetrical. The mean, median and mode coincide. The characteristic of symmetry about the ordinate at the central point of the curve implies that the size, shape and slope of the curve on one side of curve are identical to those on the other side. The value of mean, median and mode computed for a distribution following the curve are always equal.
3. The maximum ordinate of the curve occurs at the mean.
4. The curve is asymptotic. It approaches but does not meet the horizontal axis and extends from minus infinity (-∞) to plus infinity (+∞).
5. The points of infection of the curve occurs at point = 1 standard deviations unit above and below the mean. Thus, the curve changes from convex to concave in relation to the horizontal axis at these points.
6. 68.26% of the area of the curve falls within the limits ± 1 standard deviation units from the mean.
   If total area of the curve is 100, then 68.26 cases fall between the two ordinates at ± 1σ.
7. In the normal curve, the limits M ± 1.96σ include 95% and the limits M ±2.58σ include 99% of the total area of the curve i.e. 5% or 1% of the area respectively fall beyond these limits.

2.4.4 Applications of the Normal Probability Curve :

A number of problems can be solved if we assume that our obtained distributions can be treated as normal, or approximately normal. Table A at the end is made available for reference.

2.4.4.1 Determination of percentages of cases between the given standard scores from the mean.

Percentage of cases can be known from Table A. The figures given in Table A against standard scores Z are the fractions of the total area under the curve. While converting the given value into percentage one will have to multiply it by

100.

Example : Given a distribution of scores with a mean of 16, and a SD σ of 4. Assuming normality find

(a) What percentage of the cases fall between scores 12 and 20 ?

(b) What percentage of the cases lie above score 22 ?

(c) Below score 10 ?

Solution :

(a) Given Mean = 16 and σ = 4

for score 12 the sigma distance $z_1 = \dfrac{X - M}{\sigma}$

$$= \frac{12 - 16}{4} = \frac{-4}{4} = -1$$

For score 20

$$z_2 = \frac{(20-16)}{4} = \frac{+4}{4} = +1$$

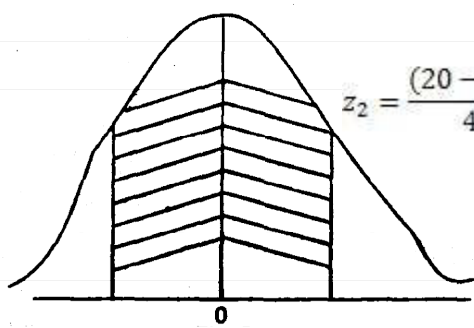


Fig.5

We have to find the area between z_1 and z_2. This is equal to area between z = 0 and z = 1, area between z = 0 and z = +1

=0.3413+0.3413

= 0.6826

Hence percentage of cases that fall between 12 and 20 = 68.26%

(See Table A)

(b) In this question, the score 18 is not included. The upper limit of score of 18 is 18.5 which in standard units is :

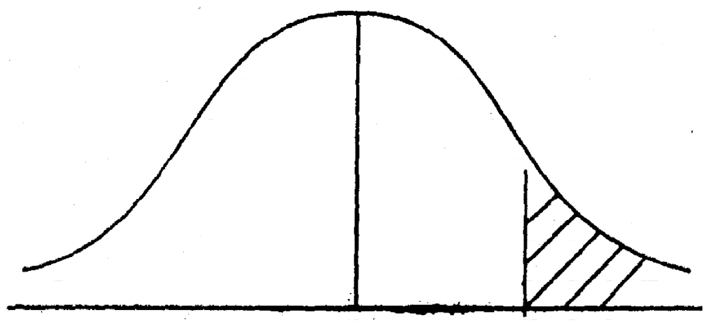
$$z = \frac{22.5 - 16}{4} = \frac{6.5}{4} = 1.625$$

Now the required number of cases which lie above score 22 = area to the right of 1.625. From table A, we find that 44.79% of cases fall between mean and 1.625.

Therefore, .0521 lie above the upper limit of 22, (50 - 44.79)

or percentage of cases that fall above 22 = 5.21%

(c) The lower limit of score of 10 is 9.5 which in standard unit is



(Fig. 6)

$$z = \frac{9.5 - 16}{4} = -1.625$$

Now the required number of cases below score 10 = half this curve - area between mean area 1.625

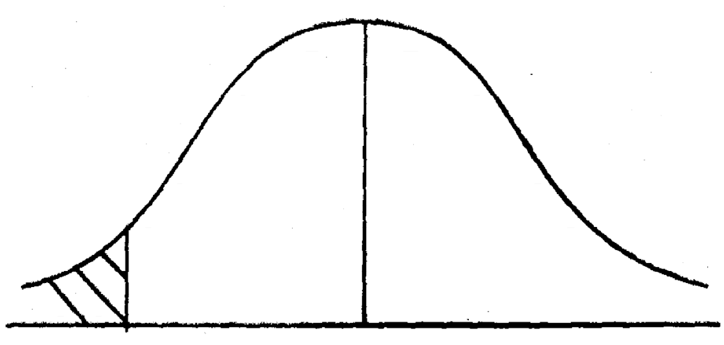
= .5 - .4479

= 0.0521

i.e. 5.21%

= Hence percentage of cases that fall below 10

= 5.21%



(Fig. 7)

2.4.4.2 To find the limits in any normal distribution which include

a given percentage of cases

Example : Given a distribution of scores with a mean of 20 and σ of 4. If we assume normality what limits will include the middle 75% of the cases.

Solution : Since a normal distribution is symmetrical about the ordinate at z = 0, the middle 75% of the cases are lying both sides of the mean, half and half i.e. 3750 on each side. Now from Table A (3749 quite close to 3750) has z = 1.15. Therefore 75% of the cases would fall between mean and ± 1.15σ.

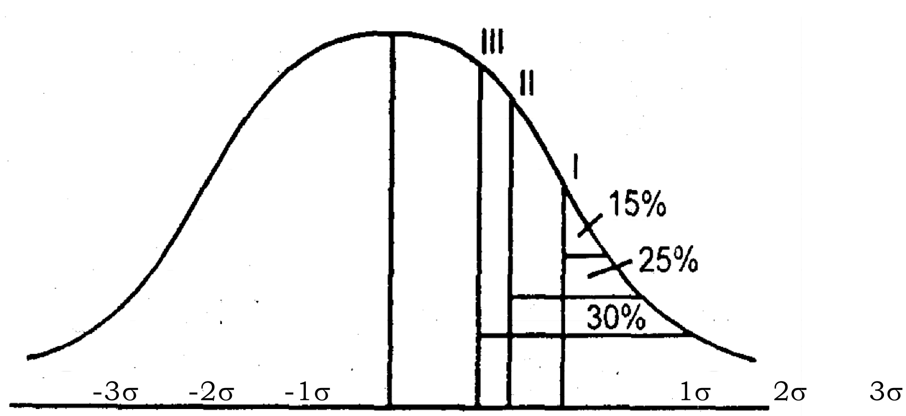
The limits would be 20 ± 4 x 1.15 = 20 ±4.6 = 24.6 and 15.4

2.4.4.3 To determine the relative difficulty of test questions, problems and other test items :

Problem A is solved by 15% of large unselected group, a second problem (B) is solved by 25 of the same and a third problem (C) solved 30%. If we assume the capacity measured by the test problem to be distributed normally, what is the relative difficulty of problem I, II and III.

Solution :

Our problem here is to find out the difficulty of each problem I, II and III. So that 15%, 25% and 30% of the entire group lies above and 85%, 75% and 70% below this cut point. For problem 15% lie above and 85% lie below the given point. The highest 15% in normally distributed group has (50-15) = 35% of cases between its lower limits and mean. From table A, we find that 35.00% (or .3508 out of 1) i.e. 35% of a normal distribution fall between the mean and 1.04σ. Hence problem I belongs to a point on the base line of the curve towards positive side, a distance of + 1.04σ from the mean and accordingly, +1.04 $\overline{0}$ may be as difficulty value of this problem.



(Fig. 8)

Calculate similarly, difficulty values of Second and Third problems and obtain the following :

Problem	Passed by	% between mean & the cut point	σValue	Difference
1	15%	35%	+1.04	0.37
2	25%	25%	+0.67	
3	30%	20%	+0.52	0.15

The difference in difficulty between questions 2 and 3 is 0.15 which is less than 1/2 or the difference in difficulty between questions I and II.

2.4.4.4 To compare two distributions in terms of overlapping :

Example : Given the distribution of the scores made on a memory test by 300 boys and 250 girls.

The boy's means score is 20.49 with σ = 3.63

The girl's means score is 21.49 with σ = 5.12

The medians are : boys 20.41, girls 21.66

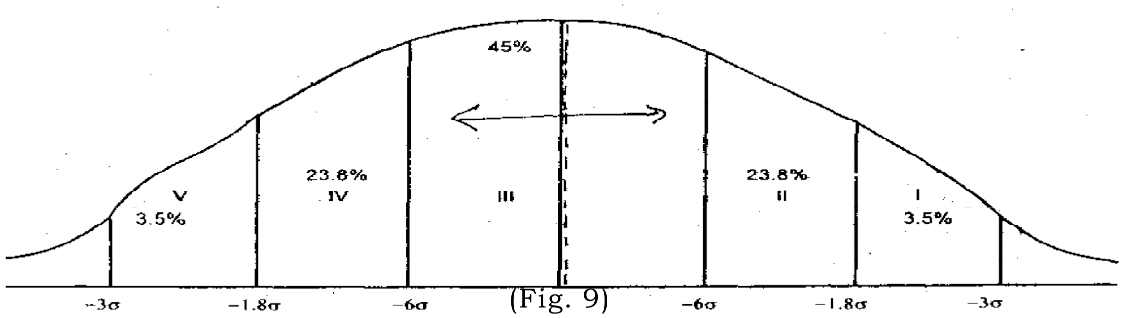
What percentage of girls exceed the median of the boys distribution?

Solution :

On the assumption that these distributions are normal, we may solve the problem with the help of Table A. The boys median is 20.41 - 21.49 = -1.08 score unit below the girls mean. Dividing 1.08 by 5.12 (The σ the girls distribution), we find that the boys median is. 21 σ below the mean of the girls distribution. Table A shows that 8.32% of the normal distribution lies between the mean and -.21σ hence 58% of the girls (50% + 8%) exceed the boys median.

2.4.4.5 To classify a given group into sub-groups according to capacity when the trait is normally distributed.

Example : Suppose, we have administered an entrance examination to 200 college students. We wish to classify our group into five sub-groups I, II, III, IV and V according to ability. The range of ability is to be equal in each subgroup. On assumption that the trait measured by our examination is normally distributed, how many students should be placed in group I, II, III, IV and V?



(Fig. 9)

Solution :

First, represent the position of the given sub-group diagrammatically on a normal curve as shown above. Consider the baseline or curve extending from -3σ to +3σ i.e. over a range of 6.

Dividing this range by 5 (the number of sub-groups, we get = 1.2 σ as the base line extent be allotted to each group. These five intervals may be laid off on the base line as shown in figure perpendiculars erected to demarcate the various subgroups.

Group I covers the upper 1.2 σ

Group II the next 1.2 σ

Group III lies 0.6 to the right, and 0.6 to the left of this mean.

Group IV and V occupy the same relative position in the lower half of curve than II and occupy in the upper half.

To find what percentage of the whole group belongs in I, we must find what percentage of normal distribution lies between 3 σ (upper limit of the I group) and 1.8σ (lower limit of group I). From Table A, 4986 i.e. 49.86% of a normal distribution is found to lie between the mean 3 and 46.41% between the mean and 1.8. Hence (49.86%-46.41% = 3.5%) of the total area under the normal curve lies between *3*σ and 1.8σ and accordingly group I comprises 3.5% of the whole group.

The percentage of the other groups are calculated in the same way. Thus, 46.41% of the normal distribution falls between the mean and 1.8 (upper limit of group B), and 22.57% OR 23.84% or our distribution belongs to sub-group II.

Group III lies from 0.6 above -0.6 below the mean. Between the mean and .6 is 22.57% of the normal distribution and the same percent lies between the mean and -6, group III, therefore, includes 45.14% (22.57 x 2) of the distribution.

Finally, group IV, which lies between -0.6σ and -1.8σ contain exactly the same percentage of the distribution of sub-group II, and group V, which lies between -1.8σ and -3σ contains the same percent of the whole distribution as group I.

The percentage and number of men in each group are given in this table below.

Table

	I	II	III	IV	V
Parentage of total individuals in each group	3.5	23.8	45	23.8	3.5
Number in each group 200 men in all)	7	48	90	48	7

Uses of Normal Probability Curve

1. In research work.
2. For measurement and tet construction.
3. In statistical treatment.
4. In administration.
5. In developing norms.

6. To determine difficulty values of test items.
7. To compare two distributions.
8. To determine levels of significance.

2.4.5 Statistics in Psychology and Education :

Table A : Fractional parts of the total area (taken as 10.000) under the normal probability curve, corresponding to distance on the basseline between the mean and successive point laid off from the mean in units of standard deviation.

Example : between the mean and a point 1.38 $\overline{0}$ $\left(\frac{x}{0} = 1038\right)$ are found 41.02% of the entire are under the curve.

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0160	0100	0230	0270	0310	0359
0.1	0398	0438	0478	0517	0557	0596	0638	0675	0714	0453
0.2	0793	0832	0871	0910	0948	0987	1026	1084	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1408	1443	1480	1517
0.4	1554	1501	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1960	1985	2010	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2488	2517	2549
0.7	2580	2611	2642	2678	2704	2734	2764	2794	2823	2852
0.8	2881	2910	2939	2987	2995	3023	3051	3078	3100	3172
0.9	3149	3180	3212	3238	3204	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3609	3621
1.1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	3840	3859	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4082	4099	4115	4131	4147	4182	4177
1.4	4192	4207	4222	4236	4251	4262	4279	4292	4306	4319
1.5	4332	4345	4357	4372	4383	4394	4406	4418	4429	4441
1.6	4452	4403	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4663
1.8	4641	4649	4656	4604	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4758	4761	4767
2.0	4772	4778	4787	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	4864	4887	4890
2.3	4893	4896	4898	4871	4875	4878	4881	4884	4887	4890
2.4	4918	4920	4922	4925	4927	4923	4931	4932	4934	

										4936
2.5	4938	4940	4941	4943	4945	4996	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	4983	4984	4984	4985	4985	4986	4986
3.0	4986.5	4986.9	4987.4	4988.2	4988.6	4988.3	4989.3	4989.7	4990.0	
3.1	4990.1	4999.3	4991.0	4991.3	4991.6	4991.8	4992.1	4992.4	4992.6	4992.9
32	4993.129	4995.160	4996.631	4997.674	4998.409	4999.519	4999.683	4999.966	4999.997132	

2.4.6 Summary :

The normal frequency distribution curve is based upon the law of probability or the probable occurrence of certain events. The 'probability' of a given event is defined as the expected frequency of occurrence of this event among events of a like sort. It may be stated mathematically as a ratio. The probability of an unbiased coin falling heads is 1/2, and probability of dice showing a four-spot is 1/6. These ratios called 'probability ratios' are defined by that fraction, the numerator of which equals the derived outcomers and the denominator of which equals the total possible outcomes. A probability ratio always falls between the limits .00 and 1.00. All possible degrees of likelihood may be expressed by appropriate ratios between these limits. Application of the normal probability curve are : (i) Determination of percentages of cases between the given stand and scores from the mean (ii) To find the limits in any normal distribution which include a given percentage of cases (iii) To determine the relative difficulty of test questions-problems and other test items (iv) To compare the distribution in terms of overlapping and
(v) To classify a given group into sub-groups according to capacity when the trait normally distributed.

2.4.7 SUGGESTED QUESTIONS :

1. (a) Define a NPC? Give its applications.
   (b) Discuss the properties of normal probability curve.
2. A test in Mathematics given to 1500 school leaving pupils, had a mean = 53.0 and a Standard Deviation of 8.0.
   Assuming normality, answer the following :
   (a) How many individuals score above 65?
   (b) How many individuals score between 45 and 75?

(c) What are the scores between which the middle 60% cases lie?

(d) What would be the chances in 100 of selecting an individual at random who would score 60 or higher?

3. In a normal distribution, M = 200, and SD = 25.

(a) What percent of scores lie between 180 and 240?