UNIT 14 ANALYSIS OF QUANTITATIVE DATA (DESCRIPTIVE STATISTICAL MEASURES : SELECTION AND APPLICATION)

Structure

14.1 Introduction

14.2 Objectives

14.3 Types of Data

- 14.3.1 Qualitative Data
- 14.3.2 Quantitative Data

14.4 Graphic Representation of Quantitative Data

- 14.4.1 The Histogram or Column Diagram
- 14.4.2 Frequency Polygon
- 14.4.3 Cumulative Frequency Curve
- 14.4.4 Cumulative Percentage Curve or Ogive

14.5 Descriptive Statistical Measures

- 14.5.1 Measures of Central Tendency or Averages
- 14.5.2 Measures of Variability or Dispersion
- 14.5.3 Normal Probability Curve
- 14.5.4 Measures of Relative Positions
- 14.5.5 Measures of Relationship

14.6 Let Us Sum Up

14.7 Unit-end Activities

14.8 Points for Discussion

14.9 Suggested Readings

14.10 Answers to Check Your Progress

14.1 INTRODUCTION

In the Block 2, you were introduced to various approaches to research: historical, philosophical, descriptive, experimental, etc. You were also explained the techniques of sampling, alongwith the development and use of tools in the collection of data. The data are of two types: Qualitative and Quantitative. Descriptive narrations, observation notes, responses to questions, etc. are qualitative data; whereas quantitative data comprise numerical figures. The quantitative data are either parametric or non-parametric in nature. Various descriptive statistical measures are used in the analysis of parametric and non-parametric data. These measures include: (i) measures of central tendency; (ii) measures of variability; (iii) measures of relative position; and (iv) measures of relationship. Descriptive statistical analysis using these measures limits generalisation to the particular group of individuals observed. No conclusions are extended beyond this group, and any similarity to those outside the group cannot be assumed.

In this unit you will study the nature of qualitative and quantitative data. The techniques which produce data of quantitative nature will also be explained. You will also be introduced to various descriptive statistical measures which are used in the analysis of quantitative data. The computation, selection and application of these descriptive statistical measures will also be explained with the help of illustrations. You will also understand the nature and characteristics of a normal probability curve, and its applications in educational research.

14.2 OBJECTIVES

After studying this unit, you will be able to:

- name and describe the nature of various types of data;
- describe the techniques/tools which generate data of quantitative type;
- describe the procedures of classifying data and their graphical representation;
- name and describe various statistical measures which are used in the analysis of quantitative data;
- define, Mean, Median and Mode as the measures of central tendency; their computation, selection and applications;
- define Range, Average Deviation, Quartile Deviation and Standard Deviation as the measures of variability; their computation, selection and application;
- discuss the nature, characteristics and applications of Normal Probability Curve;
- define Sigma Scores, T-Scores and Percentile Rank as the measures of comparing individuals; their computation, selection and application; and
- define Product moment correlation and Rank difference correlation as the measures of relationship; their computation, selection and application.

14.3 TYPES OF DATA

The data collected from various sources through the use of different tools and techniques generally consists of numerical figures, ratings, narrations, responses to open-ended questions comprising a questionnaire or an interview schedule, quotations, field notes etc. In educational research usually, the use of two types of data are recognised. These are qualitative data and quantitative data.

14.3.1 Qualitative Data

Qualitative data are verbal or other symbolic materials. The detailed descriptions of observed behaviours, people, situations and events, are some examples of qualitative data. These data are gathered by a variety of methods and techniques such as: (i) observation; (ii) interviews; (iii) questionnaires, opinionnaries and inventories; and (iv) recorded data from newspapers and other documents maintained by educational institutions, courts, clinics, government and non-government agencies etc. These data are narrations and indicate what people have said in their own words about their experiences and interactions in natural settings. The data gathered through observation may consist of patterns of action and verbal and non-verbal interaction between members of the group in their natural settings, and are qualitative in nature. The responses to open ended questions in a questionnaire are neither systematic

nor standardized. However, such responses are detailed and comprehensive which provide data of qualitative nature in the form of emotions of respondents, their thoughts and experiences about the phenomena under study.

Further details about the nature of qualitative data, their analysis and interpretation will be explained to you in the units 17 and 18.

14.3.2 Quantitative Data

Quantitative data are obtained by using various tools and tests based on scales of measurement: nominal, ordinal, interval or ratio. In educational research, we generally come across nominal, ordinal or interval scale of measurement. Ratio scale measurements are almost non-existent in the educational studies. The experiences of people are fitted into standard responses to which numerical values are attached. These data are close-ended and hardly provide depth and detail which is possible only in the case of qualitative data.

The quantitative data are either **parametric** or **non-parametric**

Parametric data are measured on interval or ratio scale measurements. Ratio scale occupies the highest level in the measurement, because it permits all the four operations of addition, subtraction, multiplication and division and hence all statistical techniques are permissible with such scales. The measurement scales have all the characteristics of the interval scale, with additional advantage of a true zero point. For example, the zero point on a centimeter scale indicates complete absense of length and height. It may be noted that ratio scales are almost non-existent in psychological and educational measurements except in the area of pscho-physical judgement where we measure the reaction time with the help of customary time unit like second and fraction of the seconds. Since in educational research we would be mostly dealing with interval scale measurement data, some examples about this scale are provided for clarifying the concept. If we actually measure the weight of the three students, A, B and C by using a weighing machine and find their weights as 41, 49 and 53 kilograms respectively, we have a measurement on a scale of equal units. This scale of measurement is called an interval scale. By this measurement we can say that C is 4 kilograms heavier than B and 12 kilograms heavier than A. We can also infer that the difference in weights between C and A is thrice than between C and B. The marks obtained by students in Hindi in a certain class provide data on interval scale.

In interval scale, the difference between consecutive points on the scale are equal over the entire scale but there is no zero point on it. The zero point (point of reference) of the scale is chosen conventionally or arbitrarily. The scores on an intelligence test or an attitude scale are also based on interval scales. They have no real zero point. To illustrate this concept, suppose a student gets "zero" score in a test of mathematics, this does not mean that the student has no knowledge of mathematics.

The operations of addition and subtraction can be performed on interval scales; the statistical techniques based on these operations are permissible. However, since operations of multiplication and division assume the existence of an exact zero point, these operations cannot be used with interval scales.

Nominal as well as ordinal scales of measurement provide non-parametric data. These data are either counted or ranked. Nominal scales are used when a set of objects among two or more categories are to be differentiated on the basis of some similar defined characteristics. Usually symbols or numericals are chosen to represent all objects in a given category. We assign students to such categories as locality (rural and urban), gender (females and males) etc. Nominal scales are non-orderable

and only arithmetical operation applicable to such scales is Counting, the mere enumeration of individuals in a particular category. Statistical analysis based on counting is permissible in this type of measurement.

The measurement based on ranks is ordinal scale measurement. In this scale, objects or individuals are ordered (ranked) on some continuum in a series from lowest to highest according to measured characteristics. The ranking of students in a class for weight, height or academic achievement are the examples of ordinal scale. Suppose we rank three students X, Y and Z in order of their height, X as the tallest, and assign the number 1, 2, 3 respectively as their ranks.

In this way we have information about the serial arrangement. We cannot infer any information as how much X is taller than Y or Z, even though the three numbers assigned to them are equally spaced on the scales of measurement. The statistical analysis based rank is permissible in this measurement.

14.4 GRAPHIC REPRESENTATION OF QUANTITATIVE DATA

Graphic representation of quantitative data is helpful in reading and interpreting data. The following four methods of graphic representation are in general use.

- Histogram or Column Diagram
- Frequency Polygon
- Cumulative Frequency Curve
- Cumulative Percentage Curve or Ogive

14.4.1 The Histogram or Column Diagram

Consider the following distribution of scores. The distribution has been obtained on the basis of the scores by a group of 40 students in a test of chemistry in which the maximum score achieved by a student was 65 and minimum score was 10.

Scores (Class Interval)	Exact Units of Class Intervals	f (Frequency)
60 - 69	59.5 - 69.5	1
50 - 59	49.5 - 59.5	4
40 - 49	39.5 - 49.5	10
30 - 39	29.5 - 39.5	15
20 - 29	19.5 - 29.5	8
10 - 19	9.5 - 19.5	2
		N = 40

Let us construct a histogram or column diagram representing the above distribution. Histogram is a graph in which the exact class intervals are represented along the horizontal axis called x-axis and their corresponding frequencies are represented by areas in the form of rectangular basis drawn on the intervals as shown in the figure (Fig. 14.1).

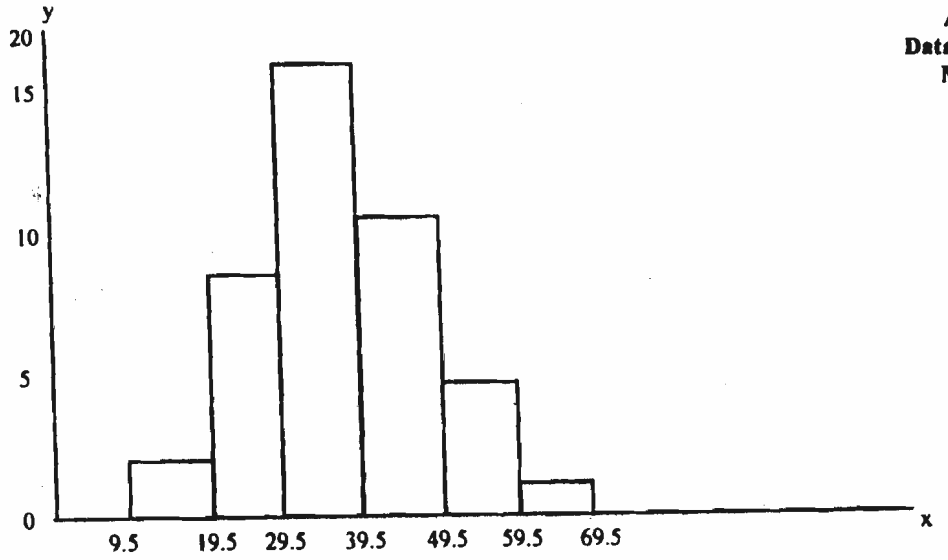


Fig. 14.1 : The Histogram or Column

14.4.2 Frequency Polygon

Another method of representing a frequency distribution graphically is 'Frequency Polygon'. It is obtained by identifying the point which represent the mid-point of each class interval and then joining all these points by straight lines as shown in the figure (Fig. 14.2).

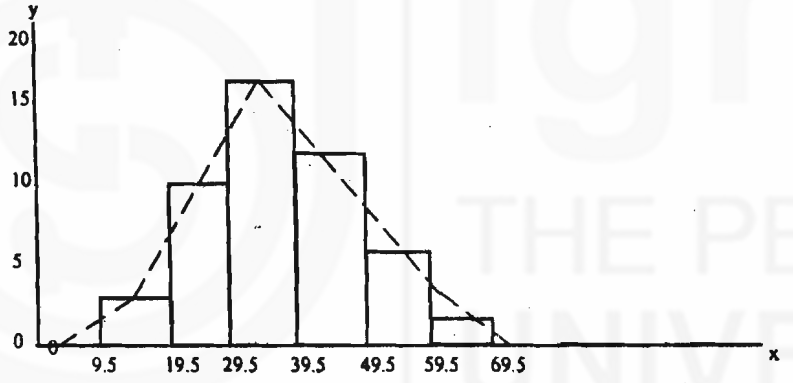


Fig. 14.2 : Frequency Polygon

14.4.3 Cumulative Frequency Curve

Cumulative frequency curve differs from frequency polygon in two ways. First, instead of plotting points corresponding to frequencies, we plot points corresponding to cumulative frequencies. Second, instead of plotting mid points of each class, we plot our points above the exact upper limit of the class interval. This is done because in this graph we wish to represent the number of cases falling above or below the particular values. Let us consider the distribution as given in 14.4.1.

Exact Units of Class Intervals	f (Frequency)	F (Cumulative Frequency)
59.5 - 69.5	1	40 = (39+1)
49.5 - 59.5	4	39 = (35+4)
39.5 - 49.5	10	35 = (25+10)
29.5 - 39.5	15	25 = (10+15)
19.5 - 29.5	8	10 = (2+8)
9.5 - 19.5	2	2
	N = 40	

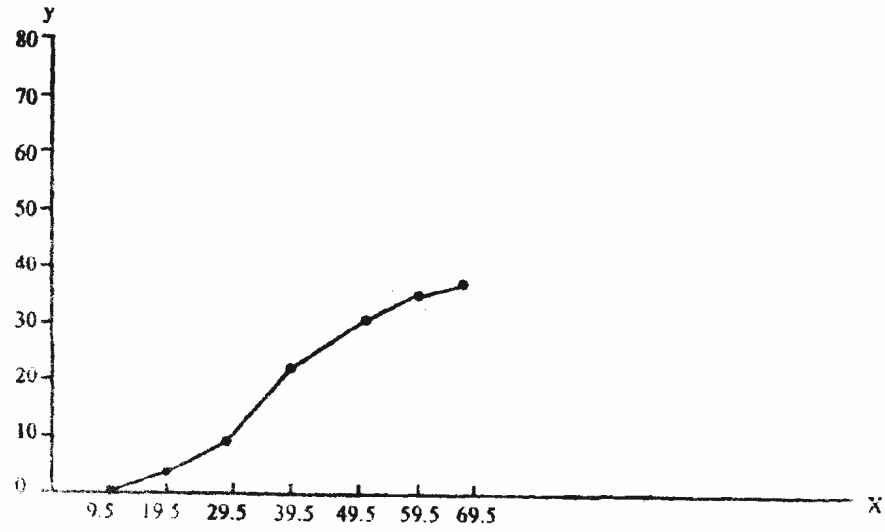


Fig. 14.3 : Cumulative Frequency Curve

For drawing the cumulative frequency curve of this distribution, each cumulative frequency is plotted at the exact upper limit of the class interval upon which it falls. For the given distribution first value of cumulative frequency corresponding to 19.5 is 2; the second value of cumulative frequency corresponding to 29.5 is 10 and so on to the last point for which value of cumulative frequency corresponding to 69.5 is 40. By joining the plotted points by lines, we get 'frequency curve' as illustrated in the Figure (Fig. 14.3). It may be noted that in order to have the curve begin on the horizontal axis, we start at the exact lower limit of the lower class (i.e. 9.5 in the present case), the cumulative frequency of which is taken to be '0'.

14.4.4 Cumulative Percentage Curve or Ogive

Cumulative percentage curve or Ogive differs from cumulative frequency curve in that frequencies are expressed as cumulative percents of N on the vertical axis instead of as cumulative frequencies, as illustrated in the given example.

Exact Units of Class Intervals	f	F (Cumulative Frequency)	Percentage (Cumulative Frequency)
59.5 - 69.5	1	40	100
49.5 - 59.5	4	39	97.5
39.5 - 49.5	10	35	87.5
29.5 - 39.5	15	25	62.5
19.5 - 29.5	8	10	25.0
9.5 - 19.5	2	2	5.0
	N = 40		

In the above distribution, the percentage cumulative frequency are expressed as percentages of N in the last column. After finding cumulative percentage frequencies we plot these frequencies corresponding to upper exact limits of class intervals, as shown in the figure (Fig. 14.4). The curve joining the points thus drawn is called cumulative percentage curve or Ogive.

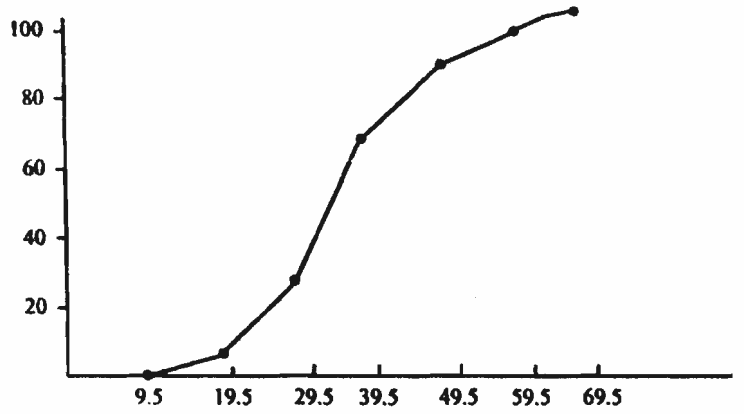


Fig. 14.4 : Cumulative Percentage Curve or Ogive

Check Your Progress

Notes : a) Space is given below for your answer.

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

1. i) What are the various types of quantitative data.

.....

.....

ii) Name various techniques of presenting data graphically.

.....

.....

14.5 DESCRIPTIVE STATISTICAL MEASURES

Statistical methods are extensively used in analyzing quantitative data in educational research. Generally, these methods are classified as 'descriptive statistics' and 'inferential statistics'. Descriptive statistics are computed to describe the characteristics (attributes) of a sample or population in totality and thus limit generalization to the particular group (sample). No conclusions are extended beyond this group whereas inferential statistical methods are used to draw generalizations beyond the sample with a known degree of accuracy.

In this unit we shall confine our discussion to the understanding of four measures of descriptive statistics commonly used in educational research. These include measures of: (i) central tendency or averages; (ii) variability or dispersion ; (iii) relative position; and (iv) relationship.

14.5.1 Measures of Central Tendency or Averages

We use averages to describe the characteristics of samples (groups) in a general way. Economic status of groups indicated by 'average income', performance of students in a class is judged by 'grade point averages', or the climate of a city is reported by 'average temperature'. But in statistical language, use of the term 'average' does not convey any specific meaning, because there are a number of types of averages, only one of which may be appropriate to use in describing given characteristics of a sample/group. Of the many averages that may be used, three are mostly used in the analysis of educational data. These include: Mean, Median and Mode.

The Mean

The Mean of a distribution is the arithmetic average. It is perhaps the most familiar, most frequently used and well understood average. It is computed by dividing the sum of all the observations by the total number of observation. If x_1, x_2, x_3, ... x_n are the N observations, the formula for computing the Mean (X) is given

$$\text{Mean} = M = \frac{x_1 + x_2 + x_3 + \ldots\ldots\ldots\ldots x_n}{N}$$

$$= \frac{\sum x}{N}$$

Suppose 10, 12, 7, 6, 4, 10, 15 are the marks obtained by 7 students in a test of Hindi, the Mean is

$$\text{Mean} = M = \frac{10+12+7+6+4+10+5}{7} = \frac{64}{7} = 9.14$$

In case of large data, the observations are arranged in a frequency distribution. Consider the following distribution which was used in 14.4.1.

Scores Class Interval	Exact Units of Class Interval	Mid Point x	f	x'	fx'
60 - 69	59.5 - 69.5	64.5	1	3	3
50 - 59	49.5 - 59.5	54.5	4	2	8
40 - 49	39.5 - 49.5	44.5	10	1	10
30 - 39	29.5 - 39.5	34.5 A.M.	15	0	0
20 - 29	19.5 - 29.5	24.5	8	–1	–8
10 - 19	9.5 - 19.5	14.5	2	–2	–4
			N = 40		$\sum$ fx'=9

These data are grouped in a frequency distribution. The following formula is used for computing the Mean in such cases:

$$\text{Mean} = M = A.M. + \frac{\sum fx'}{N} \times i$$

in which

A.M. = Assumed Mean

f = Frequency of the Class Interval

x' = Deviation of the score from the Assumed Mean Divided by the length of the class interval

i = Width of the class interval

N = Total number of observations

In the column Mid point (x) of the example, we have computed the mid-point of each class interval and taken the assumed mean (A.M.) at the interval which has the maximum frequency.

The values for the column (x') have been calculated by taking deviation of each mid point (x) from the assumed mean dividing it by the length of the class interval

i.e. 10 in the present case. The values fx' are obtained by multiplying each f with the corresponding x'.

Using the formula:

$$\text{Mean} = M = A.M. + \frac{\sum fx'}{N} \times i$$

$$= 34.5 + \frac{9}{10} \times 10$$

$$= 34.5 + 2.25$$

$$= 36.75$$

Selection and Application of Mean

Generally, Mean is a more useful measure of central tendency than its other measures.

Mean is an appropriate statistic when

i) the observations or measures are distributed symmetrically about the centre of the distribution;

ii) the most stable measure of central tendency is desired;

iii) additional statistics are to be computed later.

The Median

The Median is a point in an ordered arrangement of observations, above and below which one-half of the observations fall. It is a measure of a position rather than one of magnitude.

If the number of observations are ungrouped and small, we arrange them in order of magnitude and identify the middle observation by counting up half the value of N (the total number of observations). When the number of observations (N) is odd, the mid-observation is the Median.

For example, 45 is the median of the marks 30, 31, 45, 48, 52 obtained by 5 students in a test of Mathematics.

When the number of observations is even, the Median is the mid-point between two middle observations.

For example, $\frac{35+37}{2} = 36$ is the Median of the observations:

29, 31, 35, 37, 40, 44.

In case of data, grouped in frequency distribution we use the following formula:

$$\text{Median} = \text{Mdn} = l + \frac{(N/2 - F)}{f} \times i$$

in which

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

i = Width of class interval in which the median falls.

f = Frequency within the class interval upon which the median lies.

F = Sum of all the frequencies below l.

$\frac{N}{2}$ = One-half of the total number of observations.

To illustrate the use of this formula, consider the data which was used in 14.4.1 for computing the Mean.

Here $\frac{N}{2}=\frac{40}{2}=20$; $l=29.5$; $F=8$; $f=15$ and $i=10$

Using the formula:

$$\text{Median} = (\text{Mdn.}) = l + \frac{(N/2 - F)}{f} \times i$$

$$= 29.5 + \frac{(20-8)}{15} \times 10$$

$$= 29.5 + \frac{12}{15} \times 10$$

$$= 29.5 + 8$$

$$= 35.5$$

Selection and Application of Median

We should compute Median when

i) the distribution of the observation measures is markedly skewed;

ii) there is not sufficient time to compute a Mean;

iii) an incomplete distribution is given;

iv) we are interested in whether cases fall within the upper or lower halves of distribution and not particularly in how far they are from the central point.

The Mode

The mode is defined as the most frequently occurring observation in a distribution. It is located by inspection rather than by computation. If there is only one value which occurs a maximum number of times, then the distribution is said to have one mode or to be unimodal. In some distributions there may be more than one mode. A two mode distribution is bimodal; more than two, multimodal.

In a simple ungrouped series of measures of observations, the crude or empirical mode is that single measure which occurs most frequently. For example, in the series 19, 20, 21, 26, 28, 28, 29 and 31, the most often recurring measure, namely, 28, is the crude or empirical mode.

In grouped data distributions, the mode is assumed to be the mid score of the interval in which the greatest frequency occurs. For example, in the grouped distribution under 14.4.1 which was used for computing Mean, 34.5 is the Mode. It is the mid score of the interval 29.5-39.5 with maximum frequency of 15.

Selection and Application of Mode

The mode is the most appropriate statistics when

i) the quickest estimate of central tendency is wanted;

ii) a very rough estimate of central tendency will suffice;

iii) the purpose is to know the most typical case.

Check Your Progress

Notes : a) Space is given below for your answer.

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

2. i) Name the most stable measure of central tendency.

.....

ii) Compute mean and mode for the following distribution:

10, 12, 5, 7, 8, 9, 15, 5

.....

.....

iii) Compute median for the following distribution.

Scores	f
120 - 139	50
100 - 119	150
80 - 99	500
60 - 79	250
40 - 59	50
	N = 1000

14.5.2 Measures of Variability or Dispersion

Measures of central tendency describe location along an ordered scale. Although these statistics are very useful in describing the nature of a distribution of measures, these will not help the researcher how the observation data tend to be distributed. For this another kind of statistics, the measure of variability, is used. It is also called the measure of dispersion or spread.

There are several measures of variability, but in this unit we shall confine our discussion to the: (i) range, (ii) average deviation and (iii) standard deviation.

The Range

The range is the simplest measure of variability. It is the difference between the highest and lowest observation or score. For example, suppose the marks obtained by a group of 10 students in a test of mathematics are:

90, 80, 72, 71, 70, 70, 69, 68, 60, 50

The range for this distribution will be (90 -50) + 1 = 41

Selection and Application of Range

Although the range has the advantage of being easily calculated, it has some limitations. First, the value of range is based on only two extreme scores in the total distribution and thus it does not provide us any information of the variation of many other scores of the distribution. Second, it is not a stable statistics as its value can differ from sample to sample drawn from the same population.

The range is used when:

i) the data are too scant or too scattered;

ii) only an idea of extreme scores or of total spread is wanted.

Average Deviation (AD)

The average deviation or AD is the mean of the deviations of all of the separate observations or scores in a series taken from their mean. However, in averaging deviations to find average deviation (AD), the signs (+ and –) are not taken into consideration i.e. all the deviations, whether plus or minus, are treated as positive.

The formula for computing average deviation of an ungrouped data is:

$$\text{Average Deviation} = \text{AD} = \frac{\sum|x|}{N}$$

in which

$\sum|x|$ = Sum of deviations of scores from their mean disregarding plus and minus signs.

N = Total number of scores.

To illustrate the computation of average deviation, consider the following ungrouped distribution of marks obtained by a group of ten students in a test of spellings.

18, 19, 21, 19, 27, 31, 22, 25, 28, 20

Let us find the mean of the distribution:

X	x	|x|
18	– 5	5
19	– 4	4
21	– 2	2
19	– 4	4
27	+ 4	4
31	+ 8	8
22	– 1	1
25	+ 2	2
28	+ 5	5
20	– 3	3
$\sum x = 230$	$\sum x' = 0$	$\sum\|x'\| = 23$

$$\text{Mean} = \frac{\sum x}{N} = \frac{230}{10} = 23$$

The deviations of the scores from the mean, neglecting the negative signs, are given in the column |X|.

Using the formula:

$$\text{Average Deviation} = \frac{\sum|x|}{N} = \frac{23}{10} = 2.3$$

It may be noted that the sum of the deviations of the observations (scores) from their actual mean is zero. In the present case = $\sum x = 0$

For a grouped data distribution, the formula for computing average deviation is:

$$\text{Average Deviation (AD)} = \frac{\sum|fx|}{N}$$

in which $\sum|fx|$ = sum of the product of f and x, ignoring the minus sign.

Using the data of 14.4.1, the Average Deviation of the distribution will be:

$$\text{Average Deviation} = \frac{\sum|fx|}{N} = \frac{33}{40} = 0.825$$

Selection and Application of Average Deviation (AD)

Average deviation is used when:

i) It is desired to consider all deviations from the mean according to their size;

ii) Extreme deviations would affect standard deviation unduly.

Quartile Deviation (Q)

The quartile deviation or Q is one-half the scale distance between the 75th and 25th percentiles in a frequency distribution. The 25th percentile or Q_1 is the first quartile on the score scale, the point below which lie 25 per cent of the scores. The 75th percentile or Q_3 is the third quartile on the score scale, the point below which lie 75 per cent of the score. Mathematically,

in which

$$Q = \frac{Q_3 - Q_1}{2}$$

$$Q_1 = l + i\frac{\left[\frac{N}{4} - F\right]}{f}$$

$$Q_3 = l + i\frac{\left[\frac{3N}{4} - F\right]}{f}$$

where

l = The exact lower limit of the interval in which the quartile falls
i = The length of the interval
F = Cumulative frequency upto the interval which contains the quartile
f = The frequency on the interval containing the quartile.

To illustrate the use of the formula, consider the following distribution:

Scores	Exact Units of Class Interval	f	F
52 - 55	51.5 - 55.5	1	65
48 - 51	47.5 - 51.5	0	64
44 - 47	43.5 - 47.5	5	64
40 - 43	39.-5 - 43.5	10	59
36 - 39	35.5 - 39.5	20	49
32 - 35	31.5 - 35.5	12	29
28 - 31	27.5 - 31.5	8	17
24 - 27	23.5 - 27.5	2	9
20 - 23	19.5 - 23.5	3	7
16 - 19	15.5 - 19.5	4	4
		N = 65	

Here $\frac{N}{4} = \frac{65}{4} = 16.25$ and $\frac{3N}{4} = 48.75$

$$Q_1 = l + \frac{\left[\frac{N}{4} - F\right]}{f} = 27.5 + 4\,\frac{(16.25-9)}{8}$$

$$= 27.5 + 4\,\frac{(7.25)}{8}$$

$$= 27.5 + \frac{7.25}{2}$$

$$= 27.5 + 3.63 = 31.13$$

$$Q_3 = l + i\,\frac{\left[\frac{3N}{4} - F\right]}{f} = 35.5 + 4\,\frac{(48.75-29)}{20}$$

$$= 35.5 + 4\,\frac{(19.75)}{20}$$

$$= 35.5 + \frac{19.75}{5}$$

$$= 35.5 + 3.95 = 39.45$$

$$Q = \frac{Q_3 - Q_1}{2} = \frac{39.45 - 31.13}{2} = 4.16$$

Selection and Application of Quartile Deviation (Q)

Quartile deviation is when:

i) the mean is the measure of central tendency;

ii) there are scattered or extreme scores which would influence the standard deviation disproportionately;

iii) the concentration around the median, the middle 50 per cent of cases, is of primary interest.

Standard Deviation (SD)

The standard deviation (SD) is the most general and stable measure of variability. It is the positive square root of variance. The average of the squared deviations of the measures or scores from their mean is known as variance which is generally denoted as δ. The standard deviation is also denoted as σ.

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

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

in which

x = Deviation of the raw score or measure from the Mean.

N = Number of scores or measures.

To illustrate the use of this formula, let us consider the data:

15, 10, 15, 20, 8, 10, 25, 9

The mean or average score is:

$$M = \frac{\sum x}{N} = \frac{112}{8} = 14$$

Now we subtract the mean from each of the raw scores in the distribution and proceed as under:

Score (x)	X – M or x	x^2
15	1	1
10	– 4	16
15	1	1
20	6	36
8	– 6	36
10	– 4	16
25	11	121
9	– 5	25
		$\sum x^2 = 252$

Using the formula:

Standard Deviation = $\sigma = \sqrt{\frac{\sum x^2}{N}} = \frac{252}{8}$

$= \sqrt{31.8} = 5.64$

We can also compute standard deviation directly from the raw scores without using the deviation with the help of the formula:

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

in which

x = Raw score

N = The number of scores in the distribution.

x	x^2
15	225
10	100
15	225
20	400
8	64
10	100
25	625
9	81
$\sum x = 112$	$\sum x^2 = 1820$

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

$$= \frac{\sqrt{(8)(1820) - (112)^2}}{8}$$

$$= \frac{\sqrt{14560 - 12544}}{8}$$

$$= \frac{\sqrt{2016}}{8}$$

$$= \frac{44.90}{8}$$

$$= 5.612$$

The formula for computing standard deviation for the grouped data is:

$$S.D. = \sigma = \frac{i}{N}\sqrt{N\sum fx'^2 - \left(\sum fx'\right)^2}$$

in which

i = length of the class interval

N = total number of measures or scores

x = deviation of the raw score from the assumed mean divided by the length of the class interval.

To illustrate the use of this formula let us consider the data used for calculating S.D.

Scores	f	Mid Point (x)	x′	fx′	fx²
52 - 55	1	53.5	4	4	16
48 - 51	0	49.5	3	0	0
44 - 47	5	45.5	2	10	20
40 - 43	10	41.5	1	10	10
36 - 39	20	AM 37.5	0	0	0
32 - 35	12	33.5	–1	–12	12
28 - 31	8	29.5	–2	–16	32
24 - 27	2	25.5	–3	–6	18
20 - 23	3	21.5	–4	–12	48
16 - 19	4	17.5	–5	–20	100
	N = 65			∑fx = – 42	∑fx² = 256

Using the formula:

$$S.D. = \sigma = \frac{i}{N}\sqrt{\sum fx'^2 - \left(\sum fx'\right)^2}$$

$$= \frac{4}{65}\sqrt{(65)(256)-(-42)^2}$$

$$= \frac{4}{65}\sqrt{16640-1764}$$

$$= \frac{4}{65}\sqrt{14876}$$

$$= 7.51$$

Selection and Application of Standard Deviation (σ)

Standard deviation is used when:

i) the statistics having greatest stability is sought;

ii) extreme deviations exercise a proportionally greater effect upon the variability;

iii) co-efficient of correlation and other statistics are subsequently computed.

Check Your Progress

Notes : a) Space is given below for your answer.

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

3. i) List the situations when we compute Range and Average Deviation.

.....

.....

.....

ii) Mention the uses of quartile deviation.

.....

.....

.....

iii) Compute Standard Deviation for the following data:

18, 25, 21, 19, 27, 31, 22, 25, 28, 20

.....

.....

.....

14.5.3 Normal Probability Curve

The normal probability curve is based upon the law of probability or probable occurrence of certain events The 'probability' of a given event is defined as the expected frequency of occu ıce of this event among events of a like sort. It is expressed as a 'ratio'. For example, the probability of an unbiased coin falling heads is ½ and that of a dice showing a three-sport is 1/6. When any set of observations confirms to this mathematical form, it can be represented by a bell-shaped curve with the following characteristics.

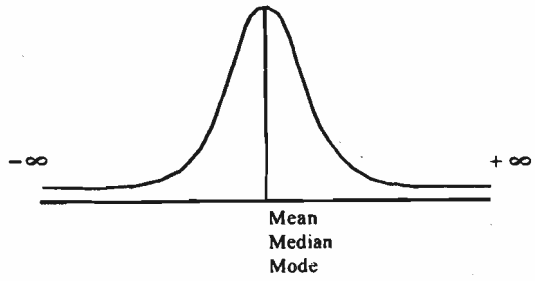


Fig. 14.5 : Normal Probability Curve

1. The curve is symmetrical around its vertical axis. It implies that the size, shape and slope of the curve on one side of the curve is identical to that on the other side of the curve.
2. The values of mean, mode and median computed for a distribution following this curve always coincide and have the same value.
3. The height of the vertical line called ordinate is maximum at the mean and in the unit normal curve is equal to 0.3989.
4. The curve has no boundaries in either direction, for the curve never touches the base line, no matter how far it is extended i.e. the curve is asymptotic and extends from – ∞ (minus infinity) to + ∞ (plus infinity).
5. The data 'cluster' around the mean. The percentages in a given standard deviation are greatest around the Mean and decrease as one moves away from the Mean.

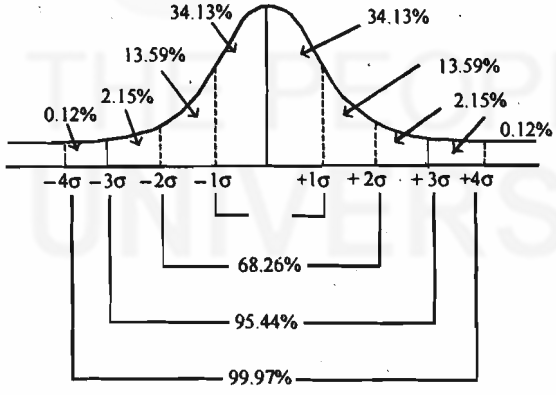


Fig. 14.6

The percentage area around the Mean are:

i) Mean ± 1.00 σ 34.13%

ii) ± 1.00 to ± 2.00 σ 13.59 %

iii) + 2.00 to ± 3.00 σ 2.15%

This implies that 68.26% of the total area of the curve falls between the limits Mean + 1σ and Mean –1; 95.44 per cent of the total area of the curve falls between limits Mean + 2 σ and Mean – 2 σ and 99.73 per cent of the total area of the curve falls between Mean + 3 σ and Mean – 3 σ.

The equation of the normal curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

in which

x = scores, expressed as deviations from the mean

y = height of the curve above the horizontal axis i.e. the frequency of a given x value

N = number of cases in the sample

π = 3.1416

e = 2.7183

When N and σ are known, it is possible from the equation to compute: (i) the frequency (y) of a given value x; and (ii) the number or percentage between two points, or above or below a given value point in the distribution.

However, these calculations are not needed as Normal Probability Tables are available from which these values can be readily obtained.

Applications of the Normal Curve

There are a number of applications of the normal curve in the field of educational research.

1. When the measures or scores are normally or nearly normally distributed, the Normal Probability Table is useful: (i) to find the percentage of cases, or the number when N is known, that fall between the mean and a given σ distance from the mean; (ii) the percentage of the total area included between the mean and a given σ distance from the mean.

2. The normal curve is used to convert a raw score into standard score. Suppose the student A gets a score of 50 in an achievement test in mathematics which are administered to random sample of 300 students. The mean and standard deviation of the test scores, assumed to be normally distributed, for the sample are 60 and 6 respectively. The standard score or sigma score of the student A is :

   $$Z = \frac{X - M}{\sigma} = \frac{50 - 60}{6} = \frac{-10}{6} = -1.67$$

   indicating that the score of 50 is 1.67 standard deviation below the Mean.

3. Normal curve is useful in calculating the percentile rank of scores in a normal distribution.

4. For normalising a given frequency distribution, the normal curve is used.

Check Your Progress

Notes : a) Space is given below for your answer.

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

4. i) What percentage of cases in a normal probability curve lie between M + 2σ and M – 2σ ?

.....

ii) State two characteristics of a normal probability curve.

.....

.....

14.5.4 Measures of Relative Positions

When we administer a test on a sample of students drawn randomly from a population, the scores obtained are raw. A raw score, taken by itself, has no meaning. It gets meaning only by comparison with some reference group or groups. We generally use the following measures for making comparisons.

1. Sigma Score (Z)
2. T Score
3. Percentile Rank

Sigma Score (Z)

In describing a score in a distribution, its deviation from the mean is more meaningful than the score itself. The unit of measurement is the standard deviation. Mathematically, it is expressed as;

$$Z = \frac{x - x}{\sigma} \text{ or } \frac{x}{\sigma}$$

in which

x = raw score

x = mean

σ = standard deviation

x = (x – x) score deviation from the mean.

To illustrate the concept, let us consider the following data:

Test 1	Test 2
x = 77 x = 84 σ = 4	x = 64 x = 58 σ = 5
$Z = \frac{77-84}{4} = \frac{-7}{4} = -1.75$	$Z = \frac{64-58}{5} = \frac{6}{5} = +1.20$

The raw score of 77 in Test 1. may be expressed as a sigma score of –1.75, indicating that 77 is 1.75 standard deviation below the mean. The raw score of 64 in Test 2 may be expressed as a sigma score of + 1.20, indicating that 64 is 1.20 standard deviation above the mean.

On sigma scale, the mean of any distribution is converted to zero and the standard deviation is equal to 1. Thus a sigma score is useful in making realistic comparison of scores and thus provides a basis for equal weighting of the scores.

Suppose a teacher wants to determine a students' equally weighted average (mean) achievement on a physics test and on a chemistry test on the basis of the following data:

Subject	Test Score	Mean	Highest Possible Score	Standard Deviation
Physics	50	60	80	6
Chemistry	65	92	90	15

It is evident that the mean of the two raw scores obtained by the student in Physics and Chemistry tests would not provide a valid picture of the student's performance because the mean would be weighted more in favour of the chemistry test score. The conversion of each test score to a sigma score makes them equally weighted and comparable as both the raw test scores have been expressed on a scale with a mean of zero and standard deviation of one.

$$Z = \frac{Z-O}{1} = \frac{x-x}{\sigma}$$

$$\text{Physics Z Score} = \frac{50-60}{6} = \frac{-10}{6} = -1.67$$

$$\text{Chemistry Z Score} = \frac{65-92}{15} = \frac{-27}{15} = -1.80$$

The teacher, on an equally weighted basis, may conclude that the performance of the student was fairly consistent, i.e., 1.67 standard deviation below the mean in physics and 1.80 standard deviation below the mean in chemistry.

In the application of normal probability curve, it was explained that the curve describes the percentage of area lying between the mean and successive deviation units. The sigma or Z score can be used in hypothesis testing and determination of percentile ranks.

T Score (T)

The value sigma (Z) score in most of the cases is expressed as scores with decimals or negatives. In order to avoid this type of situation, another version of a standard score, the T score, has been devised. It is expressed as:

$$T = 50 + 10\frac{x-x}{\sigma} \text{ or } 50 + 10Z$$

Using the scores in the previous example, we can compute the physics and chemistry T scores as:

Physics T = 50 + 10 (– 1.67) = 33.30

Chemistry T = 50 + 10 (– 1.80) = 32.00

T scores are always rounded to the nearest whole number. Thus the T scores of the student in Physics and Chemistry are 33 and 32 respectively.

Percentile Rank

The percentile rank is the point below which a given percentage of scores fall. For example, if the 80th percentile is a score of 60, 80 per cent of the scores fall below 60. The median is the 50th percentile rank, for 50 per cent of the scores fall below it.

The formula for converting a given rank into percentile ranks is:

$$\text{Percentile rank} = 100 - \left[\frac{100\,RK - 50}{N}\right]$$

in which RK = rank from the top.

For example, a student A ranks 6th in a class of 150 students. This means 5 students rank above the student A, 144 below him. The percentile rank of the student A is:

$$= 100 - \left[\frac{100 \times 6 - 50}{150}\right]$$

$$= 100 - (3.67)$$

$$= 100 - 4$$

$$= 96$$

Check Your Progress

Notes : a) Space is given below for your answer.

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

5. i) Define sigma score and T score.

.....

.....

ii) Shyam ranks twenty seventh in his class of 139 students. Find his percentile rank.

.....

iii) Find the sigma score of a student whose raw score in a test of Hindi is 76. The mean of the Hindi scores of the students in his class is 82 and standard deviation 4. What will be his T score.

.....

.....

14.5.5 Measures of Relationship

We have so far discussed the statistical description of a single variable. Now we shall study the problem of describing the degree of simultaneous variation of two variables. The data in which we obtain measures of two variables for each individual is called a bivariate data. For example, we get bivariate data if we have scores of the tests of Mathematics and Hindi for a group of school children. The essential feature of the bivariate data is that one measure can be paired with another measure for each member of the group.

When we study bivariate data we may like to know the degree of relationship between variables of such data. This degree of relationship between variables is known as 'correlation' which is represented by the co-efficient of correlation. This co-efficient may be identified by either the r, the Greek symbol rho (r), or other symbols depending upon the data distributions and the way the co-efficient has been calculated.

Students who obtain high scores on an intelligence test tend to get high scores in an achievement test in mathematics; whereas those with low scores on an intelligence test tend to score low in mathematics. When this type of relationship is obtained the two variables are said to be positively related.

When the increase in one variable is associated with decrease in the other variable, the variables are said to be negatively related.

When the relationship between two sets of variables is a pure chance relationship, we say there is no correlation.

The intensity or degree of linear relationship is represented quantitatively by co-efficient of correlation. Its value ranges from –1.00 to +1.00. A value of –1.00 indicates a perfect relationship between the two variables and +1.00 describes perfect positive relationship. A zero value indicates complete lack of correlation between the two variables. The sign (– or +) indicates the **direction** of the relationship and the numerical value its **magnitude** or strength.

Methods of Correlation

There are various methods of correlation. Their use is relative to the situation and type of data. We may have data in scores. There are many situations in which the researcher does not have scores and has to deal with data in which differences in a given attribute or characteristic can be expressed only by ranks or classifying an individual into one of several descriptive categories. In this unit we will discuss two methods of correlation, namely, product moment correlation and rank order correlation.

Product Moment Correlation (r)

In some situations the data for two variables x and y are expressed in interval or ratio level of measurement and the distribution of these variables have a linear relationship. Moreover, the distributions of the variables are uni-modal and their variances are approximately equal. In such situations we make use of product moment method of correlation. It is also called Pearson's r.

1. Calculation of Pearson's r from Ungrouped Data

When the 'N' (size of the sample or group) is small or raw scores are small numbers, there is no need of grouping the data. The formula for computing 'r' is:

$$rxy = \frac{N\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

in which

x = deviation of x measures from the assumed mean

y = deviation of y measures from the assumed mean

To illustrate the use of this formula, let us compute product-moment r form the following data in the two variables X and Y for 8 students.

X : 15, 18, 22, 17, 19, 20, 16, 21

Y : 40, 42, 50, 45, 43, 46, 41, 41

For computing the values of $\sum x, \sum y, \sum x^2, \sum y^2$ and $\sum xy$, we proceed as under:

X	Y	x	y	x^2	y^2	xy
15	40	–2	–5	4	25	10
18	42	+1	–3	1	9	–3
22	50	+5	+5	25	25	25
17 AM	45 AM	0	0	0	0	0
19	43	+2	–2	4	4	–4
20	46	+3	+1	9	1	3
16	41	–1	–4	1	16	4
21	41	+4	–4	16	16	–16
		$\sum x = +12$	$\sum y = -12$	$\sum x^2 = 60$	$\sum y^2 = 96$	$\sum xy = 19$

$$rxy = \frac{N\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[N\sum x^2 - \left(\sum x\right)^2\right]\left[N\sum y^2 - \left(\sum y\right)^2\right]}}$$

$$= \frac{(8)(19) - (12)(-12)}{\left[(8)(60) - (12)^2\right]\left[(8)(96) - (-12)^2\right]}$$

$$= \frac{152 + 144}{\sqrt{[480 - 144][768 - 144]}} = \frac{296}{\sqrt{(336)(624)}}$$

$$= \frac{296}{\sqrt{209664}} = \frac{296}{457.891} = 0.646$$

$$= 0.65$$

2. Calculation of Pearson's r from Grouped Data

When N is large or even moderate in size, the best procedure is to group data in both variables X and Y in the form of a scattergram. The Pearson's r is computed by using the following formula:

$$r = \frac{N\sum fxy - \left(\sum fx\right)\left(\sum fy\right)}{\sqrt{\left[N\sum fx^2 - \left(\sum fx\right)^2\right]\left[N\sum fy^2 - \left(\sum fy\right)^2\right]}}$$

To illustrate the use of this formula, let us consider the data in the following scattergram:

		Mathematics Test Scores (X)											
		12-13	14-15	16-17	18-19	20-21	22-23	24-25	f	y	fy	fy²	fxy
Language Test Scores (Y)	35-37					0 1 0		6 1 6	2	+3	6	18	6
	32-34					0 6 0	2 3 6		9	+2	18	36	6
	29-31		–3 1 –3	–2 2 –4	–1 6 –6	0 8 0	1 1 1		18	+1	18	18	–12
	26-28		0 4 0	0 4 0	6 0	0 11 0	0 4 0	0 1 0	30	0	0	0	0
	23-25	4 2 8	3 1 3	2 6 12	1 5 5	0 4 0	–1 1 –1		19	–1	–19	19	27
	20-22	6 3 24	43 1 4	2 1 2					7	–2	–14	28	42
	f	5	8	13	18	30	9	2	85		9	119	69
	x	–4	–3	–2	–1	0	+1	+2					
	fx	–20	–24	–26	–18	0	9	4	–75				
	fx²	80	72	52	18	0	9	8	239				
	fxy	32	12	12	1	0	6	6	62				

Using the formula:

$$r = \frac{N\sum fxy - \left(\sum fx\right)\left(\sum fy\right)}{\sqrt{\left[N\sum fx^2 - \left(\sum fx\right)^2\right]\left[N\sum fy^2 - \left(\sum fy\right)^2\right]}}$$

$$= \frac{(85)(69) - (-75)(9)}{\sqrt{\left[(85)(239) - (-75)^2\right]\left[(85)(119) - (-9^2)\right]}}$$

$$= 0.54$$

The computation for the values of Σfx, Σfx^2, Σfy, Σfy^2, and Σfxy, is done using the following steps:

Step 1

The distribution of language test scores for 85 students is found in the f column at the right of the scattergram. Assume a mean for the distribution of scores for a language test (the mid point of that interval which contains the largest frequency), and draw bold lines to mark off the row in which the assumed mean falls. Here the mean for the language test scores has been taken at 27 (mid-point of interval 26-28) and y's (deviations from the assumed mean) have been taken from this point. Fill in the values of fy and fy^2 columns.

Step 2

The distribution of the mathematics test scores of 85 students is found in the f row at the bottom of the scattergram. Assume a mean for this distribution, and draw bold lines to designate the column under the assumed mean. The mean for the mathematics test scores is taken at 20.5 (mid-point of interval 20-21) and x's (deviations from the assumed mean) are taken from this point. Fill in the fx and fx^2 rows.

Step 3

The fxy for a cell is computed by multiplying the frequency given in the particular cell with the corresponding x and y.

Rank Order Correlation (ρ)

Rank order correlation is also known as the Spearman rank order co-efficient of correlation and is denoted by rho (ρ). This method is used when the data are available only in ordinal form of measurement (ranked) rather than in interval or ratio form, or if the number of observations or measures is small.

For computing the rank order correlation (r), the following formula is used:

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

in which

D = difference between paired ranks

N = number of paired ranks

To illustrate the use of this formula, let us consider the following data:

X : 15, 18, 22, 17, 19, 20, 16, 14, 17, 22

Y : 41, 40, 42, 50, 45, 38, 46, 41, 40, 39

First, let us convert the data to ranks by assigning rank 1 to highest measure and so on. If a measure is repeated, the ranks are averaged. For example, in the variable X, 22 is the highest measure and is repeated. The ranks assigned are 1 and 2.

The average is $\frac{1+2}{2}=\frac{3}{2}=1.5$. The next measure is given the rank 3 and so on.

X	Y	Rx	Ry	D	D^2
15	41	9	6.5	2.5	6.25
18	40	5	8	–3	9.00
22	42	1.5	5	–3.5	12.25
17	50	6.5	1	5.5	30.25
19	45	4	4	0	0
20	38	3	10	–7	49.00
16	46	8	3	5	25.00
14	41	10	6.5	3.5	12.25
17	49	6.5	2	4.5	20.25
22	39	1.5	9	–7.5	56.25
					$\sum D^2 = 220.50$

Using the formula,

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

$$= 1 - \frac{(6)(220.50)}{10(100-1)}$$

$$= 1 - \frac{1323}{10\times 99}$$

$$= 1 - \frac{1323}{990}$$

$$= 1 - 1.336$$

$$= 0.336$$

$$= 0.34$$

Check Your Progress

Notes : a) Space is given below for your answer.

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

6. i) List the situations in which product moment correlation is used.

.....

.....

.....

ii) Compute product moment correlation for the following data:

X : 50, 54, 56, 59, 60, 62, 61, 65, 67, 71, 71, 74

Y : 22, 25, 34, 28, 26, 30, 32, 30, 28, 34, 36, 40

.....

iii) Compute rank difference correlation for the data given under (ii). Compare the result with the one obtained using product moment correlation.

.....

14.6 LET US SUM UP

1. In this unit, we listed the tools and techniques which generate data of a quantitative nature. We also discussed that quantitative data are either parametric or non-parametric.
2. The methods of presenting the data graphically using histogram, frequency polygon, cumulative frequency curve and Ogive were also explained.
3. The four measures of descriptive statistics, viz., measures of central tendency, variability, relative position and relationship were discussed.
4. The use, application and computation of mean, median and mode as the measures of central tendency were explained with the help of examples.
5. The measures of variability, namely, range, average deviation, quartile deviation and standard deviation were discussed. Their application, use and computation were explained with the help of examples.
6. The characteristics of normal probability curve alongwith its use in the educational research were discussed.
7. The raw test score, taken by itself, has no meaning. It is meaningful only by comparison with some reference group or groups. For which we make use of sigma score (Z), T score and percentile ranks. The method for computing sigma score, T score and percentile ranks have been explained in this unit.
8. The direction and magnitude of a relationship between the two variables is measured with the help of co-efficient of correlation. The product moment correlation co-efficient is computed when the variables (X and Y) are expressed in interval or ratio scales of measurement and the distribution of the variables have a linear relationship. In case of ordinal (ranked) data we compute rank difference correlation. The procedure for computing these co-efficients has been discussed in this unit.

14.7 UNIT-END ACTIVITIES

1. List the situations in which we obtain data of a quantitative nature.
2. What are the types of quantitative data? Illustrate with the help of examples.
3. Compute Median and Standard Deviation for the following data::

Scores	20-21	18-19	16-17	14-15	12-13	10-11	8-9
f	2	2	4	0	4	0	4

4. Compute Mean and Quartile Deviation for the following distribution:

Class Interval	f
195 - 199	1
190 - 194	2
185 - 189	4
180 - 184	5
175 - 179	8
170 - 174	10
165 - 169	6
160 - 164	4
155 - 159	4
150 - 154	2
145 - 159	3
140 - 144	1

5. Compute Mode, Median and Standard Deviation for the following scores:
   15, 20, 17, 18, 10, 11, 12, 17, 15, 9
6. Compute Product Moment Correlation (r) and Rank Difference Correlation (ρ) between the following sets of scores:

   X : 50, 42, 51, 26, 35, 42, 60, 41, 70, 55, 62, 38

   Y : 62, 40, 61, 35, 30, 52, 68, 51, 84, 63, 72, 50

14.8 POINTS FOR DISCUSSION

1. Discuss the uses of various measures of (i) central tendency and (ii) variability.
2. What are the characteristics and uses of a normal distribution curve?
3. What are the uses of sigma score, T score and percentile rank?

14.9 SUGGESTED READINGS

Best, John and James V. Kahn (1992): *Research in Education.* New Delhi: Prentice Hall of India Pvt. Ltd.

Garrett, H.E. (1962): *Statistics in Psychology and Education.* Bombay: Allied Pacific Pvt. Ltd.

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

Koul, Lokesh (1997): *Methodology of Educational Research.* Third Revised Edition, New Delhi: Vikas Publishing House Pvt. Ltd.

14.10 ANSWERS TO CHECK YOUR PROGRESS

1. i) There are two types of quantitative data : parametric and non-parametric.

   ii) The data can be represented graphically using Histogram, Frequency Polygon, Frequency Cumulative Curve, and Ogive.

2. i) Mean

   ii) Mean = 8.88

   Mode = 5

   iii) Median = 87.5

3. i) Range is used: (a) when the data are too scant or too scattered; and (b) a knowledge of extreme scores or of total spread is wanted.

   We use average deviation when:

   a) it is desired to consider all deviations from the mean according to their size; and (b) extreme deviations would influence the standard deviation unduly.

   ii) Quartile deviation is used: (a) when the mean is the measure of central tendency; and (b) when the data are scattered or there are extreme score which would influence the standard deviation.

   iii) SD = 4.05.

4. i) 95.44%

   ii) a) The normal curve is bell shaped and symmetrical around its vertical axis called ordinate; and

   b) the curve is highest at the mean — the mean, median, and mode have the same value.

5. i) Sigma score is defined as = $Z \frac{X-X}{\sigma} or \frac{X}{\sigma}$ in which x is the mean and σ is the standard deviation of the distribution.

   T score is defined as:

   $T = 50 + 10 \frac{x - x}{\sigma}$

   $= 50 + 10 Z.$

ii) percentile rank = 81

iii) sigma score = – 1.50

T score = 35

6. i) Product moment correlation is used when:

a) The data for two variables (X and Y say) are expressed in interval or ratio scale of measurement.

b) The distributions of the variables have a linear relationship.

c) The distributions of the variables are unimodal and their variances are approximately equal.

ii) $r = 0.78$

iii) $\rho = 0.78$

UNIT 15 ANALYSIS OF QUANTITATIVE DATA: INFERENTIAL STATISTICS BASED ON PARAMETRIC TESTS

Structure

15.1 Introduction

15.2 Objectives

15.3 Inferential Statistics

15.4 Parametric Tests: Uses and Assumptions

15.5 Statistical Inference Based on Parametric Tests

- 15.5.1 Statistical Inference Regarding Means of Large Samples
- 15.5.2 Statistical Inference Regarding Means of Small Samples

15.6 Testing the Statistical Significance of the Difference Between Means

- 15.6.1 Significance of the Difference Between the Means of Two Independent Large and Small Samples
- 15.6.2 Significance of the Difference Between the Means of Two Dependent Samples
- 15.6.3 Significance of the Difference Between the Means of Three or More Samples

15.7 Statistical Inference Regarding Pearson's Co-efficient of Correlation

15.8 Significance of the Difference Between Pearson's Co-efficients of Correlation of Two Independent Samples

15.9 Let Us Sum Up

15.10 Unit-end Activities

15.11 Suggested Readings

15.12 Answers to Check Your Progress

15.1 INTRODUCTION

In the previous unit, you studied the nature of quantitative data and various descriptive statistical measures which are used in the analysis of such data. These include measures of central tendency, variability, relative position and relationship. The concept, characteristics and applications of normal probability curve were also explained.

The computed values of various statistics are used to describe the properties of particular samples. In this unit we shall discuss inferential or sampling statistics, which are useful to a researcher in making generalisations or inferences about the populations from the observations of the characteristics of samples. For making inferences about various population values (parameters), we generally make use of parametric and non-parametric tests. The concept and assumptions of parametric tests will be explained to you in this section along with the inference regarding the means and correlations of large and small samples, and significance of the difference between the means and correlations in large and small independent samples. The assumptions and applications of analysis of variance and co-variance for testing the

significance of the difference between the means of three or more samples will also be discussed.

15.2 OBJECTIVES

After studying this unit, you will be able to:

- define the concept of statistical inference;
- explain the nature of parametric tests;
- state the assumptions on which the use of parametric tests are based;
- define the sampling distribution of means and product moment correlation;
- state the characteristics of central limit theorem;
- define the standard error of mean;
- define the confidence intervals and levels of confidence;
- compute the .95 and .99 confidence intervals for the true mean from the large sample mean.
- define and illustrate the concept of degrees of freedom;
- compute the .95 and .99 confidence intervals for the true mean from the small sample mean.
- test the significance of the difference between the means of: (i) two independent large samples, and (ii) two independent small samples involving: (a) one tailed test and (b) two tailed test;
- test the significance of the difference between the means of two dependent samples;
- test the significance of the difference between the means of three or more samples using the technique of: (i) analysis of variance and (ii) analysis of covariance;
- compute the .95 and .99 confidence intervals for the true co-efficient of correlation from the sample Pearson's co-efficient of correlation;
- test the significance of Pearson's co-efficient correlation;
- test the significance of the difference between Pearson's co-efficient of correlations of the independent samples.

15.3 INFERENTIAL STATISTICS

It was explained to you in Unit 14 that the values of descriptive statistics, namely, mean, median, mode, standard deviation, correlation etc. are used to describe properties of particular samples. It is not possible to infer or make generalizations about the populations from the measures of the samples. The sampling or inferential statistics enable the researcher to make generalisations about a population characteristic (parameter) from a probability sample (statistics). The researcher computes certain statistics (sample values) as the basis for inferring what the corresponding parameters (population values) might be as it is not possible to measure all of the members/units of a given population. It may be noted that the values of the parameters for a given population are unknown.

In research situations, a researcher may draw a single sample using any of the probability sampling methods studied by you earlier. The choice of the sampling method depends upon a given set of conditions. His problem is to determine how well he can infer or estimate the parameter from the one sample statistics. Under specified conditions, it is possible for the researcher to forecast the population values (parameters) from the sample values (statistics) with a known degree of accuracy. The degree to which a sample statistics represents its parameter is an index of the significance of the computed sample statistics. While it is possible to make inferences about various parameters, in this unit we shall limit our discussion to mean and product moment correlation using parametric tests.

15.4 PARAMETRIC TESTS : USES AND ASSUMPTIONS

Parametric tests are useful as these tests are most powerful for testing the significance or trustworthiness of the computed sample statistics. However, their use is based upon certain assumptions. These assumptions are based on the nature of the population distribution and on the way the type of scale is used to quantify the data measures. It may be mentioned that there are some parametric tests, namely, t-test and F-test, which are quite robust and are appropriate even when some assumptions are not met.

The assumptions for most parametric tests are the following:

1. The variables described are expressed in interval or ratio scales and not in nominal or ordinal scales of measurement.
2. The samples have equal or nearly equal variances. This condition is known as equality or homogeneity of variances and is particularly important to determine when the samples are small..
3. The population values are normally distributed.
4. The observations are independent. The selection of one case in the sample is not dependent upon the selection of any other case.

15.5 STATISTICAL INFERENCE BASED ON PARAMETRIC TESTS

Suppose a large number of researchers selected 50 random samples each of 200 students from the population of all B.Ed. students enrolled with IGNOU and obtained their scores on a scale measuring their attitude towards teaching. The mean attitude scores of the samples would not be identical. A few would be relatively high, a few relatively low, but most of the means would tend to cluster around the population mean. The variation of sample means is due to an error which is known as 'sampling error'. This type of error explains the chance variations which are inevitable when a number of randomly selected sample means are computed.

Parametric tests are useful in making the inferences about the population mean with the help of the sample means on a probability basis. In the following discussion, you will learn how to draw statistical inference about the means of large and small samples.

15.5.1 Statistical Inference Regarding Means of Large Samples

Suppose we wish to measure the teaching aptitude of the B.Ed. distance trainees enrolled with IGNOU using a verbal aptitude teaching test. It is not possible and convenient to measure the teaching aptitude of all the enrolled B.Ed. trainees and hence we must usually be satisfied with a sample drawn from this population. However, this sample should be as large and as randomly drawn as possible to represent adequately all the B.Ed. trainees of IGNOU. If we select a large number of random samples of 100 trainees each from the population of all trainees, the mean values of teaching aptitude scores for all samples would not be identical. A few would be relatively high, a few relatively low, but most of them would tend to cluster around the population mean. The sample means due to 'sampling error' will not vary from sample to sample but will also usually deviate from the population mean. Each of these sample means can be treated as a single observation and these means can be put in a frequency distribution which is known as sampling distribution of the means.

An important principle, known as the 'Central Limit Theorem', describes the characteristics of sample means. According to this theorem, if a large number of equal-sized samples, greater than 30 in size, are selected at random from an infinite population:

1. The means of the samples will be normally distributed.
2. The average value of the sample means will be the same as the mean of the population.
3. The distribution of sample means will have its own standard deviation. This standard deviation is known as the 'standard error of the mean' which is denoted as SE_M or σ_M. It gives us a clue as to how far such sample means may be expected to deviate from the population mean. The standard error of a mean tells us how large the errors of estimation are in any particular sampling situation.

   The formula for the standard error of the mean in a large sample is:

   $$SE_M \text{ or } \sigma_M = \frac{\sigma}{\sqrt{N}}$$

   where

   σ = the standard deviation of the population

   N = the size of the sample

To illustrate the application of the central limit theorem, let us assume that the mean of the teaching aptitude scores of a sample of 64 B.Ed. trainees is 16 and the standard deviation 5. The standard error of the mean is:

$$SE_M = \sigma_M = \frac{5}{\sqrt{64}} = 5/8 = 0.625 \text{ or } 0.63$$

The value 0.63 of SE_M can be thought of as the standard deviation of a distribution of sample mean around the fixed population mean of all the B.Ed. trainees. Since the sampling distribution of sample means is assumed to be normal, it possesses all the characteristics of a normal distribution.

The normal curve in the Figure 15.1 illustrates that this sampling distribution of means is centred at the unknown population mean with its standard deviation 0.63. The sample means falls equally often on the positive and negative sides of the population mean. About 2/3 of sample means (exactly 68.26%) will lie within ± 1.00 σ_M of the population mean i.e., within a range of ± 1 × 0.63 or ± 0.63.

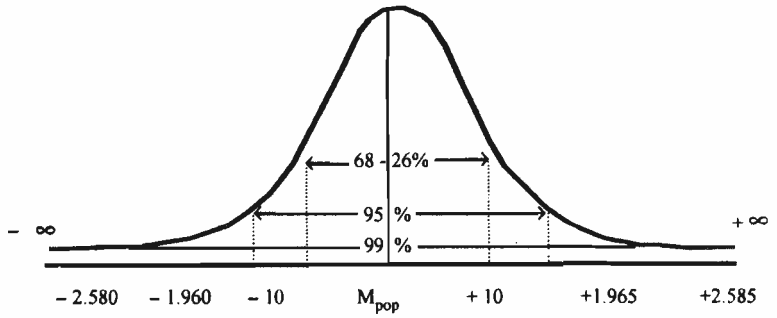


Fig. 15.1: Sampling Distribution of Means Showing Variability of Obtained Means Around Population (M_{pop}) in terms of σ_M.

Furthermore, 95 out of 100 sample means will lie within ± 2.00 σ_M (more exactly ± 1.96 σ_M) of the population mean i.e. 95 out of 100 sample means will lie within ± 1.96 × 0.63 or ± 1.23 of the population mean. The probability (P) is .95, that the sample mean of 64 does not miss the population mean (M_{pop}) by more than ± 1.23. In other words the probability is .05 that the sample mean does miss the M_{pop} by more than ± 1.23. Also, 99 of 100 sample means will lie within ± 3.00 σ_M (more exactly ± 2.58 σ_M) of the population mean i.e. 99 of sample means will lie within ± 2.58 × 0.63 or ± 1.63 of the population mean. The probability (P) is .99 that sample mean of 64 does not miss the M_{pop} by more than ± 1.63. In other words, the probability is .01 that the sample mean of 64 does miss the M_{pop} by more than ± 1.63. The magnitude of probable deviation of a sample mean from its population mean gives us a measure of the probability with which we are able to estimate the population mean (M_{pop}) from the sample mean.

Using the central limit theorem and the normal-curve concept, we may infer that the sample mean of 64 has a 95 per cent chance of being within 1.96 standard error units from M_{pop}. In other words, the mean of 64 for the random sample has a 95 per cent chance of being within 1.96 units from M_{pop}. We may also infer that there is a 99 per cent chance that the sample mean of 64 lies within 2.58 σ_M units from M_{pop}. More specifically, we can say that there is a 95 per cent probability that the limits M ± 1.96 $\acute{o}_M$ spans the population mean M_{pop}. Also, there is a 99 per cent probability that the limits M ± 2.58σ_M encloses M_{pop}.

The limits M ± 1.96 σ_M and M ± 2.58 σ_M are called '**confidence intervals**' or '**fiduciary limits**'. These limits help us to adopt particularly two **levels of confidence**. One is known as 5 per cent level or .05 level, and the other as the 1 per cent or .01 level. The .05 level of confidence indicates that the probability is .95 that M_{pop} lies within the interval M ± 1.96 σ_M and .05 that it falls outside these limits. Similarly, .01 level of confidence indicates that the probability is .99 that M_{pop} lies within the interval M ± 2.58 σ_M, and .01 that it falls outside of these limits.

In our example, the limits of M ± 1.96 σ_M will be 64 ± 1.96 × 0.63 i.e. a confidence interval marked off by the limits 62.77 and 65.23. Our confidence that this interval contains M_{pop} is expressed by a probability of .95. For a higher degree of confidence, we can take .99 level of confidence, for which the limits are M ± 2.58 σ_M i.e. a confidence interval given by the limits 62.37 and 65.63. This indicates that M_{pop} is not lower than 62.37 nor higher than 65.63 i.e., the chances are 99 in 100 that the M_{pop} lies between 62.37 and 65.63.

15.5.2 Statistical Inference Regarding Means of Small Samples

In case of small samples, the sampling distribution of means is not normal. It was in about 1815 when William Seely Gosset developed the concept of small sample size.

He found that the distribution curves of small sample means were some what different from the normal curve. This distribution was named as t-distribution. When the size of the sample is small, the t-distribution lies under the normal curve. Figure 2 shows that the t-distribution does not differ greatly from the normal curve unless the sample size is quite small. As the sample increases in size, the t-distribution approaches more and more closely to the normal curve.

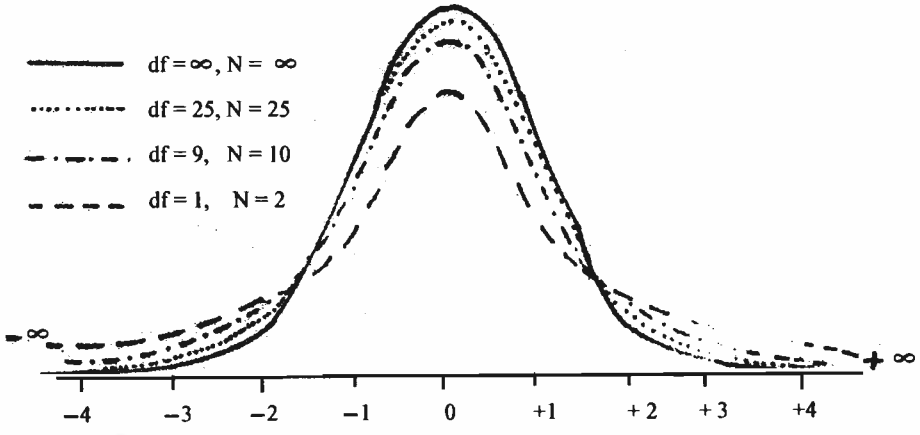


Fig. 15.2: Distribution of t for Degrees of Freedom from I to ∞ (when df is very large, the distribution of t approaches normal).

In case of small samples, we make use of t-table, given in the Appendix, using respective 'degrees of freedom' depending upon the size of the sample for getting t-values instead of 1.96 for .05 and 2.58 for .01 levels of confidence. You can observe from the t-table that the t-values approach z-values (1.96 or 2.58) of normal probability table as the sample size increases. For the application of t-table it is necessary for you to understand the concept of 'degree of freedom'.

The number of 'degrees of freedom' in a distribution is the number of observations or values that are independent of each other and which cannot be deduced from each other. When a mean is computed for data involving a number of observations, the sum of the measures is calculated and divided by N.

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

But computing a mean, 1 degree of freedom is used up or lost, and subsequent calculations of variance and standard deviation will be based on N–1 independent observations or N–1 degrees of freedom.

Let us understand the concept with the help of the following example, using original distribution (0) and altered distribution (A):

O Original Distribution	A Altered Distribution
+8	12
+7	9
+6	14
+5	8
+4	−13
$\sum x = 30$	$\sum x = +30$
N = 5	N = 5
Mean = +6	Mean = +6

These four terms can be altered in any way.

This term is dependent on, or determined by, the other four terms.

When the deviations are taken from the mean, the sum of the deviations is zero. Hence, in the altered distribution (A), the fifth term must have a value of –13 for the sum to equal +30, the mean to be +6, and the sum of the deviations from the mean is equal to zero. Thus the four terms, 12, 9, 14, 8 in the altered distribution are independent and can be altered. But one term is fixed which is deduced from the other four. There are N-1 i.e. (5-1) or 4 degrees of freedom.

When a sample statistics is used to estimate a population parameter, the number of degrees of freedom depends upon the restrictions placed. One df is lost for each restriction imposed. Thus the number of degrees of freedom (df) will vary from one statistics to another.

In estimating population mean (M_{pop}) from the sample mean (M), we lose 1 df and so the number of degrees of freedom is (N–1). By way of illustration, let us determine the .95 and .99 confidence intervals for the population mean (M_{pop}) of the scores 8, 8, 10, 7 and 9 achieved by 5 students in a test of mathematics. The mean of scores is:

$$= \frac{8+8+10+7+9}{5} = 8.40$$

When the number of cases in the sample is small (less than 30), the formula for computing standard deviation (S) is:

$$S = \frac{\sqrt{\sum x^2}}{\sqrt{N-1}}$$

in which

$\sum x^2$ = Sum of the squares of deviations from the mean.

N = Number of cases in the sample.

Using this formula, we compute the value of standard deviation as:

X	x = X – M	x^2
8	–0.40	0.16
8	–0.40	0.16
10	+1.60	2.56
7	–1.40	1.96
9	+0.60	0.36
M = 8.40	$\sum x = 0$	$\sum x^2 = 5.20$

$$S = \sqrt{\frac{\sum x^2}{N-1}}$$

$$S = \sqrt{\frac{5.20}{5-1}} = 1.14$$

The standard error of the mean (SE_M) is:

$$SE_M = \frac{S}{\sqrt{N}}$$

$$= \frac{1.14}{\sqrt{5}} = 0.51$$

For estimating the M_{pop} from the sample mean of 5.40, we determine the value of t at .05 and .01 points using appropriate number of degrees of freedom. The df for determining t are N–1 or 4. Entering t-table with 4 df, the value of t at 0.05 point is 2.78 and 4.60 at 0.01 point.

For t = 2.78, 95 of our 100 sample means will lie within ± 2.78 SE_M or ± 2.78 × 0.51 of the population mean and that 5 out of 100 fall outside these limits. The probability (P) is .95 that the sample mean 8.40 does not miss the M_{pop} by more than ± 2.78 × 0.51 or ± 1.92. The limits of the .95 confidence interval are 8.40 ± 2.78 × 0.51 or 6.48 and 10.32. The probability (P) is .95 that M_{pop} is not less than 6.48 and greater than 10.32.

For t=4.60, we know that 99 of our 100 sample means will lie between M_{pop} and ± 4.60 SE_M or ± 4.60 × 0.51 and that 1 out of 99 fall beyond these limits. The probability (P) is .99 that our sample mean of 8.40 does not miss the M_{pop} by more than ± 4.60 × 0.51 or ± 2.35. The limits of .99 confidence interval are 8.40 ± 4.60 × 0.51 or 6.05 and 10.75. The probability (P) is .99 that M_{pop} is not less than 6.05 and greater than 10.75.

Check Your Progress

Notes : a) Space is given below for your answer.

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

1. i) State the assumptions upon which parametric tests are based.

.....

.....

ii) Compute: (i) .95 and (ii) .99 confidence intervals for the population mean for a sample with mean = 30, SD = 5 and N = 100.

.....

.....

iii) Find the .99 confidence interval for the population mean for a sample of 5 students whose scores on a test are 6, 5, 7, 4, 5.

.....

15.6 TESTING THE STATISTICAL SIGNIFICANCE OF THE DIFFERENCE BETWEEN MEANS

In some research situations we require the use of a statistical technique to determine whether a true difference exists between the population parameters of two samples. The parameters may be means. standard deviations, correlations etc. For example, we wish to determine whether the population of male B.Ed. trainees enrolled with IGNOU differ from their female counterparts in their attitude towards teaching. In this case we would first draw samples of male and female B.Ed. trainees. Next, we would administer an attitude scale measuring attitude towards teaching on the selected samples, compute the means of the two samples, and find the difference between them. Let the mean of the male sample be 55 and that of the females 59. Then it has to be ascertained if the difference of 4 between the sample means is large enough to be taken as real and not due only to sampling error or chance. In order to test the significance of the obtained difference of 4, we need to first find

out the standard error of the difference of the two means because it is reasonable to expect that the difference between two means will be subject to sampling errors. Then from the difference between the sample means and its standard error we can determine whether a difference probably exists between the population means. In the following sections we will discuss the procedure of testing the significance of the difference between the means and correlations of the samples.

15.6.1 Significance of the Difference Between the Means of Two Independent Large and Small Samples

Means are said to be independent or uncorrelated when computed from samples drawn at random from totally different and unrelated groups.

Large Samples

You have learnt that the frequency distribution of large sample means, drawn from the same population, fall into a normal distribution around the population mean (M_{pop}) as their measure of central tendency. It is reasonable to expect that the frequency distribution of the difference between the means computed from samples drawn from two different populations will also tend to be normal with a mean of zero and standard deviation which is called the standard error of the difference of means. The standard error is denoted by σ_{dM} which is estimated from the standard errors of the two sample means, σ_{M1} and σ_{M2}. The formula is:

$$\sigma_{dM} = \sqrt{\sigma_{M1}^2 + \sigma_{M2}^2}$$

in which

σ_{M1} = SE of the mean of the first sample

σ_{M2} = SE of the mean of the second sample

N_1 = Number of cases in the first sample

N_2 = Number of cases in the second sample

To illustrate the application of this formula, let us consider a problem in which a teaching aptitude test was administered on two randomly selected groups of teachers, one of 32 males and the other 34 females. The data are summarized in the following table:

Statistics	Male Teachers	Female Teachers
N	32	34
Mean	87.43	82.58
Standard Deviation	6.27	6.39

For testing the significance of the difference between the means, 87.43 and 82.58, of the two groups, we first compute the standard error of the difference of these means. Using the formula of σ_{dM}, we have:

$$\sigma_{dm} = \sqrt{\frac{(6.27)^2}{32} + \frac{(6.39)^2}{34}}$$

$$= \sqrt{\frac{39.40}{32} + \frac{40.80}{34}}$$

$$= \sqrt{1.23 + 1.20}$$

$$= \sqrt{2.43} = 1.56$$

In order to find out whether the male and female teachers actually differ in teaching aptitude, it is needed to set up a 'Null-Hypothesis'. This type of hypothesis asserts that the difference between the population means of male and female teachers is zero and that, except for sampling errors, mean differences from sample to sample will all be zero. In accordance with the null hypothesis, it is assumed that the sampling distribution of differences is normal with the mean at zero, or at [M_{pop} (males) – M_{pop} (females)] = 0. The deviation of each sample difference, (M_{males} – $M_{females}$), from this central reference point is equal to [M_{males} – $M_{females}$] – [M_{pop} (males) – M_{pop} (females)], or [M_{males} – $M_{females}$] = 0. The deviation of each sampled difference given in terms of standard measure would be deviation divided by the standard error, which gives us a C.R. value (Z-value) in terms of a general formula:

$$CR = Z = \frac{M_1 - M_2}{\sigma dm}$$

$$CR = Z = \frac{87.43 - 82.58}{1.56} = \frac{4.85}{1.56} = 3.11$$

Since the CR value of 3.11 exceeds 2.58, the null hypothesis may be rejected at the .01 level of significance i.e. the mean teaching aptitude of male and female teachers differs significantly at .01 level.

For the sake of convenience, the researchers have .05 and .01 levels of significance as two arbitrary standards for accepting or rejecting a null hypothesis. When the value of CR or Z is 1.96 or more, we reject a null hypothesis at .05 level of significance. If the value is 2.58 or more, we reject the null hypothesis at .01 level and the probability (P) is that not more than once in 100 trials would a difference of this size arise if the true difference (M_{pop1} – M_{pop2}) were zero.

In testing a null-hypothesis, we make use of a two tailed test or a one-tailed test. If we set up a hypothesis that there is no difference, other than a sampling error difference between the means, we would be concerned only with a difference, and not in superiority or inferiority of either group. For testing this type of hypothesis, we apply a two tailed test since the difference between the obtained means may be as often in one direction (plus) as in the other (minus) from the true difference of zero; and in determining probabilities we take both tails of sampling distribution.

When we are hypothesizing a direction of difference, rather than the mere existence of a difference, we make use of a one tailed test. In determining probabilities, we take either the upper tail or the lower tail of the sampling distribution.

Small Samples

It has already been explained to you that the frequency distribution of small sample means drawn from the same population forms a t-distribution. It is reasonably expected that the sampling distribution of the difference between the means computed from small samples drawn from two different populations will also fall under the category of t-distribution.

The formula for testing a difference between means computed from independent or uncorrelated samples is:

$$t = \frac{|M_1 - M_2|}{\sqrt{\left(\frac{\sum x_1^2 + \sum x_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 N_2}\right)}}$$

in which:

M_1 and M_2 = means of the two samples

$\sum x_1^2 + \sum x_2^2$ = Sums of squares of the deviations from the means in the two samples

N1 and N2 = Number of cases in the two samples

In the use of t-test, the following conditions need to be fulfilled:

1. The population distribution should be normal. If the population distribution is badly skewed, t-test should not be used.
2. The samples under investigation should have the same variability. This condition is known as 'homogeneity of variances.'

To illustrate the use of t-test, let us consider the following data obtained from the two samples of elementary school in-service teachers enrolled in a distance teacher training programme, one of 10 males and the other of 12 females. The data in respect of means and sum of the square of deviations from means of the intelligence test scores of the samples are as under:

Teachers	Mean	$\sum x^2$	N
Males	9	20.44	10
Females	14	19.60	12

Using t-test:

$$t = \frac{|9-14|}{\sqrt{\left(\frac{19.60+20.44}{12+10-2}\right)\left(\frac{1+1}{12\times10}\right)}}$$

$$= \frac{5}{\sqrt{3.67}} = 8.26$$

Using a two-tailed test, the t critical values for rejection of the null hypothesis at .05 and .01 level of significance for (12 + 10 –2) or 20 degrees of freedom are 2.09 and 2.84 respectively. Since the obtained t value 8.26 is greater than 2.09, the null hypothesis is rejected at the .05 level for 20 degrees of freedom. We may conclude that the difference between the means of intelligence test scores of the male and female teachers is significant. The value 8.26 of t is also greater than 2.84, the null hypothesis is rejected at .01 level for 20 degrees of freedom concluding that there is a significant difference in the means of the two samples at the higher level of confidence.

15.6.2 Significance of the Differences Between the Means of Two Dependent Samples

Means are said to be dependent or correlated when obtained from the scores of the same test administered to the same sample upon two occasions, or when the same test is administered to equivalent samples in which the members of the group have been matched person for person, by one or more attributes.

When the same test is administered to the same sample on two occasions, the formula used for testing the significance of the difference between means obtained in the initial and final testing is:

$$t = \frac{|M_1 + M_2|}{\sigma M_1^2 + \sigma M_2^2 - 2 r_{12} \sigma M_1 \sigma M_2}$$

in which

M_1 and M_2 = Means of the scores of the initial and final testing.

σ_{M1} = Standard error of the initial test mean.

σ_{M2} = Standard error of the final test mean.

r_{12} = Correlation between the scores on initial and final testing

To illustrate the use of this formula, let us consider the scores of 12 sixth grade students on trial 1. After providing remedial instruction to the students the scores on trial 2 on the mastery learning test in mathematics were also obtained.

Trial 1	Trial 2
53	65
77	87
70	78
85	99
56	66
75	83
57	67
50	45
41	50
66	76
57	55
65	77

$M_1 = 62.66$ $M_2 = 70.66$

$\sigma_1 = 9.44$ $\sigma_2 = 15.15$

$\sigma_{M1} = 3.46$ $\sigma_{M2} = 4.37$

$r_{12} = 0.94$

Using formula:

$$t = \frac{|62.66 - 70.66|}{\sqrt{(3.46)^2 + (4.37)^2 - (2)(0.94)(3.46)(4.37)}}$$

$$= \frac{8.00}{\sqrt{2.6425}}$$

$$= \frac{8}{1.6255}$$

$$= 4.92$$

Since there are 12 students in the sample, we have 12 pairs of scores, the degrees of freedom (df) becomes 12 – 1 or 11. To test the hypothesis that remedial instruction has enhanced the achievement of students in mathematics, we would make use of a one tailed test. In the one-tailed test, for 11 df the .05 level is read from .10 column $\left(\frac{P}{2}=.05\right)$ of the t-table to be 1.80 and .01 level from .02 column $\left(\frac{P}{2}=.01\right)$ is 2.72.

Since our t of 4.92 is greater than 2.72, we may conclude that the gain from Trail 1 to Trial 2 is significant and the hypothesis that the remedial instruction in mathematics has increased the achievement is accepted.

Check Your Progress

Notes : a) Space is given below for your answer.

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

2. i) The means and standard deviations of two independent large samples for 95 rural and 64 urban seventh grade students of intelligence test scores are given below:

	Rural	Urban
N	95	64
Mean	24.5	23.4
Standard Deviation	5.12	4.07

Assuming that the samples are random, test the difference between means at .05 level of confidence.

.....

.....

.....

ii) The means and Σx^2 of achievement scores in respect of two dependent small groups matched for intelligence are given below. Compute t-value for testing the significance between the means.

Group 1	Group 2
$N_1 = 20$	$N_2 = 20$
$\sigma_1^2 = 54.76$	$\sigma_2^2 = 42.25$
$M_1 = 53.20$	$M_2 = 49.80$

$r_{12} = 0.60$

.....

.....

.....

.....

15.6.3 Significance of the Difference Between the Means of Three or More Samples

We compute CR and t-values to determine whether there is any significant difference between the means of two random samples. Suppose we have N (N > 2) random samples and we want to determine whether there are any significant differences among their means. For this we have to compute $\frac{N(N-1)}{2}$ values of t to determine the significance of the difference between N means by taking two means at a time. This procedure is time consuming and hence, it is advisable to use the technique of 'Analysis of Variance' which would make it possible to determine if any of the two of the N means differ significantly from each other by a single test, called F test, rather than computing $\frac{N(N-1)}{2}$ values of t.

Analysis of variance has the following basic assumptions underlying it which should be fulfilled in the use of this technique.

1. The population distribution should be normal. This assumption, however, is not especially important. Eden and Yates showed that even with a population departing considerably from normality, the effectiveness of the normal distribution still held.

2. All the groups of a certain criterion or of the combination of more than one criterion should be randomly chosen from the sub-population having the same criterion or having the same combination of more than one criterion. For instance, if we wish to select two groups in a population of B.Ed. trainees enrolled with IGNOU, one of males and the other of females, we must choose randomly from the respective sub-populations. The assumption of randomness is the key stone of the analysis of variance technique. There is no substitute for randomisation.

3. The sub-groups under investigation should have the same variability. This assumption is tested by applying F_{max} test.

$$F_{max} = \frac{\text{Largest Variance}}{\text{Smallest Variance}}$$

In analysis of variance, we have usually three or more groups i.e. there will be three or more variances. Unless the computed value of F_{max} equals or exceeds the appropriate F critical value at .05 level in the Table N of the Appendix, it is assumed that the variances are homogeneous and the difference is not significant.

Suppose we want to apply the technique of analysis of variance to test the significance of the difference between the means of the scores of a test of five groups, one group of greater variance (28.40) with 10 degrees of freedom and the other group of smaller variance (18.06) with 12 degree of freedom. The value of F_{max} is:

$$F_{max} = \frac{28.40}{18.60} = 1.57$$

Since the value of F_{max} 1.57 is less than the value 3.28 of F_{max} at .05 level in the Table N, it may be concluded that the variances fulfilled the condition of homogeneity. It may be noted the table of F_{max} used for testing the homogeneity of variances is different from the F table which is used for testing the significance of the difference between the means of the samples.

Analysis of Variance for Randomized Group Design (One-Way Analysis of Variance)

In an analysis of variance for a randomized group design' (one-way analysis of variance), we examine the relationship between one independent and one dependent variable. The analysis consists of following operations:

i) The variance of the measures (scores) for the randomized groups are combined into one composite group, known as the 'total groups variance' (V_t).

ii) The mean value of the variances of each of the three groups computed separately, is known as the 'within-groups variances' (V_w).

iii) The 'difference between the total groups variance and within-group variance' is known as the 'between-groups variance' ($V_t - V_w = V_b$).

iv) The F-ratio is computed as:

$$F = \frac{V_b}{V_w} = \frac{\text{Between Groups Variance}}{\text{Within Groups Variance}}$$

The within-groups variance represents the sampling error in the distributions. It is also referred to as the error variance or residual. The between-groups variance represents the influence of the variable under study or the experimental variable. If the 'between-groups variance' is not substantially greater than the 'within-groups variance', we may conclude that the difference between the means is probably due to sampling error. If the F-ratio is substantially greater than one, it means that the ratio of the 'between-groups variance' and 'within-groups variance' is probably too large to attribute to sampling error.

The critical values of the F-ratio for .05 and .01 levels of significance are read from the F-table, given in the Appendix. The table indicates two different degrees of freedom, one for V_b (the numerator) and one for V_w (the denominator). The degrees of freedom for the within a groups variance (V_w) is determined in the same way as it is for the t-test i.e., the sum of the subjects (individuals) for all the sample groups minus the number of the groups. If there are K number of groups with number of subjects as N_1, N_2, N_3, N_4 ..... in each group, then ($N_1 + N_2 + N_3 + N_4 + \ldots\ldots - K$) represents the degrees of freedom for within groups variance.

Let us illustrate the computation of F ratio with help of following example. In a remedial instruction programme, 5 subjects were randomly assigned to each of the three groups. Each group was administered the same mastery or criterion test but under slightly different remedial treatments. The scores of the three groups are presented as under:

Group 1		Group 2		Group 3	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
14	196	16	256	12	144
11	121	16	256	19	361
17	289	12	144	20	400
20	400	21	441	24	576
18	324	16	256	14	196
$\sum X_1 = 80$	$\sum X_1^2 = 1330$	$\sum X_2 = 81$	$\sum X_2^2 = 1353$	$\sum X_3 = 89$	$\sum X_3^2 = 1677$

To test the hypothesis that the groups do not differ in mean performance, let us apply the technique of analysis of variance.

The basic assumptions underlying the technique are tested as under:

1. Assumption of Normality

In the light of the finding of Eden and Yates discussed earlier, the assumption of normality may not be considered important for the data of the study.

2. Assumption of Randomness

The requirement of randomness has been amply fulfilled in this study as the subjects were randomly assigned to three treatment groups from the same pool sample.

3. Assumption of Homogeneity of Variance

The F critical value exceeds F max value for the data, it may be assumed that the variances are homogeneous.

After testing the data for the basic assumptions we may proceed with the computations for the analysis of variance using the following steps:

Step 1 : Correction $= \frac{\left(\sum x\right)^2}{N} = \frac{\left(\sum x_1 + \sum x_2 + \sum x_3\right)^2}{N}$

$$= \frac{(80+81+89)^2}{15}$$

$$= \frac{(250)^2}{15}$$

$$= 4166.67$$

Step 2 : Total Sum of Squares (Total SS_t)

$$SS_t = \Sigma x_1^2 + \Sigma x_2^2 + \Sigma x_3^2 - \text{Correction}$$

$$= 1330 + 1353 + 1677 - 4166.67$$

$$= 4360 - 4166.67$$

$$= 193.33$$

Step 3 : Sum of Squares Between Means of Treatment (SS_b)

$$SS_b = \frac{\left(\sum x_1\right)^2}{N_1} + \frac{\left(\sum x_2\right)^2}{N_2} + \frac{\left(\sum x_3\right)^2}{N_3} - \text{Correction}$$

$$= \frac{(80)^2}{5} + \frac{(81)^2}{5} + \frac{(89)^2}{5} - 4166.67$$

$$= 1280 + 1312.2 + 1584.2 - 4166.67$$

$$= 4176.40 - 4166.67$$

$$= 9.73$$

Step 4: Sum of Squares Within Treatments (SS_w)

$$SS_w = SS_t - SS_b$$

$= 193.33 - 9.73$

$= 183.60$

Step 5 : Calculation of Variances for Each SS and Analysis of the Total Variance into Components

Each SS becomes a variance when divided by the degrees of freedom (df) allotted to it. There are 15 scores in all and hence there are (N−1) or (15−1) or 14 df in all. These 14 df are allotted in the following way:

If N = number of scores in all and K = number of groups, we have df for total sum of squares (SS_t) = 15 − 1 = 14; df_w for within treatments (SS_w) = 15 − 3 = 12 and df_3 for between means of treatments (SS_b) = K − 1 = 3 − 1 = 2.

To find the mean square between (MS_b) and mean square within (MS_w), we divide the sum of squares between (SS_b) and the sum of squares within (SS_w) by their respective degrees of freedom (df).

$$F = \frac{MS_b}{MS_w} = \frac{SS_b}{df_b} \Big/ \frac{SS_w}{df_w}$$

$$= \frac{\frac{9.73}{2}}{\frac{183.60}{12}}$$

$$= \frac{4.865}{15.30}$$

$$= 0.318$$

Summary : Analysis of Variance of Three Groups

Source of Variance	df	Sum of Squares (SS)	Mean Square (Variance)	F
Between Groups	2	9.73	4.865	0.318
Within Groups	12	183.60	15.30	
Total	14	193.33		

In our problem the null hypothesis asserts that three sets of scores are in reality the score of three random samples drawn from the same population, and that the means of the treatments of the three groups 1, 2 and 3 differ only through the fluctuations of sampling. To test this hypothesis we divided the 'between means' variance by the 'within treatments' variance. The resulting variance ratio, called F, is to be compared with the F values in the F table presented in the Appendix. The F value in our problem is 0.318 and the df are 2 for the numerator (df_1) and 12 for the denomination (df_2). Entering the F table we read from column 2 and row 12 that an F of 3.88 is significant at .05 level and F of 6.93 is significant at the .01 level. Since the obtained F value of 0.318 is less than the table values, we accept the null hypothesis and conclude that the means of three groups do not differ significantly.

Analysis of Variance for Factorial Design

In many experimental studies, we are concerned with the effect of two or more independent variables, usually called **factors,** on a dependent variable. The number

of ways in which a factor is varied is called the number of levels of the factor. A factor that is varied in two ways is said to have two levels and a factor that is varied in three ways is said to have three levels. With two or more factors each with two or more levels; a treatment in some experiments consists of a combination of one level for each factor.

In such multiple classification or factorial designs, both the independent and interactive effects of two or more independent variables on one dependent variable are analysed. The total variance is divided into more than two parts: one part for each independent variable (main effect), one part for each interaction of two or more independent variables, and one part for the residual, or within-group variance. Thus, in a design with two independent variables, the variance is divided into four parts. For example, in case of a three group randomized design we could also divide each of the three treatment groups into males and females. We then have a factorial design with two independent variables, treatments and sex. Since there are three treatment conditions and two conditions of sex, this is a 3 × 2 factorial design. In the analysis of variance of this design, the variance is divided into four parts: (i) the main effect of treatments; (ii) the main effect of sex; (iii) the interaction effect of treatments with sex; and (iv) residual. Thus we are able to compute three sepearate values for F; one to test the difference among the treatments, one to test the difference between males and females, and one to test the interaction of sex and treatments.

With the availability of computers, analysis of variance can be used with any number of independent variables. However, the analysis does not take into account the correlation between the dependent variables and the pertinent intervening variables. For example, in the studies of memory and learning, for certain reasons it is not possible to equate experimental groups on some pertinent intervening variables at the start of the experiment. In such situations, we may use the technique of '**analysis of co-variance**'.

Analysis of co-variance is an extension of analysis of variance that tests the significance of the difference between means of the final experimental data by taking into account the correlation between the dependent variable and one or more co-variates or pertinent control variables, and by adjusting initial mean differences in the groups. The technique is particularly appropriate when the subjects in two or more groups are found to differ on a pretest of other initial variable. In such a situation, the effects of the pretest and other relevant variables are partialled out and the resulting adjusted means of the posttest scores are compared. The initial status of the experimental groups may be determined by the pretest scores in a pretest – post-test study, or in post-test only studies, by such measures as previous knowledge of subject matter, intelligence, academic achievement etc. The differences in the initial status of groups in terms of such pertinent variables can be removed statistically so that they can be compared as though their intial status had been equated.

To illustrate, suppose in an experiment the dependent variable was the performance in mathematics of three groups of seventh grade distance learners enrolled with National Institute of Open Schooling. One group was taught through computer assisted instruction, second through PSI and the third through conventional method. The three groups were framed by randomly assigning 10 subjects to each of the groups and subjects were equated on the pre-test scores on the criterion (mastery) test in mathematics. Since intelligence was also considered to be a significant factor related to the scores on the criterion test, the scores on the intelligence test were also taken into account so as to provide a statistical control for testing the significance of the difference between the means of criterion test of the three groups at the end of the experiment.

Check Your Progress

Notes : a) Space is given below for your answer.

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

3. i) List the assumptions of analysis of variance.

.....

.....

ii) Thirty subjects were randomly assigned to three groups. Each group performed the same task under three experimental treatments. The scores of the subjects on the task are as under:

Group I: 26, 27, 18, 22, 23, 19, 27, 26, 24, 26

Group II: 18, 22, 18, 23, 19, 24, 20, 21, 19, 25

Group III: 18, 14, 15, 14, 19, 21, 17, 17, 18, 19

Apply analysis of variance and test the significance of the difference between the means of the three groups.

.....

.....

15.7 STATISTICAL INFERENCE REGARDING PEARSON'S CO-EFFICIENT OF CORRELATION

The mathematical basis for standard error of a Pearson's co-efficient of correlation (r) is rather complicated because of the difficulty in its nature of sampling distribution. The sampling distribution of r is not normal except when population r is near zero and size of the sample is large (N = 30 or greater). When r is high (0.80 or more) and N is small, the sampling distribution of r is skewed. It is also true when r is low (0.20 or less).

In view of this, a sound method for making the inference regarding Pearson's r, especially when its magnitude is very high or very low, is to convert r into Fisher's Z coefficient using conversion table provided in the Appendix and find the standard error (SE) of Z. The sampling distribution of Z co-efficient is normal regardless of the size of sample N and the size of the population r. Furthermore, the SE of Z depends only upon the size of sample N.

The formula for standard error of Z (σ_z) is:

$$SE_z = \sigma_z = \frac{1}{\sqrt{N-3}}$$

To illustrate, let us consider the correlation of 0.55 between scores of an achievement test in science obtained from a random sample of 39 female students of sixth grade.

Using the formula of standard error of Z, we have:

$$SE_z = \sigma_z = \frac{1}{\sqrt{39-3}} = \frac{1}{6} = 0.17$$

For the r = 0.55, the corresponding value of Z is 0.62. Since the sampling distribution of Z is normal, the confidence interval at .95 level for the population of true Z is Z $\pm \sigma_Z \times 1.96$ i.e. $0.62 \pm 0.17 \times 1.96$ or 0.29 and 0.95. The corresponding r's are 0.28 and 0.74, which give a well estimated interval within which we expect the population r with .95 confidence. The chances are 95 in 100 that population r lies between 0.28 and 0.74. For a higher degree of confidence we can take .99 level, for which limits are $Z \pm \sigma_Z \times 2.58$ i.e. $0.62 \pm 0.17 \times 2.58$ or 0.18 and 1.06. The corresponding r's from the conversion table are 0.18 and 0.78. The chances are 99 in 100 that population r lies within 0.18 and 0.78.

The significance of r is also tested by using the formula:

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

With N – 2 degrees of freedom, a co-efficient of correlation is judged as statistically significant when t_r value equals or exceeds the t critical value in the t table.

To illustrate, for r = 0.60 and N = 27:

$$t_r = \frac{(0.60)\left(\sqrt{27-2}\right)}{\sqrt{1-(0.60)^2}}$$

$$= \frac{(0.60)(5)}{\sqrt{1-36}} = 3.75$$

The obtained t_r value of 3.75 exceeds the t critical values of 2.06 and 2.79 at .05 and .01 levels for 25 degrees of freedom which indicates that the correlation of 0.60 is significant.

We can also test the significance of r directly by using the Table K presented in the Appendix. This table presents critical values of r which can be read directly without computing the t value using N – 2 degrees of freedom.

15.8 SIGNIFICANCE OF THE DIFFERENCE BETWEEN PEARSON'S CO-EFFICIENTS OF CORRELATION OF TWO INDEPENDENT SAMPLES

The method of determining the standard error of the difference between Pearson's co-efficients of correlation of two samples is first to convert the r's into Fisher's Z co-efficients and then to determine the significance of the difference between the two Z's.

When we have two correlations between the same two variables, X and Y, computed from two totally different and unmatched samples, the standard error of a difference between two corresponding Z's is computed by the formula:

$$SE_{DZ} = \sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

in which

N_1 and N_2 = Sizes of the two samples

The significance of the difference two Z's is tested with the help of formula:

$$CR = \frac{Z_1 - Z_2}{SE_{DZ}}$$

Illustrating the use of the formula, let us consider the correlations of 0.80 and 0.85 computed between teaching aptitude and teaching attitude scores on the basis of the data obtained from two groups, one males and other females, of B.Ed. trainees enrolled with IGNOU. The data are presented as under:

Male B.Ed. Trainees	Female B.Ed. Trainees
$N_1 = 208$	$N_2 = 220$
$r_1 = 0.80$	$r_2 = 0.85$

To test the significance of the difference between the two r's, we have to first convert these r's into Z-co-efficients. The corresponding Z-coefficients from the Table F given in the Appendix are 1.10 and 1.26 respectively.

Using formula of SE_{DZ}:

$$SE_{DZ} = \sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{208-3} + \frac{1}{220-3}}$$

$$= \sqrt{\frac{1}{205} + \frac{1}{217}}$$

$$= \sqrt{.0048 + .0046}$$

$$= \sqrt{.0094}$$

$$= .097$$

$$CR = \frac{|1.10 - 1.26|}{.097}$$

$$= 1.65$$

The obtained CR value is less than 1.96, the difference between $Z_1 = 1.10$ and 1.26 is not significant at .05 level which indicates that the difference between the corresponding correlations 0.80 and 0.85 is also not significant. We may infer that the correlation between teaching aptitude and teaching attitude does not differ in the two groups drawn from two populations of male and female B.Ed. trainees.

Check Your Progress

Notes : a) Space is given below for your answer.

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

4. i) List the advantages of converting Pearson's r into Fisher's Z.

.....

.....

.....

.....

ii) Find the .95 confidence interval for population Pearson's correlation co-efficient of 0.78 obtained from a random sample of 84 cases.

.....

.....

iii) Test the significance of r for the following :

r = 0.40

N = 25

.....

.....

15.9 LET US SUM UP

1. In this unit we discussed the statistical inference based on parametric tests. It included the assumptions on which the use of parametric tests are based; inferences regarding means of large and small samples; significance of the difference between the means of two large and small independent samples; significance of the difference between means of the two dependent samples; significance of the difference between means of three or more samples; significance of Pearson's coefficients of correlation; and significance of the difference between Pearson's coefficients of correlation of two independent samples.

2. Parametric statistical tests are based on certain assumptions about the nature of distribution and types of measurement scales used in the collection of data. These tests should be used: (i) when the data are expressed in interval or ratio scales of measurement; (ii) the population values are normally distributed; (iii) the samples have equal or nearly equal variances (homogeneity of variances); and (iv) selection of cases is random.

3. Statistical inferences about sample means are not made with certainty but are based upon probability estimates making use of 'Central Limit Theorem', standard error of the mean, null hypothesis, levels of significance, and one-tailed and two-tailed tests.

4. In testing the significance of the difference between two means, we make use of Z or CR value or t-test depending upon the size of samples.

5. F test is used for testing the significance between the means of three or more sampling. It involves the use of analysis of variance or analysis of co-variance.

6. Analysis of variance is based on the assumptions of: (i) randomness; (ii) homogeneity of variances and (iii) normality. However, the assumption of randomness is the key stone of this technique; failure to fulfil this assumption gives biased results.

7. For testing the significance of Pearson's r, we make use of Fisher's Z transformation or t-test.

8. The inference regarding the significance of the difference between Pearson's r based on two independent sample is made by computing the value of Z using Fisher's Z transformation.

15.10 UNIT-END ACTIVITIES

1. List the assumptions on which the use of Parametric Tests is based.
2. Describe the characteristics of Central Limit Theorem.
3. Define the standard error of mean.
4. Define the confidence intervals and levels of confidence.
5. Given a sample with mean = 35.80, SD = 6.35 and N = 100 Compute the .95 and .99 confidence intervals for the true mean.
6. The mean of 12 independent observations of a test is 80 and SD is 14. Compute the .99 confidence intervals for the true mean.
7. In a study of attitude measurement, a sample of 120 rural school teachers and a sample of 130 school urban teachers scored as below:

Number	Rural Teachers	Urban Teachers
112	115	14.52
108	119	9.81

Assuming that our samples are random, is the difference between means significant at .01 level.

8. Ten subjects were given 4 successive trials upon a mastery test in physics, of which only the scores for trails 1 and 4 one shown below. Test the hypothesis that remedial programmes between the trials have increased the test scores from initial to final trials significantly:

Trial 1 : 80, 72, 81,56, 65, 72, 80, 71, 60, 85, 82, 68

Trail 4 : 91, 85, 79, 62, 71, 84, 91, 82, 71, 87, 91,72

9. Apply Analysis of Variance and test the significance of the difference between the means of scores of the following five groups.

Group 1	Group 2	Group 3	Group 4	Group 5
27	41	34	19	41
37	29	37	20	42
24	35	51	25	38
20	62	52	31	39
18	36	39	38	37

10. State the Assumptions of Analysis of Variance.
11. The Pearson's coefficient of correlation between the height and weight of 100 tenth grade boys is 0.85 and that of 175 tenth grade girls is 0.73. Is the difference between the correlations significant at .01 level?

15.11 SUGGESTED READINGS

Best, John and James V. Kahn (1992): *Research in Education.* New Delhi: Prentice Hall of India Pvt. Ltd.

Edwards, A.L. (1971): *Experimental Design in Psychological Research.* New Delhi: Amerind Publishing Co. Pvt. Ltd. Indian Edition.

Garrett, H.E. (1962): *Statistics in Psychology and Education.* Bombay: Allied Pacific Pvt. Ltd.

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

Koul, Lokesh (1997): *Methodology of Educational Research.* Third Revised Edition, New Delhi: Vikas Publishing House Pvt. Ltd.

15.12 ANSWERS TO CHECK YOUR PROGRESS

1. i) Parametric tests should be used when the:
   a) Variables described are expressed in interval or ratio scales;
   b) Samples have equal or nearly equal variances; and
   c) Observations are independent.
   ii) a) 29.02 and 30.98
   b) 28.71 and 31.29
   iii) 3.05 and 7.70
2. i) CR = 1.51, not significant at .05 level.
   ii) t = 2.43, significant at .05 level for 19 df.
3 i) a) The population distribution should be normal.
   b) The sub-groups under investigation should have the same variability.
   c) The groups should be selected at random from the same population.
   ii) F = 14.77, which is significant at .01 level.
4. i) a) The sampling distribution of Z - coefficient is normal regardless of the size of sample N.
   b) The standard error of Z depends only upon the size of sample N.
   ii) 0.68 and 0.85.
   iii) t_r = 2.08, which is not significant at .05 level. Hence r = 0.40 is not significant.

UNIT 16 ANALYSIS OF QUANTITATIVE DATA: INFERENTIAL STATISTICS BASED ON NON-PARAMETRIC TESTS

Structure

16.1 Introduction

16.2 Objectives

16.3 Non-parametric Tests

16.4 Statistical Inference Based on Non-parametric Tests: Unrelated Samples

16.4.1 The Chi Square (χ^2) Test

16.4.2 The Median Test

16.4.3 The Mann-Whitney U Test

16.5 Statistical Inference Based on Non-Parametric Tests: Related Samples

16.5.1 The Sign Test

16.5.2 The Wilcoxon Matched – Pairs Signed – Ranks Test

16.6 Statistical Inference Regarding Correlations Using Non-parametric Data

16.6.1 Significance of Sprearman's Rho (ρ) Correlation Coefficient

16.6.2 Significance of Phi (ϕ) Correlation Coefficient

16.6.3 Signifance of Contingency Coefficient (C)

16.7 Let Us Sum Up

16.8 Unit-end Activities

16.9 Suggested Readings

16.10 Answers To Check Your Progress

16.1 INTRODUCTION

In the previous unit you learnt about the use of parametric tests in making inferences about the means computed from large and small samples. The use of Z and T tests were also explained to you for testing the significance of the difference between the means of two large and small samples. The application and use of analysis of variance and co-variance for testing the differnce between the means of three or more samples were also discussed with the help of examples. The signifance of the Pearson's co-efficient of correlation using Fisher's Z conversion were also explained alongwith the the use of Z test for testing the signifance of the difference between Pearson's coefficients of correlation computed from two samples.

In the use of parametric tests for making statistical inferences, we need to take into account certain assumptions about the nature of the population distribution, and also the type of the measurement scale used ot quantify the data. In this unit you will learn about another category of tests which do not make stringent assumptions about the nature of the population distribution. This category of test is called distribution free or non-parametric tests. The use and application of various of non-parametric tests involving unrelated and related samples will be explained in this unit. These would include chi-square test, median test, Man-Whitney U test, sign test and Wilcoxon-matched pairs signed-ranks test.

16.2 OBJECTIVES

After studying this unit, you will be able to:

- explain the nature of non-parametric tests;
- state the use of non-parametric tests;
- draw statistical inference pertaining to unrelated samples using the: (i) chi-square test: (ii) median test; and (iii) man-whitney U test;
- make statistical inferences pertaining to related samples using the: (i) sign test and (ii) wilcoxon matched-pairs signed-ranks test; and
- test statistical signifance of spearman's correlation co-efficient, phi-co-efficient and contingency co-efficient.

16.3 NON-PARAMETRIC TESTS

In the last unit you learnt that parametric tests are generally quite robust and are useful even when some of their mathematical assumptions are violated. However, these tests are used only with the data based upon ratio or interval measurements. In case of counted or ranked data, we make use of non-parametric tests. It is argued that non-parametric tests have greater merit because their validity is not based upon assumptions about the nature of the population distribution, assumptions that are so frequently ignored or violated by researchers using parametric tests. It may be noted that non-parametric tests are less precise and have less power than the parametric tests.

The use of non-parametric tests do not make numerous or stringent assumption about the nature of the population distribution and hence they are called distribution free tests.

Non-parametric are used when:

1. The nature of the population from which samples are drawn is not known to be normal.
2. The variables are expressed in nominal form.
3. The data are measures which are ranked or expressed in numerical scores which have the strength of ranks.

16.4 STATISTICAL INFERENCE BASED ON NON-PARAMETRIC TESTS : UNRELATED SAMPLES

The most frequently non-parametric tests which are used in drawing statistical inferences in case of unrelated or independent samples are: (1) chi square test; (ii) median test; and (iii) man-whitney test. The use and application of these tests are discussed below:

16.4.1 The Chi Square (χ^2) Test

The chi square test is applied only to discrete data. The data that are counted rather than measured. It is a test of independence and is used to estimate the likelihood that some factor other chance accounts for the observed relationship. The chi square (χ^2) is not a measure of the degree of relationship between the

variables under study, the chi square test merely evaluates the probability that the observed relationship results from chance. The basic assumption, as in case of other statistical significance, is that the sample observations have been randomly selected.

The formula for chi-square (χ^2) is:

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

In which

fo = frequency of occurrence of observed or experimentally determined facts.

fe = expected frequency of occurrence.

To evaluate the significance of chi square, we use the Table E of chi square presented in the Appendix with computed value of chi square and the appropriate number of degrees of freedom (df). The number of df = (r–1) (c–1) r is number of rows and c is the number of columns, in which the data are tabulated.

To illustrate the use of formula, let us consider the following data based on the judgements of 390 judges.The judgements have been classified into five categories taken to represent a continuum of opinion:

Categories

	I	II	III	IV	V	**Total**
Judgements	55	63	82	93	57	350

The hypothesis to be tested is 'equal probability hypothesis 'i.e. whether the judgments expressed in five categories differ significantly or not. For this we have to compute the distribution of answers to be expected on the equality or null hypothesis. Since the total judgements are 350 and the number of categories is 5, the expected judgements in each category would be 350/5 = 70. The data in respect of observed (fo) and expected frequencies (fe) alongwith the values (fo–fe), (fo–fe) 2 etc. can be arranged as under

	Categories					**Total**
	I	II	III	IV	V	
Observed Judgements (fo)	55	63	82	93	57	350
Expected Judgements (fe)	70	70	70	70	70	350
(fo–fe)	15	7	12	23	13	
$(fo-fe)^2$	225	49	144	529	169	
$\frac{(fo-fe)^2}{fe}$	3.21	0.70	2.06	7.56	2.41	

$$\chi^2 = \sum\left(\frac{(fo-fe)^2}{fe}\right)$$

$= 3.21 + 0.70 + 2.06 + 7.56 + 2.41$

$= 15.94$

The degrees of freedom in the table may be calculated from the formula df = (r–1) (c–1) to be (5–1) (2–1) or 4.

Using chi square Table E in the Appendix, we find in row df = 4, χ^2 of 9.488 in the column headed .05. Since the obtained value of χ^2 = 15.94 is greater than the table value of 9.488, we reject the equal judgement hypothesis and conclude that judgements in terms of various categories differ significantly.

Suppose instead of the 'hypothesis of equality', we may wish to test the data expressed in various judgement categories against the hypothesis of a normal distribution. In that case our hypothesis may assert that the judgement frequencies which we have observed really follow the normal distribution instead of being equally probably. Using the data of the above example, we have to find out how many of the 350 (total of the categories of judgements) may be expected to fall in each categories on the hypothesis of a normal distribution. These are found by first dividing the base line of a normal curve (taken to extend over 6 σ) into 5 equal segments each of 1.20 σ each. From the normal table (Table A) of the Appendix, the proportion of the normal distribution to be found in each of these segments would be as follows:

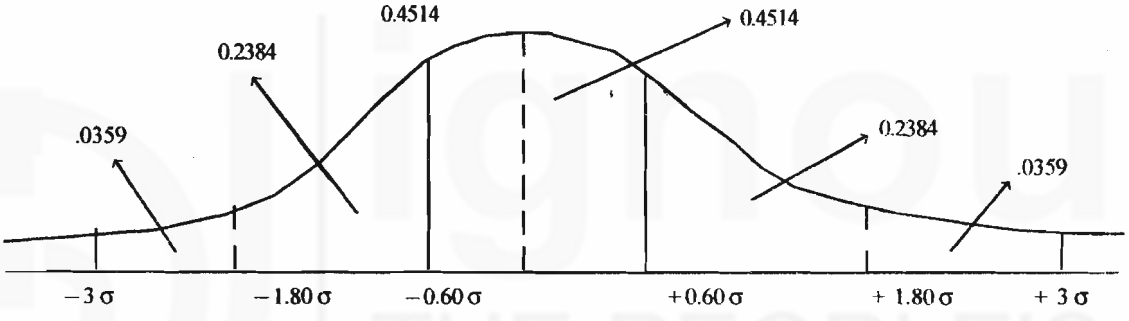


Fig.16.1: Normal Distribution of judgment frequencies.

Between + 3.00 σ and + 1.80 σ = .0359

+ 1.80 σ and + 0.60 σ = .2384

– 0.60 σ and – 0.60 σ = .4514

– 0.60 σ and – 1.80 σ = .2384

– 1.80 σ and – 3.00 σ = .0359

These proportions of 350 have been calculated as 12.56, 83, 44, 157.99, 83.44 and 12.56 and are entered in the row fe in following table:

	Categories					Total
	I	II	III	IV	V	
(fo)	55	63	82	93	57	350
(fe)	12.56	83.44	158.00	83.44	12.56	350
(fo–fe)	42.44	20.44	76	9.56	44.44	
$(fo-fe)^2$	1801.15	417.79	5776	91.39	1974.91	
$\frac{(fo-fe)^2}{fe}$	143.40	5.01	36.56	1.10	157.24	

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

$= 143.40 + 5.01 + 36.56 + 1.10 + 157.24$

$= 343.31$

The value of χ^2 in the Table E in the Appendix is 9.488 for df = 4 in the column headed by .05, which is less than the computed χ^2 value of 343.31. The difference between observed and expected values is so great that the hypothesis of normal distribution of judgement categories must be rejected.

Let us use chi square test to the data, which represent the number of boys and the number of girls who chose each of the three possible answers to an item on a personality inventory, to test whether the item differentiates significantly between boys and girls.

	Yes	No	Undecided	Total
Boys	14	66	10	90
Girls	27	66	7	100
Total	41	132	17	190

For each of the observed frequency in the table, let us compute the expected frequency in the following way:

Row I (Boys): $\frac{41\times 90}{190} = 19.42$; $\frac{132\times 90}{190} = 62.53$; $\frac{17\times 90}{190} = 8.05$

Row 2 (Girls): $\frac{41\times 100}{190} = 21.58$; $\frac{132\times 100}{190} = 62.47$; $\frac{17\times 100}{190} = 8.95$

The data in respect of observed and expected frequencies are arranged in the following table. The values in parentheses within the different cells are expected frequencies.

Responses

	Yes	No	Undecided	Total
Boys	14 (19.42)	66 (62.53)	10 (8.05)	90
Girls	27 (21.58)	66 (69.47)	7 (8.95)	100
Total	41	132	17	190

The hypothesis to be tested is the null hypothesis namely, that the item does not differentiate between the groups of boys and girls.

Using the formula of χ^2

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

$$= \frac{(14-19.42)^2}{19.42} + \frac{(66-62.53)^2}{62.53} + \frac{(10-8.05)^2}{8.05}$$

$$= \frac{(27-21.58)^2}{21.58} + \frac{(66-69.47)^2}{69.47} + \frac{(17-8.95)^2}{8.95}$$

$$= 1.51 + 0.19 + 0.47 + 1.36 + 0.17 + 7.24$$

$$= 10.94$$

$$df = (r-1)(c-1) = (2-1)(3-1) = 2$$

The χ^2 critical values for 2 df as given in the Table E are 5.991 and 9.210 respectively for .05 and .01 levels of signifance and the obtained value 10.94 of χ^2 is higher than these values. This indicates that the item of the personality inventory differentiate between boys and girls and the null hypothesis is rejected.

Check Your Progress

Notes : a) Space is given below for your answer.

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

1. The following judgements were classified into six categories taken to represent a continuum of opinion:

	Categories						
	I	II	III	IV	V	VI	**Total**
Judgements	48	61	82	91	57	45	384

Test the given distribution versus normal distribution hypothesis.

.....

.....

2. The following table represents the number of boys and the number of girls who choose each of the possible answers to an item in an attitude scale.

	Strongly Approve	Approve	Indifferent	Disapprove	Strongly Disapprove	**Total**
Boys	25	30	10	25	10	100
Girls	10	15	5	15	15	60
Total	35	45	15	40	25	160

Do these date indicate a significant sex difference in attitude towards this question?

.....

.....

16.4.2 The Median Test

The median test is used for testing whether two independent samples differ in central tendencies. It gives information as to whether it is likely that two independent samples have been drawn from populations with the same median. It is particularly useful when even the measurements for the two samples are expressed in an ordinal scale.

In using the median test, we first calculate the combined median for all measures (scores) in both samples. Then both sets of scores at the combined median are dichotomized and the data are set in a 2×2 table presented below:

Table for Use of Median Test

	Group I	Group II	Total
No. of measures (scores) above combined Median	A	B	A + B
No. of measures (scores) below combined Median	C	D	C + D
Total	A + C	B + D	

Under the null hypothesis, we would expect about half of each group's (scores) to be above the combined median and about half to be below , that is, we would expect frequencies A and C to be about equal, and frequencies B and D to be about equal. In order to test this hypothesis, we calculate $\aleph^2$ using the following formula:

$$\aleph^2 = \frac{N\left[\left|AD-BC\right|-\frac{N}{2}\right]^2}{(A+B)(C+D)(A+C)(B+D)}$$

Let us illustrate the use of this formula with the help of the following example:

Twenty male and fifteen female teacher educators of a teacher training institute were asked to express their attitude towards teacher education programmes offered through distance mode at the B. Ed. Level. Both the groups were administrated an attitude scale and common median attitude score was computed. The number of cases from both groups falling above and below the median score is shown in the following table:

Distribution of Male and Female Teacher Educators Below and Above the Common Median Attitude Score

	Below Median	Above Median	Total
Female Teachers Educators	9	6	15
Male Teachers Educators	6	14	20
Total	15	20	35

Using the formula for $\aleph^2$

$$\aleph^2 = \frac{35\left[\left|(9)(14)-(6)(6)\right|-\frac{35}{2}\right]^2}{(15)(20)(15)(20)}$$

$$= \frac{35(\left|126-36\right|-17.5)^2}{(300)(300)}$$

$$= \frac{35(\left|126-36\right|-17.5)^2}{90000}$$

$$= \frac{35(90-17.5)^2}{90000}$$

$$= \frac{35(72.5)^2}{90000}$$

$$= 2.044$$

The critical χ^2 value for (2–1) (2–1) df or 1 df is 3.84 for. .05 level. Since the obtained χ^2 value of 2.044 is less than 3.84, the null hypothesis is retained and we may conclude that there is no difference in the attitude of male and female teacher educators towards teacher education programmes at the B.Ed. Level.

16.4.3 The Mann-Whitney U Test

The Mann-Whitney U test is more useful than the Median test. It is a most useful alternative to the parametric t test when the parametric assumptions cannot be met and when the measurements are expressed in ordinal scale values.

Suppose N_1 is the number of individuals in one of the two independent groups and N_2 the number of the individuals in the other. In using Mann-Whitney U test, we first combine the measures or scores from both groups, and rank these in order of increasing size. In this ranking, we have to consider the algebraic sign, that is, the lowest ranks are assigned to the largest negative numbers, if any. The ranks of each sample group are then summed individually and represented as ΣR_1 and ΣR_2.

There are two Us: U1 and U2 which are calculated using the following formula:

$$U_1 = N_1 N_2 + \frac{N_1 (N_1 + 1)}{2} - \sum R_1$$

$$U_2 = N_1 N_2 + \frac{N_2 (N_1 + 1)}{2} - \sum R_2$$

in which:

N_1 = number in one group

N_2 = number in second group

$\sum R_1$ = sum of ranks in one group

$\sum R_2$ = sum of ranks in second group

The two U's are related by the equation:

$$U_1 = N_1 N_2 - U_2$$

Thus only one U needs to be calculated, for the other can be easily determined by this education.

The Z value of U can be computed by the formula:

$$Z = \frac{U - \frac{N_1 N_2}{2}}{\sqrt{\frac{(N_1)(N_2)(N_1 + N_2 + 1)}{12}}}$$

It does not matter which U (the larger or smaller) is used in the computation of Z. The sign of Z will depend on which U is used, but the numerical value will be identical.

The following example used by Koul (1997) illustrated the application of Mann-Whitney U test in which a researcher wished to evaluate the effectiveness of micro-teaching and simulation in developing certain teachings skills among student-teachers of a college into two groups A and B by randomly assigning 20 to each of the groups. Group A was trained in various skills of teaching through micro- teaching and the group B was trained through simulation technique. After a period of two months training, the student-teachers were rated in the teaching skills by supervisors. The rating scores of the student teachers are given in table 16.1:

Table 16.1: Scores of Student Teachers

Group	Rank	Group B	Rank
90	39	46	4
78	29	42	1
75	25.5	65	15
72	22	61	12
75	25.5	64	14
83	33.5	82	32
73	23	69	19
80	31	66	16
74	24	56	9
67	17	48	6
63	13	68	18
45	3	44	2
55	8	85	36
84	35	83	33.5
89	38	71	21
77	28	87	37
70	20	76	27
58	10	50	7
47	5	59	11
92	40	79	30
$N_1=20$	$\sum R_1=469.50$	$N_2=20$	$\sum R_2=350.50$

All rating measures are ranked from lowest to highest and the Mann-Whitney U test is used to test the null hypothesis at the .05 of significance using the formula of U_1 and U_2.

$$U_1 = (20)(20) + \frac{(20)(21)}{2} - 469.50$$

$$= 140.50$$

$$U_2 = (20)(20) + \frac{(20)(21)}{2} - 350.50$$

$$= 259.50$$

Using the equation $U_1 = N_1N_2 - U_2$, we check:

$$140.50 = 400 - 259.50$$

$$140.50 = 140.50$$

$$Z = \frac{140.50 - \frac{400}{2}}{\sqrt{\frac{(20)(20)(41)}{12}}}$$

$$= -1.61$$

The obtained Z value of –1.61 does not exceed the Z critical value of 1.96 at .05 level, the null hypothesis is accepted. It may be concluded that micro teaching approach and simulation technique are equally effective in developing certain teaching skills among student teachers.

Check Your Progress

Notes : a) Space is given below for your answer.

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

3. In answering a questionnaire the following scores were achieved by 10 men and 20 women:

   Men: 22, 31, 38, 47, 48, 48, 49, 50, 52, 61

   Women: 22, 23, 25, 25, 13, 33, 34, 35, 37, 40, 41, 42, 43, 44, 44, 46, 48, 53, 54

   Do men and women differ significantly in their answers to this questionnaire? Apply Median test by taking the Median = 41.5.

   .....

   .....

   .....

   .....

4. The performance scores of the students taught by method A and method B are given below:

Method A	Method B
50	49
60	90
89	88
94	76
82	92
75	81
63	55
52	64
97	84
95	51
83	47
80	70
77	66
80	69
88	87
78	74
85	71
79	61
72	55
68	73

   Apply Mann-Whitney U test and test the significance between the performance of the students taught by method A and method B.

   .....

   .....

   .....

16.5 STATISTICAL INFERENCE BASED ON NON-PARAMETRIC TEST: RELATED SAMPLES

Various tests are used in drawing statistical inferences in case of related samples. In this section we shall confine our discussion to the use of Sign Test and Wilcoxon Matched-Paris Signed-Ranks Test Only.

16.5.1 The Sign Test

The sign test is the simplest test of significance in the category of non-parametric tests. It makes use of plus and minus signs rather than quantitative measures as its data. It is particularly useful in situations in which quantitative measurement is impossible or inconvenient, but on the basis of superior or inferior performance it is possible to rank with respect to each other, the two members of each pair.

The sign test is used either in the case of single sample from which observations are obtained under two experimental conditions are obtained under two experimental conditions and the researcher wants to establish that two conditions are different or to the case of two equivalent samples in which the subjects are matched with respect to the relevant extraneous variables. The use of this test does not make any assumption about the form of the distribution of differences. The only assumption underlying this test is that the variable under investigation has a continuous distribution.

If the number of the individuals in the single sample or in each of the equivalent or related samples is less or equal to 25, we make use of the 'Table of Probabilities Associated with Values as Small as Observed Values of x (number of fewer signs) in the Binomial Test' (Siegal, 1956).

When the number of individuals in the group is larger than 25, the normal approximation to the binomial distribution is used. The significance of difference is tested by calculating the value of Z given by the formula:

$$Z = \frac{(X \pm .5) - \frac{1}{2}N}{\frac{1}{2}\sqrt{N}}$$

Where (x + .5) is used when x < 1/2 N and (x–5) is used when x > 1/2 N. The significance of the obtained value of Z is determined by reference to the normal Table A.

16.5.2 The Wilcoxon Matched – Pairs Signed – Ranks Test

The Wilcoxon matched-pairs signed – ranks test is more powerful than the sign test because it tests not only direction but also magnitude of differences within pairs of matched groups. This test, like the sign test, deals with dependent groups made up of matched pairs of individuals and is not applicable to independent groups. The null hypothesis would assume that the direction and magnitude of pair difference would be about the same.

The application of wilcoxon matched-pairs signed-ranks test involves the following steps:

1. Let d_i be the difference scores for any matched pair, representing the difference between a pair's scores under two treatments A and B. There would be one d_i for each pair of scores.
2. Delete all such pairs for which $d_i = 0$
3. Rank all the d_i's without regard to sign, giving rank 1 to the smallest difference d_i rank 2 the next smallest, etc. If two or more d_i's are of the same size, assign the same rank to such tied cases. The rank assigned would be average of the ranks which would have been assigned if the d_i's had differed slightly

from each other. For example, if three pairs yield d_i's of –1, –1 and +1, then each pair would be assigned the rank of 2, for $\frac{1+2+3}{2}=2$, and next d_i on order would be assigned the rank of 4 because ranks 1, 2 and 3 have already been exhausted.

4. Indicate which ranks arose from negative d_i's and which ranks arose from positive d_i's by affixing to each rank the sign of difference.

5. Sum the ranks for the positive differences and sum the ranks for the negative differences. Under the null hypothesis we would expect the two sums to be equal. In other words, if the sum of the positive ranks equals the sum of the negative ranks, we would conclude that the treatments A and B are not different. But if the sum of the positive ranks is very much different from the sum of the negative ranks, we would infer that the treatment A differs from treatment B and thus we would reject the null hypothesis.

Let us illustrate the application of Wilcoxon test with the help of the following example used by Koul (1997). Suppose a group of 26 delinquent children were initially rated for their social adjustment by psychiatrist and sent to a juvenile jail. After a year they were rated again by a psychiatrist for social adjustment and then initial and final adjustment rating scores were compared. The rating data are presented in the following Table 16.2:

Table 16.2: Rating Scores of Delinquent Children

Post-Rating Score	Pre-Rating Score	d	Rank of d	Rank with less Frequent Sign
50	47	3	7.5	
54	52	2	5.5	
62	63	–1	–2.5	2.5
39	31	8	17.5	
56	52	4	10	
51	42	9	19	
58	51	7	16	
60	49	11	23	
48	42	6	14	
46	42	4	10	
42	40	2	5.5	
53	56	–3	–7.5	7.5
40	36	4	10	
51	50	1	2.5	
56	42	19	26	
60	48	12	24	
57	47	10	21	
43	48	–5	–12	12
45	37	8	17.5	
52	39	13	25	
61	60	1	2.5	
62	52	10	21	
48	42	6	14	
39	45	–6	–14	14
41	42	–1	–25	25
39	49	–10	–21	21

The null hypothesis that there was no difference in initial and final adjustment rating scores of the group was tested at .05 level of significance using the following formula:

$$Z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}}$$

in which

N = number of pairs ranked

T = sum of ranks of the smaller of the like-signed ranks

In the example, T, the smaller of sums of the like-signed ranks = 2.5 + 7.5 + 12.0 + 14.0 + 2.5 + 21 = 59.5.

$$Z = \frac{59.5 - \frac{(26)(26+1)}{4}}{\sqrt{\frac{(26)(26+1)(52+1)}{24}}}$$

$$= 2.95$$

Since the obtained Z value of 2.95 exceeds Z critical value of 1.96 at .05 level, the null hypothesis is rejected and we may conclude that the environment in juvenile jail has considerable improved the social adjustment of delinquent children.

Check Your Progress

Notes : a) Space is given below for your answer.

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

5. List the uses of : (i) sign test and (ii) Wilcoxon matched-pairs signed-ranked test.

.....

.....

.....

.....

16.6 STATISTICAL INFERENCE REGARDING CORRELATION USING NON-PARAMETRIC DATA

The correlations which are computed from the measurements based on nominal (enumerative) and ordinal (ranking) data give rise to spearman's rho (ρ), phi (ϕ) contingence (C) coefficients. In this section, we will discuss the procedure for testing the statistical significance of these coefficients.

16.6.1 Significance of Spearman's Rho (ρ) Correlation Coefficient

There is no generally accepted formula for estimating the standard error of ρ which we need for testing its significance and determine its confidence limits. However, we can test the null hypothesis that the two variables under study are not associated in the population and that the observed value of rho (ρ) differs zero only by chance, in two ways.

1. When the size of sample N is from 4 to 30, the interpretation is best made by the aid of Table L given in the Appendix, in which are given ρ coefficients significant at .05 and .01 levels of confidence. This is a one tailed table, that is, the stated probabilities apply when the observed value of ρ is in the predicted direction, either positive or negative. For a one tailed test, if an observed value of ρ equals or exceeds the value of ρ shown in the Table L, the observed value is significant at the level indicated.

2. When N is 10 or large, the significance of an obtained ρ under null hypothesis may tested by the formula:

$$t = \rho \sqrt{\frac{N-2}{1-\rho^2}}$$

The interpretation of the obtained value of t is made with the use of Table C of the Appendix, using (N–2) degrees of freedom (df).

16.6.2 Significance of Phi (ϕ) Correlation Coefficient

The significance of ϕ coefficient is determined using the relationship of ϕ to χ^2 as:

$$\chi^2 = N\phi^2$$

This formula helps us to test an obtained ϕ against the null hypothesis. First we convert ϕ to an equivalent χ^2 and then test the significance of χ^2 by referring to the chi-square Table E. If χ^2 comes out to be significant for a particular level of confidence, the corresponding value of s is also significant.

16.6.3 Significance of Contingency Coefficient (C)

The significane of contingency coefficient C is also determined through the relationship which it bears to χ^2 using the formula:

$$C = \sqrt{\frac{\chi^2}{N+\chi^2}}$$

This formula helps us to test the significance of the obtained value of C coefficient against the null hypothesis by first converting C to χ^2. The interpretation of an obtained chi-square is made with the use of chi-square Table E. If chi-square is significant at a particular level of confidence, C is also significant.

Check Your Progress

Notes : a) Space is given below for your answer.

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

6. Test the significance of rho (ρ) 0.76 for N = 2

.....

.....

7. The coefficient of contingency between father's eye colour and son's eye colour computed on the basis of 4×4 contingency table came out to be 0.46. Test its significance at .05 level.

.....

.....

16.7 LET US SUM UP

1. The use of non-parametric tests do not make stringent assumptions about the nature of the population distribution. Non parametric tests are distribution free tests.
2. Non-parametric tests are used when: (i) the nature of the population, from which samples are drawn, is not known to be normal; (ii) the variables are expressed in nominal scale of measurement; and (iii) the data are measures which are ranked or expressed in numerical scores which have the strength of ranks.
3. Chi-square test, median and Mann-Whitney U test are most frequently non-parametric tests of significance which we use in case of unrelated or independent samples.
4. In case of related or dependent samples, we make use of sign test and Wilcoxon-matched-pairs signed-ranks test.
5. The procedure for testing the significance of rho (ρ), phi (ϕ) and contingency (C) coefficients of correlations were also discussed.

16.8 UNIT-END ACTIVITIES

1. Discuss the uses of non-parametric tests.
2. Describe the uses of chi square test, median test and Man-Whitney U test.
3. Illustrate the use of Wilcoxon matched-pairs signed-ranks test with the help of an example.

16.9 SUGGESTED READINGS

Garrett, H.E. (1962): *Statistics in Psychology and Education*. Bombay: Allied Pacific Pvt. Ltd.

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

Koul, Lokesh (1997): *Methodology of Educational Research*. New Delhi; Vikas Publishing House Pvt. Ltd. (Third Revised Edition).

Siegal, S. (1956): *Non-Parametric Statistics for Behavioural Sciences*. Tokyo: McGraw Hill Hoga Kusna Ltd.

16.10 ANSWERS TO CHECK YOUR PROGRESS

1. $\chi^2 = 346$, the deviation from the normal distribution is significant.

2. $\chi^2 = 7.03$, no signifance sex difference in attitude towards the question.

3. No, $\chi^2 = 1.35$

4. No, $Z = -1.61$

5. i) Sign test is particularly useful in the situations is which quantitative measurements is impossible or impracticable, on the basis of superior or inferior performance. It is applicable either to the case of single sample from which observations are obtained under two experimental conditions and one wished to establish that two conditions are different or to the case of the equivalent samples in which the subjects are matched with respect to the relevant extraneous variables.

   ii) Wilcoxon test is more powerful than the sign test because it tests not only direction but also magnitude of differences within pairs of matched groups.

6. Significant at .01 level.

7. Not significant at .05 level.

UNIT 17 ANALYSIS OF QUALITATIVE DATA

Structure

17.1 Introduction

17.2 Objectives

17.3 Nature of Qualitative Data

17.4 Sources of Qualitative Data Collection

17.5 Deciding Data Analysis Strategies

- 17.5.1 Ongoing Analysis Versus Analysis at the End
- 17.5.2 Structured and Open-ended Analysis
- 17.5.3 Analysis Related to Qualitative Genre

17.6 Components of Qualitative Data Analysis

- 17.6.1 Data Reduction
- 17.6.2 Data Display
- 17.6.3 Drawing and Verifying Conclusion

17.7 Let Us Sum Up

17.8 Unit-end Activities

17.9 Points for Discussion

17.10 Suggested Readings

17.11 Answers to Check Your Progress

17.1 INTRODUCTION

Analysis and interpretation of data is the most crucial phase in social science research. But, the challenge faced by the social science researcher is to make sense of a massive amount of data, reduce the volume of information, identify significant patterns, and construct a framework for communicating the essence of what the data reveal. In case of qualitative data "few agreed on canons for these data analysis, in the sense of shared ground rules for drawing conclusions and verifying their sturdiness" (Miles and Huberman, 1984). There are no formulae for determining significance. There are no ways of perfectly replicating the researcher's analytical thought processes. There are no straightforward tests for reliability and validity. In short, there are no absolute rules except to do the very best with our full intellect, to fairly represent the data and communicate what the data reveals given the purpose of the study.

This does not mean that there are no guidelines to assist in analyzing qualitative data. But guidelines and procedural suggestions are not rules. Applying guidelines requires judgement and creativity. Because each qualitative study is unique, the analytical approach used will be unique. Because qualitative inquiry depends at every stage on the skills, training, insights, and capabilities of the researcher, qualitative analysis ultimately depends on the analytical intellect and style of the analyst. The human factor is both the great strength and the fundamental weakness of qualitative inquiry and analysis.

In this unit, you will be provided with the knowledge of nature of qualitative data, sources of qualitative data collection, how to decide data analysis strategies and components of qualitative data analysis.

17.2 OBJECTIVES

After studying this unit, you will be able to:

- discuss the complexities involved in the analysis of qualitative data;
- identify and discern quantitative and qualitative data;
- describe the variety of sources through which qualitative data can be collected;
- list the qualitative data analysis strategies;
- explain the uniqueness of different data analysis strategies; and
- analyze data using qualitative techniques.

17.3 NATURE OF QUALITATIVE DATA

In Unit 14 you were told that there are two types of data: (i) quantitative and (ii) qualitative. In quantitative data, numerical values are assigned to the characteristics or properties of objects or events, according to logically accepted rules. Whereas, in qualitative data no such numerical values are assigned and the description or narration of events or situations are taken as they are.

As you already know quantitative data describe an empirical event or phenomenon in a numerical system with the help of different scales of measurement: nominal, ordinal, interval and ratio. Nominal scales of measurement are used when a set of objects among two or more categories is to be differentiated on the basis of certain clearly known characteristics, such as gender, nationality etc. The ordinal scales of measurement correspond to quantitative classification of a set of objects done with the help of ranking on a continuum. The interval scale of measurement is based on equal units of measurement. It includes how much or how little of a given characteristic or attribute is present. Ratio scale is the highest level of measurement. Since this scale assumes the existence of absolute zero, this type of measurement is almost non-existent in educational and psychological measurement.

Qualitative data consists of 'detailed descriptions' of situations, events, people, interactions, observed behaviours, still or moving images and artifacts. These data are also available in the form of 'direct quotations' from people about their attitudes, beliefs, and thoughts; 'excerpts' or 'entire passages' from documents, correspondence records and case histories. In addition to these, verbal data gathered through open ended questionnaires, observations, and interviews are also mostly qualitative in nature. It may be noted that all these data are not usually immediately accessible for analysis, but require some processing. Raw field notes needs to be corrected, edited, typed; tape recordings need to be transcribed and corrected; video filming needs to be technically edited, and so on.

Check Your Progress

Notes : a) Space is given below for your answer.

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

1. State briefly the nature of qualitative data.

.....

.....

.....

.....

2. Give three examples of qualitative data.

.....

.....

.....

.....

3. State any two characteristics of qualitative data.

.....

.....

.....

.....

17.4 SOURCES OF QUALITATIVE DATA COLLECTION

The main source of qualitative data collection is the 'Field' or 'Research Site'. When a researcher enters the field, practically every aspect of the field provides scope for qualitative data collection. For example, the people and their interaction patterns, the social event / s, the cultural practices, the language and literature, the art and artifacts, the general ambiences of the research site and the researcher's field notes and diary. Apart from these, the archieval materials also form another major source of qualitative data collection.

17.5 DECIDING DATA ANALYSIS STRATEGIES

Analysis of qualitative data is the process of systematically organizing the information collected from the field. In a study typically following qualitative approach, data analysis is a fixed stage / step and the data analysis techniques are more or less decided in advance. Whereas in a qualitative study, deciding about the data analysis strategies is very crucial and much depends upon the thinking and creativity of the researcher. There is typically not a precise point at which data collection ends and analysis begins. However, before taking any decision about qualitative data analysis strategies, the researcher makes a review of the following:

i) The conceptual framework of the study.

ii) The research questions raised in the study.

iii) The strategy for research and design adopted.

Thus, the above review sets the stage for deciding about qualitative data analysis strategies.

The qualitative data analysis strategies are based on several decisions which include: (i) Ongoing Analysis versus Analysis at the End; (ii) Structured or Open-ended Analysis; and (iii) Analysis Related to Qualitative Genre.

17.5.1 Ongoing Analysis Versus Analysis at the End

In terms of when the researcher formally starts analyzing, there are basically two options available to him – ongoing analysis or analysis at the end of data gathering. In the ongoing analysis phase, the analysts formally reflect about the data, ask analytic questions and write analytic notes throughout the study. Those who analyse at the end, wait until all (or most) of data are gathered and then begin the task of asking analytic questions. There are advantages and disadvantages to both. More experienced field workers tend to analyse as they go. The beginners try to do some analysis as the study unfolds because it makes the final analysis easier and less daunting. However, in either case, Rossman and Rallis (1998) have suggested the following points which may be kept in view by a researcher:

- Keep your questions in mind. Remember what you are trying to learn about.
- Modify your data gathering based on what you are learning, not chance. Ask analytic questions as you go along.
- Write all the time. Note hunches, thoughts, impressions; write analytic notes.
- Talk your ideas through with people.
- Read and read and read what others have said about the topic.
- Practice good management skill for keeping the data organized and accessible.
- Discipline yourself to log the day's activities, noting the date, what you did, names, times, and places.

These points are invaluable suggestions which will facilitate ongoing analysis or analysis at the end.

Check Your Progress

Notes : a) Space is given below for your answer.

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

4. What are the points to be kept in mind before deciding about Data Analysis Strategies?

.....

.....

.....

.....

5. What does a researcher do at the:

   i) Ongoing analysis state

   .....

   .....

   .....

   .....

   ii) Analysis at the end.

   .....

   .....

   .....

   .....

17.5.2 Structured and Open-ended Analysis

Another decision to be made is how structured or open-ended is the analysis. In a structured strategy, analytic categories have been identified at the conceptualization and design phase. Open-ended analysis, in contrast, pose grand questions, are open to the unexpected, and let the analytic direction of the study emerge as it progresses. For example, suppose you are engaged in a study of **School Improvement Programme**. In this study, the structured analytic categories would be as follows:

School Improvement Programme

Structured Analytic Categories	Impinging Factors	Internal Context	Adoption Decision	Cycle of Transformations	Outcomes

In the same example, if you have decided to use open ended analysis strategies the procedure would be as follows:

School Improvement Programme

Impinging Factors

- What are the assumptions or characteristics of this innovative programme?
- How far does the External Context – the community or district education office, act as an impinging factor?

Internal Context

- What is the prior history of innovations in this school?
- To what extent do the demographic factors affect the school innovation programme?
- What are the organizational rules, norms, work arrangements, practices within the class and the school at large?

Adoption Decision

- Do the school authorities have concurrence regarding the decisions to be adopted?

- What are the support systems made available for implementation?

Cycle of Transformations

- How frequently does the school plan for improvement?

Outcomes

- To what extent is the school improvement programme institutionalized?
- What are the perceived gains and losses – individual and institutional?
- What are the side effects – positive, negative, anticipated and unanticipated?

Depending upon the responses to the above categories, further categories and themes will emerge. Forecasting or closely stipulating analytic categories does not occur with open-ended studies. In practice, studies fall somewhere along a continuum, as with the assumptions you make about social science and society. Most studies are more or less either open-ended or structured; few are all one way or the other.

Check Your Progress

Notes : a) Space is given below for your answer.

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

6. What is the main difference between structured analysis and open-ended analysis?

.....

.....

.....

.....

17.5.3 Analysis Related to Qualitative Genre

Analysis is also shaped by the genre framing your study. For example, phenomenological studies are open-ended, searching for themes of meaning in participant's lives. Broad categories are sought, with sub themes to elaborate the topography of meaning. A feminist phenomenological study would search for the deep meaning of women's experiences – what are women's experiences of patriarchy, what role does oppression play in their lives, specific instances of discrimination and so on. Ethnographic studies usually begin with broad domains for gathering data that then shape analysis; they are balanced between structure and openness. Case studies are uniquely intended to capture the complexity of a particular event, programme, individual or place. They focus holistically on the organisation or individual.

Analysis, then, proceeds logically and systematically from decisions made earlier on; it does not begin at some point in the study. Decisions made about the specific strategy, assumptions about knowledge and truth, and the genre most congruent with these decisions, all forecast analysis. Whether you analyse as you go on or hold off until the end of a study; whether you have stipulated analytic categories at the conceptualization stage or let them emerge; whatever the qualitative genre framing the study, the process of analyses are similar.

Check Your Progress

Notes : a) Space is given below for your answer.

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

7. Give an example to show how analysis is shaped by the genre framing the study.

.....

.....

.....

.....

17.6 COMPONENTS OF QUALITATIVE DATA ANALYSIS

Once the data analysis strategy is decided, the next task to be done involves three concurrent flows of activity – data reduction, data display, and drawing conclusion and verification (Miles Huberman, 1994). It may be mentioned here that these three activities are by no means exclusive and that they form a part of data analysis strategies. Brief explanations of those activities are as follows:

17.6.1 Data Reduction

It refers to the process of selecting, focusing, simplifying, abstracting, and transforming the data that appear in written – up field notes or transcriptions. Data reduction is not something separate from analysis. It is a part of analysis. Data reduction is a form of analysis that sharpens, sorts, focuses, discards, and organizes data in such a way that meaningful conclusions can be drawn and verified.

17.6.2 Data Display

The second major flow of analysis activity is data display. Generally, a display is an organized, compressed assembly of information that permits conclusion drawing and action. Looking at display helps us to understand what is happening and to do something – either analyse further or take action – based on that understanding.

As with data reduction, the creation and use of displays is not separate from analysis, it is a part of analysis. Designing a display – deciding on the rows and columns of a matrix for qualitative data and deciding which data, in which form, should be entered in the cells – are analytic activities.

17.6.3 Drawing and Verifying Conclusion

The third stream of analysis activity is conclusion drawing and verification. From the start of data collection, the researcher in the field makes notes about the activities and their explanations. Subsequently, these notes reveal some possible configurations from where tentative conclusions are drawn and they are verified and cross checked, while in the field, to check about the credibility of those conclusions.

Given below is an example to illustrate how the above three activities have been done.

Note: The example given below is only one interaction pattern pertaining to the event which has been presented.

Field Note

Date : 24/6/1986 **Time:** 8 a.m. **Place:** Thuamul – Rampur

Event: A school drop out working as a labourer in a construction site.

Rohit is a school dropout of VII standard. I met him earlier in his house and wanted to talk to him, but he was shy and refused to talk. I then decided to observe him, how he spends / utilizes his time and in the process find out the reasons of his dropping out.

It is a construction site. One individual (who I don't know) is giving directions to a group of people (Rohit included) to attend to some task. In the mean time, he calls Rohit, gives him some money and sends him to fetch tea for him from a nearby teastall. On his return, he stares at him (why I don't know), scolds him and sends him to the same teastall again and after some time Rohit returns, speaks something to him and goes back to the work he was doing. On my enquiring, the gentleman explained that Rohit had committed some error in calculating and as a result, he had scolded him. Also, he expressed surprise that despite having studied upto Std. VII, Rohit did not know even simple calculation! Meanwhile, I had some other discussion with him (though it was not on my agenda) and by then he ordered all people working at the site for a lunch break. It was about 12.00 noon when I left the place.

The field note presented above contains a wide variety of information pertaining to my task. For the purpose of my analysis, I may not need them all. So through the process of **data reduction** the following text has been arrived at.

Date: 24/6/1986 **Time:** 8 a.m. **Place:** Thuamul – Rampur

Event: A school dropout working as a labourer in a construction site.

Rohit is a school dropout of Std. VII. When questioned by the investigator, he refused to divulge the reasons for his dropping out. Consequently, the investigator wanted to observe him to know how he utilizes his time.

Rohit was engaged as a labourer in a construction site.

One day he got a scolding (from the owner perhaps) for committing a calculation mistake.

The owner's perception was that a VIIth standard student was not supposed to commit errors in simple calculation. What is meant by simple calculation – no clarity exists.

If the above transcribed field note is presented in the form of **data display** then the text appears somewhat like this.

Date : 24/6/1986 **Time:** 8 a.m. **Place:** Thuamul - Rampur

Event : A school dropout working as labourer in a construction site.

Name of the individual	–	Rohit
Dropout at which class	–	VIIth
Current Engagement	–	Labourer at a construction site
Other than as a labourer, what was the work he was engaged in due to which he got a scolding?	–	Fetching tea from the teastall

Why was he get the scolding?	—	Error in calculating
Why was he at fault?	—	No information
What was the perception of the owner?	—	VIIth standard educands are not supposed to commit mistakes in simple calculation
What is simple calculation?	—	No clarity
Did the owner verify whose fault it was?	—	No information
What did Rohit have to say?	—	No information

Thus, the data display presented above reveals some tentative conclusions that Rohit is a school dropout of VIIth standard engaged as a labourer in a construction site. One day, due to an error in calculation, he was ridiculed by the owner. In the absence of adequate information, it is very difficult to say whether he knows calculation or not and whether it is because of this lack of mathematical knowledge that he left the school. Such tentative conclusions, help the investigator to probe further till all the unanswered questions are answered in order to arrive at a final conclusion about one aspect of Rohit's dropout behaviour.

Check Your Progress

Notes : a) Space is given below for your answer.

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

8. State the meaning of the following:

i) Data Reduction

.....

.....

.....

.....

ii) Data Display

.....

.....

.....

.....

iii) Conclusion

.....

.....

.....

.....

17.7 LET US SUM UP

In this unit, we discussed the challenges faced by the social and behavioural science researcher in doing qualitative analysis of data; the necessity of skills, training, insights, and capabilities of the researcher in analyzing qualitative data, and the human factor which is both a great strength and weakness of qualitative inquiry and analysis. Also, we have discussed the meaning and characteristics of qualitative data, the sources and the points to be kept in mind while deciding about qualitative analysis strategies. The main points of discussion in the entire unit are as follows:

- We have a few agreed – on canons for qualitative data analysis.
- There are no absolute rules for qualitative analysis of data except to do the very best with one's full intellect and fairly represent the data.
- Qualitative analysis ultimately depends on the analytical intellect and style of the analyst.
- Data are mainly of two types – Quantitative and Qualitative.
- Quantitative data describes an empirical event or phenomenon in a numerical system with the help of different scales of measurement – nominal, ordinal, interval and ratio.
- Qualitative data consists of detailed descriptions of situations, events, people, interactions, and observed behaviours.
- The main source of qualitative data is the field or research site. Other than people and their socio-cultural interaction patterns, the art, literature and artifacts too are the sources of qualitative data collection. In addition, archieval material also serves as a source.
- Decision about qualitative data analysis strategies is dependent upon the conceptual framework of the study, the attendant research questions, and the strategy for research and design.
- There are three alternatives available for deciding qualitative data analysis strategies. They are, (i) Ongoing Analysis Versus Analysis at the end, (ii) Structured or Open-ended Analysis, and (iii) Analysis Related to Qualitative Genre.
- The decision about data analysis is followed by three concurrent flows of activity viz, Data Reduction, Data Display, and Drawing Conclusion and Verification.
- Data reduction is the process of selecting, focusing, simplifying, abstracting, and transforming the data that appear in the field notes.
- Data display is an organized, compressed assembly of information that permits drawing of conclusion and action.
- Conclusion drawing and verification stems from data display which is verified and cross-checked while the researcher is in the field.

17.8 UNIT-END ACTIVITIES

1. Identify three dropout cases in your locality at school level. Observe them and display their data in qualitative terms.

2. Observe children who do not attend school and prepare a display record of these children.

17.9 POINTS FOR DISCUSSION

Take the following points for discussion with your counsellor.

1) Analysis of Qualitative data or Qualitative analysis of data – Are they the same or are they different? Discuss

2) Why is it necessary for the beginner to try to do some analysis as the study unfolds?

3) Can there be any other way to decide about qualitative data analysis strategies?

17.10 SUGGESTED READINGS

Denzin, Normann K. and Lincoln, Y. S. (eds.) (1994): *Handbook of Qualitative Research.* New Delhi: Sage Publications.

Patton, Michael, Quinn (1990): *Qualitative Evaluation and Research Methods.* New Delhi: Sage Publications.

Rossman, G.B., and Rallis, S.F. (1998) *Learning in the Field – An Introduction to Qualitative Research.* New Delhi: Sage Publications.

Tesch, Renota (1990): *Qualitative Research: Analysis Types and Software Tools.* London: The Falmer Press.

Miles, Mattew. B. and Hubarman, A. Michael (1994): *Qualitative Data Analysis – An Expanded Source Book.* 2nd ed., New Delhi: Sage Publications.

17.11 ANSWERS TO CHECK YOUR PROGRESS

1. Qualitative data is the detailed descriptions of situations, events, people, interactions, and observed behaviours.

2. i) Quotations, ii) Narration of experiences by an Individual, iii) Responses to an open-ended question.

3. Qualitative data are words, that is, language in the form of extended text. They are at times direct quotations from people, excerpts or entire passages from records and documents.

4. The conceptual framework of the study, the attendant research questions, and the strategy for research and design.

5. i) The Ongoing analyst formally reflects about data, asks analytic questions, and writes analytical notes.

   ii) Those who analyse at the end, wait until all (or most) of the data are gathered and then begin the task analytic questions.

6. In structured analysis the analytic categories are identified at the conceptualization and design phase where as in the open – ended analysis the analytic direction emerges as the study progresses.

7. For instance, a feminist phenomenologic study would search for the deep meaning of women's experiences. The conceptual framework would seek information pertaining to women's experiences of patriarchy, how oppression plays out in their lives and so on.

8. i) Data Reduction refers to the process of selecting, focusing, simplifying, abstracting, and transforming the data that appear in the field notes.

   ii) Data Display is an organized, compressed assembly of information that permits conclusion drawing and action.

   iii) Conclusion drawing and verification stems from display which is cross-checked while the researcher is in the field.

UNIT 18 DATA ANALYSIS TECHNIQUES IN QUALITATIVE RESEARCH

Structure

18.1 Introduction

18.2 Objectives

18.3 Data Analysis Techniques in Qualitative Research

18.4 Codification

- 18.4.1 Types of Codes
- 18.4.2 Some Do's and Don'ts about Codes

18.5 Categorization and Classification

- 18.5.1 Categorization
- 18.5.2 Classification

18.6 Content Analysis

- 18.6.1 Approaches to Content Analysis
- 18.6.2 The Procedure of Qualitative Content Analysis
- 18.6.3 Techniques of Qualitative Content Analysis

18.7 Triangulation

- 18.7.1 Triangulation of Methods
- 18.7.2 Triangulation of Sources
- 18.7.3 Analyst Triangulation
- 18.7.4 Theory / Perspective Triangulation

18.8 Let Us Sum Up

18.9 Unit-end Activities

18.10 Points for Discussion

18.11 Suggested Readings

18.12 Answers to Check Your Progress

18.1 INTRODUCTION

The analysis of qualitative data is best described as a progression, not a stage; an ongoing process, not a one time event. Marshall and Rossman (1989) explain, "Data analysis is the process of bringing order, structure, and meaning to the mass of collected data. It is a messy, ambiguous, time consuming, creative, and fascinating process. It does not proceed in a linear fashion; it is not neat. Qualitative data analysis is a search for general statements about relationship among categories of data; it builds grounded theory," (P. 112).

The interaction between data collection and analysis is one of the major features that distinguish qualitative research from traditional research. The researcher collects the first available data from the field, and immediately forms very tentative working hypotheses that cause adjustments in interview questions, observational strategies, and other data collection procedures. New data, obtained through refined procedures, test and reshape the tentative hypotheses that have been formed and refined by the data collection procedures. This interactive refining process never really ceases

until the final report has been written. This whole process is facilitated through the decision about data analysis strategies that have been decided early and the concurrent activities viz., data reduction, data display, conclusion drawing and verification help the researcher finally analyse and interpret the data. However before the so called 'final' (term given by Glaser and Strauss, 1967) analysis is attempted by the analyst, the following activities have to be carried out which are mandatory in nature. The activities are (i) Organization of Data and (ii) Familiarizing oneself with the Data.

Organization of Data

The qualitative data gathered through open-ended questionnaires, indepth interviews and participant observations are voluminous. Sitting down to make sense out of those pages and field notes seems pretty difficult. Therefore, before you actually begin the formal analysis, it is important to clean up and organize as you go along. It saves time, creates a more complete record, and stimulates analytical thinking.

The first thing to do is to make sure it's all there. Are the field notes ready? Be sure that you know where and when you took the field notes, who was there, and what the event was all about. Check to be sure that interview transcriptions are dated and who it was, whom you interviewed. Also be sure that you have written down hunches and analytic ideas throughout the study; this shapes and refines your thinking and provides insights for analysis.

Once the researcher is certain that all data are there, he/she has checked out the quality of the data, and has filled in any missing gaps, formal analysis can begin. It may be noted here, that since data are unique and priceless and it is not possible to obtain the same conversational or observational notes once again; for safety purpose it would be desirable to make at least a few copies. Once copies of the data have been made the formal analysis can begin. Since analysis of qualitative data is a creative process, there is no right way to go about organizing qualitative data. Therefore, the data analysis techniques that will be discussed in this block will be suggestive rather than prescriptive.

Familiarizing Oneself with the Data

This is the most crucial aspect of data analysis process. Very often, due to our unfamiliarity with the data, we tend to drift along the way and get confused. Therefore, it is necessary for the researcher to read, reread, and once more read through the data. This intense and often tedious process enables you to become familiar with what you have learned in intimate ways. There is no substitute for transcribing. It familiarizes you with the data, provides leads for further data gathering, provokes insight, and stimulates analytical thinking. Furthermore, not all people are good visual learners or oral learners. So, transcribing is not a mere technical exercise, it caters to both types of persons. Thus this task requires extraordinary discipline or the part of the researcher.

18.2 OBJECTIVES

After having studied this unit, you will be able to:

- identify the basic requirements of qualitative data analysis techniques;
- discuss the different qualitative data analysis techniques such as codification, categorization and classification, content analysis, and triangulation procedure;
- describe the procedure involved in different qualitative data analysis techniques;

- differentiate amongst various qualitative data analysis techniques;
- appreciate that qualitative data analysis requires intellectual rigor and creativity; and
- use qualitative data analysis techniques in your research work.

18.3 DATA ANALYSIS TECHNIQUES IN QUALITATIVE RESEARCH

A wide variety of data analysis techniques are used in qualitative research. Familiarity about all those techniques would be highly desirable on the part of every researcher. But, for a beginner it would be pertinent to know about a few techniques which are most commonly used. Some of those techniques are presented below.

18.4 CODIFICATION

Coding is the formal representation of analytical thinking. Codes are tags or labels for assigning units of meaning to the descriptive or inferential information compiled during a study. Codes are usually assigned in the form of an alphabet or alphabets, words, phrases or sentences. Use of symbols and numbers are generally avoided because data retrieval becomes very difficult and the analyst finds it cumbersome to be able to get back to the concept as quickly as possible.

Example

In a programme evaluation research, suppose you are coding an interview data then coding will appear somewhat like the following:

Code	Meaning
Prog	Programme
Org	Organization
Ob	Observation
Obs	Observations
P	Participant
Ps	Participants
S	Staff
Prog. Ideology	Programme Ideology
Prog. Process	Examples of Programme Process
Group Proc.	Group Process
Sub-G	Subgroup formations

Using the above codes the whole set of interview data can be codified.

18.4.1 Types of Codes

In the literature on qualitative research, for qualitative data analysis, especially for coding, a variety of types of codes exist but, for a beginner the following two are suggestive – Descriptive Codes and Pattern Codes.

Descriptive Codes

They entail little interpretation. Rather, you are attributing a class of phenomena to a segment of text. For example, in a study on school innovation, you might use the following codes – 'MOT' for motivation, and suppose you want to distinguish between teacher's motivation from administrators', you can use the following code – 'TR-MOT' for teacher's motivation and 'ADM-MOT' for administrator's motivation. However, in course of field work, you gain some more knowledge about the local people and their involvement in the school innovation programme, you may then use the following codes – 'PUB-MOT' for public motivation and 'PRIV-MOT' for private motivation.

Pattern Codes

Pattern codes are inferential or explanatory that identify an emergent theme or explanation. They pull together a lot of material into meaningful and parsimonious units of analysis. A coded segment of field notes illustrates an emergent leitmotiv or pattern that you have discerned in the field. The codes can be assigned, for example, LM (Leitmotiv), PATT (Pattern), TH (Theme), COMT (Commitment), ATT (Attitude), CL (Causal link) and so on for a pattern you have inferred during your field work on school innovation programme. In this case the discerned pattern would be about the teacher's commitment and attitude towards the school innovation programme.

When to Code

This is important. Some analysts reserve coding for the end of data collection. This is not desirable because late coding enfeebles analysis. Coding is not just something you do to "get the data ready" for analysis, but it is something that drives ongoing data collection.

18.4.2 Some Do's and Don'ts about Codes

Miles and Huberman (1994) has suggested the following do's and don't's in the form of advice.

- Codes are efficient data-labeling and data retrieval devices. They empower and speed up analysis.
- Creating a start list of codes prior to field work is helpful.
- Do not wait too long or move capriciously from one coding scheme to another.
- Make sure all of the codes fit into a structure. Do not add, remove, or reconfigure codes.
- Keep the codes semantically close to the terms they represent. Do not use numbers as codes.
- Define codes operationally.

Remember codes are category labels, but they are not a filling system. Every research needs a systematic way to store coded field data and a way to retrieve them easily during analysis.

Check Your Progress

Notes : a) Space is given below for your answer.

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

1. Define Codes and Coding.

.....

.....

.....

2. State the different types of codes.

.....

.....

.....

18.5 CATEGORIZATION AND CLASSIFICATION

The literature on qualitative research does not distinguish between the concepts categorization and classification; rather they are used interchangeably. However, for the purpose here let us understand them separately so that we will be able to appreciate why they are inseparable concepts in so far as qualitative data analysis is concerned. It may be pertinent to note here that both categorization and classification are processes and the analyst uses them simultaneously for qualitative data analysis. (Eventually this will become clear as you read through this section).

18.5.1 Categorization

It is a process of identifying patterns in the data; recurring ideas, themes, perspectives and descriptions that depict the social world you are studying. As you read through field notes or listen to interview tapes, your task is to identify salient themes, recurring ideas and patterns of belief that help you respond to your research questions. This is an intellectually challenging phase of data analysis. To some extent, it is an art and is creative as well. Patton (1990) describes inductive analysis as a strategy to identify salient categories or themes. Inductive analysis means that the patterns, themes, and categories of analysis come from the data; they emerge out of the data rather than being imposed on them, prior to data collection. In the inductive analysis, data can be categorized using typologies – Indigenous typologies and Analyst constructed typologies. Each of these typologies is described below.

Indigenous Typologies

Typologies are classification systems made up of categories that divide some aspects of the world into parts. Indigenous typologies (classification schemes) are expressed by participants and are generated through their analysis of how they use language and what they express. This approach requires an analysis of the verbal categories used by participants in a project, in order to break up the complexity of reality into parts. It is a fundamental purpose of language to tell us what is important by giving it a name and, therefore, separating it from other things with other names. Once these labels have been identified from an analysis of what people in the programme have said, the next step is to identify the attributes or characteristics that distinguish

one thing from the other. The purpose of this analysis is to discern and report how people construe their world of experience from the way they talk about it.

Example

Let us suppose that you are engaged in a project on reducing dropout rate among tribal school children. While collecting the data from one of the targeted schools, you realized that teachers can provide more valuable information. You then directed your observation and focused interviews on them and learnt that teachers had a peculiar way of categorizing students. It then became necessary for you to understand the ways in which teachers categorized students. With regard to the problem of disobedience, absenteeism, tardiness, and skipping classes, the teachers had come to label students as either "Chronics" or "Borderlines". One teacher described the chronics as "the ones who do not obey teachers at all and remain frequently absent in the class". Another teacher said "tribal children are like that, everything you do to get them does not work". Another teacher said "tribal children do not understand our language and there always remains a communication gap; thus to avoid unpleasant situations they show all these behaviours. You can always pick them out, the chronics. The borderlines, on the other hand, skip a few classes, remain absent only on certain times of the academic year". Another teacher said, "They are disobedient and get influenced by the chronics". Another teacher said, "Borderlines are those who are erratic in their attendance in the class".

In the above example, you will observe that all teachers used only two categories to label the students but, their definitions differed. As an analyst, you have to portray and understand the teachers' views about dropouts as represented by this indigenous typology. This is possible only when you know what are the educational inputs provided in the school / s, how these are implemented, and what are the roles of the teachers in this endeavour. Such efforts will pave the way for a clear understanding about this indigenous typology of 'Chronics' and 'Borderlines' used by the teachers and it will have important implications for the research project. Thus, these categories, then become the themes which will be important throughout the data analysis and the final report.

Analyst – Constructed Typologies

These typologies are created by the researcher and do not necessarily correspond directly to the categories of meaning used by the participants. This process entails describing patterns, themes, and categories that you see in the data and may well be subject to the "legitimate charge of imposing a world of meaning on the participants that better reflects the observer's world than the world under study" (Patton, 1990). Through logical reasoning, you may identify ways of conceptualizing data that then generate new insights or typologies for further exploration.

Example

Let us take the example used in the case of indigenous typology to have a fuller understanding about analyst – constructed typology. From the example, as an analyst you can identify four different categories of dropouts (many more and different categories are possible). They are:

The Truants

These are the students who do not obey the teachers and remain frequently absent in the class.

The Submissives

These are the students who develop complexity due to a communication gap and are non-responsive to any extent of cajoling by the teachers. Consequently, they turn out to be docile.

The Easy Preys

These are the students who cannot decide what is good or bad for them, have no goal, and get carried away by the behaviours of truants. As a result, they become victims of the situation.

The Vulnerables

These are the students who are subjected to all kinds of victimization both by their peers and teachers.

The typology presented above describes characteristics of dropouts. These typologies can later be used to make interpretations about the nature of the project.

18.5.2 Classification

Classification is a way of knowing; we have to be cognizant of the attributes of things to be able to group them (Tesch, 1990). In fact, in the early days, science was considered largely a way of developing classification systems. For some scholarly fields, classification is still a major objective. Botanical classification schemes and the periodic table of elements are well-known examples of classification schemes.

While the periodic table is an example of a classification system that is so neat that the properties of elements not yet discovered could be deduced from the rules of the system, not all things on earth can be classified so neatly. In fact, as a rule, classification has its problems. For instance, to decide which objects fall into the class of 'chair' we need criteria by which to assess them. Usually, such criteria are called the 'properties' of a category. The first property that comes to mind when I think of the class 'chair' is that it is something to sit on. But it is not good enough, since I can sit on a bench as well, or a stool, a sofa or even a log of wood. To rule out the log, I can decide that the next property is: 'it must have legs'. Like this I can go on deciding the properties and even then the class 'chair' would remain hazy, because for all the above mentioned properties different people (French, German, etc.) would have special words for the same class.

If classification becomes difficult in such a simple case, imagine classifying the vast amount of research data. Thus, on the one hand, it is obvious that classification is a helpful, probably essential tool; on the other hand, it seems artificial in many, perhaps most cases.

The qualitative analyst, using classification as a tool for analysis, then looks for convergence in the data that have been categorized using typologies. Such an activity leads to a classification system of data. Guba (1978) suggests several steps for converting field notes and interview transcripts into systematic categories for the purpose of analysis. The analyst begins by looking for "recurring regularities" in the data. These regularities represent patterns that can be sorted into categories. Categories should then by judged by two criteria: "internal homogeneity and "external heterogeneity". The first criterion concerns the extent to which the data that belong to a certain category hold together in a meaningful way. The second criterion concerns the extent to which differences among categories are bold and clear. The existence of a large number of unassignable or overlapping data items is good evidence of some basic fault in the category system. The analyst then works back and forth between the data and classification system to verify the meaningfulness and accuracy of the categories and the placement of data in categories. When several different classification systems have been developed, some priorities must be established to determine which category systems are more important than others. Prioritizing is done according to the salience, credibility, uniqueness, feasibility, and materiality of the classification schemes. Finally, the category system or class of

categories are tested for completeness. The completeness can be tested by examining the categories internally or externally. Internal examination means – the individual categories should appear to be consistent. External examination means – the class of categories should seem to comprise a whole picture.

The steps and procedures suggested by Guba for categorizing and classifying the qualitative data are not mechanical and rigid. It involves both technical and creative dimensions. No infallible procedure exists for performing it. However, for the beginner, the following points are suggestive for doing classification of data. These have been presented in the form of basic requirements.

Basic Requirements for Classification

- Identify, generate and develop categories using any one typology or both.
- Judge the categories for internal homogeneity and external heterogeneity.
- Arrange or sort out the data as per the developed categories.
- There might be possibility of some data remaining out of the categories.
- Check, recheck and verify the meaningfulness and accuracy of the categories and the placement of data in categories.
- Prioritise the categories according to salience, credibility and uniqueness of the classification schemes.
- Test the categories for completeness – both internal and external.
- Label the categories with some identification mark that will represent a class.

Example

Let us take the same example used in categorization to understand classification system and at the same time to know how categorization and classification of data cannot be seen in isolation and separate from each other.

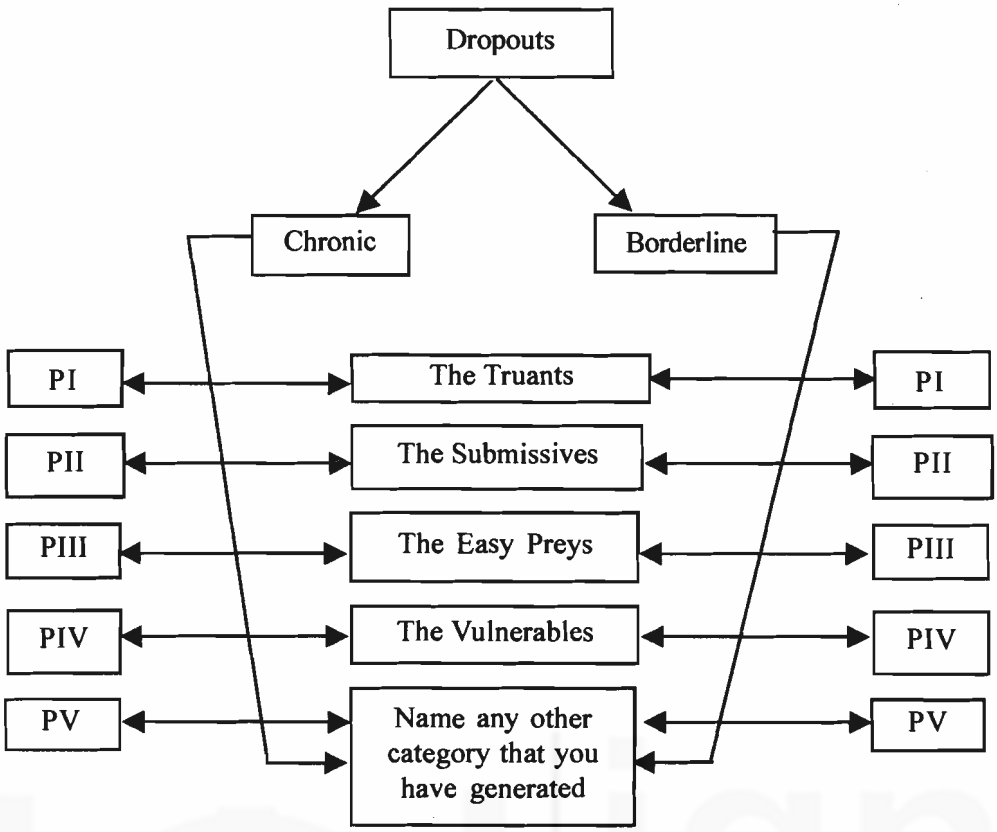
Now, according to the first requirement we have generated six categories (Chronic and Borderline through Indigenous Typology and The Truants, The Submissives, The Easy Preys and The Vulnerables through Analyst Constructed Typology).

According to the second requirement - the developed categories appear quite meaningful and at the same time distinct from one another. There is no ambiguity in understanding who are chronic or not and so on.

According to the third requirement – let us suppose that we have sorted out the data into the identified categories. Even then some data still remains to be put into categories and some questions still remain to be answered. In this case you can see that those teachers who opine "chronics are those who do not obey teachers at all and remain frequently absent in the class", find it difficult to label students who obey them but continuously remain absent in the class. Further, they are not sure whether such a description is gender free or not. Thus, while sorting out the data into categories many such data can be further refined and new categories can also be generated.

According to the fourth requirement – the rechecking and verification of categories indicated that the categories are quite accurate and the placement of data into categories is quite appropriate. (you can also verify the same on your own).

According to the fifth requirement – when we prioritise the categories, the scheme may appear somewhat like the following:



Note: P indicates Priority

According to the sixth requirement – if you examine internally, you will notice that the categories are quite consistent in the sense that they have emerged from the two types of typology adopted by the investigator. From external examination you will notice that the generated categories reveal one composite picture about the behaviour of dropouts.

Now according to the last requirement – if you give identification marks as CD (Chronic dropouts) and BD (Borderline dropouts) you have automatically created two classes.

Check Your Progress

Notes : a) Space is given below for your answer.

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

3. State the meaning of the following:

   i) Categorization

   .....

   .....

   .....

   ii) Indigenous Typologies

   .....

   .....

   .....

iii) Analyst Constructed Typologies

.....

.....

.....

iv) Classification

.....

.....

.....

18.6 CONTENT ANALYSIS

Documentary data (textual data, records, interview transcripts, excerpts from people's speech, case histories, field notes or diary, biographies and observation records) have always been central to social science analysis but modes of analyzing them vary within the social sciences. For instance, in historical study, texts are analyzed with a combination of chemical analysis and textual analysis; in psychology texts are analyzed with a combination of projective tests, clinical analysis and more precise modes of assessing the content of the written document, and so on. However, amongst other types of textual analysis, content analysis is commonly used by researchers in the field of education. Content analysis is concerned with the classification, organization and comparison of content of the document or communication. A central idea in content analysis is that the many words of the text are classified into much fewer content categories (Weber, 1985). While the 'recording units' are sometimes larger than words (sentences, paragraphs or themes), the basic procedure in content analysis is to design categories that are relevant to the research purpose and to sort all occurrences of relevant words or other recording units into these categories. (For categorization see the previous section of this unit). Then the frequency of occurrences in each category is counted and inferences are drawn.

Another regularly used technique in content analysis is the taking of 'inventories' of words in a document or transcription (Tesch, 1990). This is done by constructing exhaustive indices. Since the indices are then frequency ordered (usually with the most often used word at the top of the list), this strategy ends up also being numeric, as categorizing does. Inferences made from these inventories would also be similar to those drawn from the results of categorization. The only qualitative procedure in content analysis, i.e., the one in which numbers don't play any role at all, is the exploration of word usage, where researchers want to discover the range of meanings that a word can express in normal use. The method employed is the establishment of a 'Key – word – in - context' index. The target words are extracted with a specified amount of text immediately preceding and following them. Researchers can then group together those words in which the meaning is similar and establish how broadly or narrowly a certain term is construed by the author of the text, or they may compare word uses among groups of authors.

18.6.1 Approaches to Content Analysis

There are three approaches that a researcher may adopt in content analysis. They are: (i) characteristics of content, (ii) procedures or causes of content, and (iii) audience or effects of content.

In the first approach, the researcher is interested primarily in the characteristics of the content itself. He / she may focus either on the 'substantive nature' of the content or upon the 'form' of the content. For instance, if you are content analyzing a historical writing, you may concentrate upon the substantive aspect of the writing. Whereas, if you are content analyzing any archival material, you may concentrate both on the substantive dimension and form of the content.

In the second approach, the researcher attempts to draw valid inferences about the nature of the procedures of the content or the causes of the symbolic material from the characteristics of the material itself. For instance, if you are content analyzing a video recorded situation, you may concentrate upon the procedures adopted by the participants to create the situation.

In the third approach to content analysis, the researcher interprets the content so as to reveal something about the nature of its 'audience' or its effects'. He / she takes the content material as the basis for drawing inferences about the characteristics of the 'audience' for whom the material (content) is designed or about the effects of communication which it brings about. For instance, if you are content analyzing an interview transcript of juvenile delinquents, you will be able to interpret from the data about the probable causes that has led the children to be juvenile delinquents.

18.6.2 The Procedure of Qualitative Content Analysis

Mayring (1983) has developed a procedure for qualitative content analysis, which includes a procedural model of text analysis and different techniques for applying it. The first is to define the material, to select the interviews or those parts which are relevant for answering the research question. The second step is to analyse the situation of data collection (How was the material generated? Who was involved? Who was present in the interview situation? Where do the documents to be analyzed come from? etc.) In the third step, the material is formally characterized (How was the material documented? How was it edited-influence of the transcription on the texts? etc.) In the fourth step, decision is taken about the direction of the analysis for the selected texts and 'what one actually wants to 'interpret out of them'. This is followed by defining the analytical technique – the researcher defines which of the three techniques (presented in the next section) can be applied concretely. Finally, analytic units are defined. Here, Mayring differentiates between the units in the following way. The 'loading unit' defines what is 'the smallest element of material' which may be analyzed, the minimal part of the text which may fall under a category. The 'contextual unit' defines what is the largest element in the text which may fall under a category. The 'analytic unit' defines which passages are analyzed one after the other. In the last but one step, the actual analyses are conducted before their results are interpreted finally.

18.6.3 Techniques of Qualitative Content Analysis

The concrete methodological procedure basically includes three techniques viz., Summarizing content analysis, Explicative content analysis, and Structuring content analysis.

Summarizing Content Analysis

In this the material is paraphrased, which means that the less relevant passages and paraphrases with the same meanings are skipped (first reduction) and similar paraphrases are bundled and summarized (second reduction).

Example

An interview with a retired school teacher. 'I' – Investigator, 'RT' Retired Teacher

I – How are you sir?

RT – Fine, Thank you.

I – What's your current preoccupation?

RT – General reading in diverse areas

I – How do you plan to utilize your future time?

RT – Well, I plan to open a free coaching class for XII standard students.

I – You mean the type available / found in the society.

RT – No not exactly like that. I do not have profit as a motive.

I – How will you manage the expenses?

RT – By nominal charge

I – Can it not be attributed to profit motive

RT – No, not exactly, because I shall be charging exactly the amount of expenditure.

I – When do you plan to start?

RT – Next June

I – Is it the right time to start?

RT – Yes, precisely, I will get nearly a years time to prepare students for the exam.

I – Thank you Sir, for your noble idea

RT – You're welcome.

For the above interview transcript, if you are doing the 'summarizing the content analysis,' then the text will appear somewhat like the following:

A retired school teacher / wishes to start a coaching class / XII th standard students / no profit motive / charging only the actual expenditure incurred by the teacher / to prepare students for XII th standard board examination. (First reduction)

If you paraphrase further (second reduction) then the text will appear somewhat like the following:

A retired school teacher is looking forward to starting a coaching class for XII th standard students / as a voluntary effort / to prepare students for examination.

Now if you further paraphrase and make the final summary statement, then the text will appear like the following:

A retired school teacher is looking forward to starting a coaching class on a voluntary basis for preparing students to appear for XII th standard examination.

The Explicative Content Analysis

It clarifies ambiguous or contradictory passages by involving context material in the analysis. Definitions taken from dictionaries or based on grammar are used or formulated.

Example

Let us take the same example used in "Summarizing Content Analysis". In this interview transcript you will notice that the retired teacher is looking forward to

starting a coaching class. Also, although he is going to charge the students, he says he does not have a profit motive. Now, if you want to find out how a coaching class can be run without any profit motive, then collect the definitions pertaining to the two words 'tuition class' and 'profit motive'. Based on the meaning of these words you will paraphrase the content like "the retired teacher is intending to start a coaching class on no profit and no loss basis".

The Structuring Content Analysis

This analysis looks for types or formal structures in the material. Structuring is done on the formal aspects of the content by filtering out the internal structure. Further, the analyst looks for a single salient feature that describes more exactly the internal structure. Finally, the salient feature is interpreted in the light of the research questions of the analysis stated in the beginning.

For doing such a type of content analysis, no general rule can be defined. It depends on the respective research question and the analytical power of the analyst. However, an attempt has been made in the following example to make you familiar about structuring Content Analysis.

Example

Let us take the same example used in Summarizing Content Analysis.

The research question stated for the analysis of the interview transcript was – "What was the ulterior motive of the retired school teacher?" From the analysis of interview transcript you can filter out the internal structure like the following:

"The retired school teacher was contemplating to open a coaching class for XIIth standard students on a no profit and no loss basis." The salient feature in the internal structure can be identified as the following:

"Coaching class on no profit and no loss basis".

Check Your Progress

Notes : a) Space is given below for your answer.

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

4. Define content analysis.

.....

.....

.....

5. State the different approaches to content analysis.

.....

.....

.....

6. State the techniques of qualitative content analysis.

.....

.....

.....

18.7 TRIANGULATION

Triangulation may be defined as the use of two or more methods of data collection in the study of some aspect of human behaviour. It is a technique of research to which many subscribe in principle, but which only a minority use in practice. Denzin (1978) has identified four basic types of triangulation: (i) Data Triangulation - the use of a variety of data sources in a study; (ii) Investigator Triangulation – the use of several different researchers; (iii) Theory Triangulation - the use of multiple perspectives to interpret a single set of data; and (iv) Methodological triangulation – the use of multiple methods to study a single problem.

The term triangulation is taken from level of surveying. Knowing a single landmark only locates you somewhere along a line in a direction from the landmark, whereas with two landmarks you can take bearings in two directions and locate yourself at their intersection (Fielding and Fielding, 1986). By analogy, triangular techniques in social sciences attempt to map out, or explain more fully, the richness and complexity of human behaviour by studying it from more than one stand point. The logic of triangulation is based on the premise that "no single method ever adequately solves the problem of rival causal factors .... Because each method reveals different aspects of empirical reality, multiple methods of observation must be employed. This is termed triangulation. I now offer as a final methodological rule the principle that multiple methods should be used in every investigation" (Denzin, 1978).

The presentation on triangulation so far reveals the vast scope of the use of this technique in educational research. It would have been ideal to have a discussion on the various types of triangulation techniques. But for the purpose here let us focus our attention to triangulation as an analysis strategy. There are basically four kinds of triangulation that contribute to verification and validation of qualitative analysis: (i) Checking out the consistency of findings generated by different data – collection methods, that is, 'methods triangulation'; (ii) Checking out the consistency of different data sources within the same method, that is, 'triangulation of sources'; (iii) Using multiple analysts to review findings, that is analyst triangulation; and (iv) Using multiple perspectives to interpret the data, that is, theory / perspective triangulation.

By combining multiple observers, perspectives, methods and data sources, researchers can hope to overcome the intrinsic bias that comes from single – method, single – observer, and single – perspective studies (Denzin, 1970).

18.7.1 Triangulation of Methods

Triangulation of methods will most often revolve around comparing data collected through some kind of qualitative methods with data collected through some kind of quantitative methods. This is seldom a straight forward process because, it is likely that quantitative methods and qualitative methods will eventually answer different questions that do not easily come together to provide a single, well integrated picture of the situation. Further more, it may be stated here that there is no magic in triangulation. The researcher using different methods to investigate the same problem should not expect that the findings generated by those different methods will automatically come together to produce some nicely integrated whole. Indeed, the evidence is that one ought to expect initial conflicts in findings from qualitative and quantitative data and expect those findings to be received with varying degrees of credibility. Given below is an example of multi-method approach to study the education of tribal women. Conducted by Panigrahi, et. al. in 1987. Title: - Education of Tribal Women: A Socio – Ecological Perspective.

Objective: To explore the predominant factors operating in the tribal women's socio-ecological field with special reference to education.

Methods: Development of village profile

Examination of school records

Classification of economic and household activities

Interview with parents and teachers.

Observation of tribal women's activities through Field work.

In the above study, the use of quantitative and qualitative methods give three sets of findings which are as follows:

- Between 1981 – 1986 there were only 21 girls out of a total list of 337 tribal students in the Ashram school.
- No girls' enrolment was found in the same school after 1983.
- About 90 per cent of the cultivable land was owned by people of other communities.
- Girls were engaged in a variety of activities like rearing children, cooking, cleaning the house, washing clothes, fetching water and collecting fire wood.
- Educational attainment of the tribal women was determined mainly by her economic and social conditions.
- The school curriculum, as such, does not have any direct bearing on their economic and social life.
- Children (especially girls) before 6 years of age are not considered fit to go to school and after 10 years they are potential earners, thus they have only 3 to 4 years left in between to go to school, that too if they wish.
- In spite of the lodging and boarding facilities provided by the Government in the Ashram school, girl students do not enroll.
- There was a growing feeling among tribal people that educating their children will alienate them.

When the above findings were triangulated the researchers observed that it is the economic condition coupled with traditional life-styles and social norms for women that define the position and role of women. Further, the passive acceptance by the tribals of their age old life styles seems to continue in a cyclic way. And finally, formal schooling and curriculum has hardly any relevance to the tribal women and her life. Neither does it provide the tribal women with any alternatives to improve herself economically or socially.

In the end it may be noted here that the above observations would not have been possible without the use of triangulation technique.

18.7.2 Triangulation of Sources

The second type of triangulation involves triangulating data sources. This means comparing and cross-checking the consistency of information derived at different times by different means within qualitative methods. It means as Patton (1990) says, comparing observational data with interview data; comparing what people say in public with what they say in private; checking for the consistency of what people say about the same thing over time; and comparing the perspectives of people from different points of view.

As with triangulation of methods, triangulation of data sources within qualitative methods will seldom lead to a single, totally consistent picture (Patton, 1990). It is best not to expect everything to turn out the same. The point is to study and understand when and why there are differences. The fact that observational data produce different results than interview data does not mean that either or both kinds of data are invalid, although that may be the case. More likely, it means that different kinds of data have captured different things and so the analyst attempts to understand the reasons for the differences.

Example

Let us take the same study by Panigrahi, et. al. 1987. (Only two instances of triangulated data sources are presented here). In this study the different data sources were teachers, parents, community teachers, girl children, school records, educational officials, field notes. With regard to women's economic activities, when these different data sources were triangulated, the authors observed that their interview data revealed equal wages for men and women and observation data revealed less wages for women. When this point was further explored, the researchers came to understand that, although, legislation on equal remuneration exists, the presence of a number of exploitative forces makes it difficult for them to receive adequate wages and it is more so for women, who are paid less than men.

With regards to the education of women, the triangulated data sources indicated that the outlook regarding the education of girls was more conservative. While they endorse the view that a little bit of reading and writing is essential for every child, prolonged schooling is not necessary for the said purpose. Further, triangulation in this regard revealed that even if free lodging and boarding facilities are provided in the Ashram schools, it is only the economically and educationally better off families that were sending their children to school.

18.7.3 Analyst Triangulation

A third kind of triangulation is investigator or analyst triangulation, that is, using multiple as opposed to singular observers or analysts. Triangulating observers provides a check on bias in data collection. A related strategy is triangulating analyst – that is having two or more persons independently analyse the same qualitative data set and then compare their findings (Patton 1990).

Example

Given below are a list of studies, please refer to them for understanding Analyst triangulation.

Blease, D and Cohen, L. (1990): *Coping with Computers: An Ethnographic Study in Primary Classrooms.*

Douglass, J. K. (1976): *Investigative Social Research: Individual and Team Field Research.*

Patton, Michael. Quinn (1986): *Utilization-Focused Evaluation.*

Note: Detailed references are provided at the end of this block.

18.7.4 Theory / Perspective Triangulation

A fourth kind of triangulation involves using different theoretical perspectives to look at the same data. Observations of a group, community, or organization can be examined from a Marxian or Webrian perspective, a conflict or functionalist point of view. The point of theory triangulation is to understand how findings are affected by different assumptions and fundamental premises. For example, if you are observing

a social institution from a functionlist point of view, with regard to the concept "Conflict" you tend to believe that it is a social construct and society develops ways to control it but, from a conflict theorist point of view, you tend to believe conflict as an arena in which groups fight for power and the control of conflict means that one group is able, temporarily, to suppress its rivals. Take another example – existence of civil law in the society. A functionalist, sees it as a way of increasing social integration; but a conflict theorist sees civil law as a way of defining and upholding a particular order that benefits some groups at the expense of others.

Thus, from the above two examples you will notice how perspectives bring drastic changes in the approach to a particular phenomenon.

The discussion on triangulation so far has indicated that as an analysis strategy, it is quite robust and guards against all odds of investigator's bias. While it would be ideal to use all these types of triangulation analysis strategies, many a times it's neither desirable nor feasible. Thus, for a beginner it would be appropriate to begin with the use of the first two types viz., method triangulation and data sources triangulation as analysis strategies. With experience and maturity, all the four types can be used as data analysis strategies.

Check Your Progress

Notes : a) Space is given below for your answer.

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

7. Define triangulation.

.....

.....

.....

8. What are the four basic type of triangulation?

.....

.....

.....

9. State the different types of triangulation procedures used in qualitative data analysis.

.....

.....

.....

18.8 LET US SUM UP

In this unit we discussed the following:

- Analysis of qualitative data is best described as a progression, not a stage; an ongoing process, not a one time event.
- Before the analyst attempts to analyze the data he / she has to do two activities viz., (i) Organization of data and (ii) Familiarizing oneself with the data.

- A wide variety of data analysis techniques are used in qualitative research and the most commonly used techniques are – (i) Codification, (ii) Categorization and classification, (iii) Content analysis, and (iv) Triangulation procedures.
- Coding is the formal representation of analytical thinking. Codes are tags or labels for assigning units of meaning to the descriptive or inferential information compiled during a study.
- Codes are two types – Descriptive Codes and Pattern Codes.
- Categorization is a process of identifying patterns in the data; recurring themes, , perspectives and descriptions that depict the social world.
- Inductive analysis is a strategy to identify salient categories or themes.
- In inductive analysis data can be categorized using typologies – Indigenous typologies and Analyst - Constructed typologies.
- Classification is a way of knowing; we have to be cognizant of the document or communication.
- There are three approaches to content analysis viz., (i) Characteristics of content, (ii) Procedures or causes of content, and (iii) Audience or effects of content.
- There are three techniques used in qualitative content analysis. They are (i) Summarizing content analysis, (ii) Explicative content analysis, and (iii) Structuring content analysis.
- There are four kinds of triangulation procedure used in qualitative analysis of data. They are (i) Methods triangulation, (ii) Triangulation of sources, (iii) Analyst triangulation, and (iv) Theory / Perspective triangulation.

18.9 UNIT-END ACTIVITIES

1. Identify 3/4 dropout from a school near to your area. Conduct informal interviews with them, primarily to focus on the reason of their dropping out. Also conduct interviews with their parents and teachers. After collecting or while collecting information do the following for analysis of data.
   - — Codify the data
   - — Categorize and classify the data
   - — Triangulate the data
   - — Interpret the analysis.
2. If you have access to any library, identify a dissertation or thesis, collect information from them and try to apply the data analysis techniques you have learned in this unit.

18.10 POINTS FOR DISCUSSION

1. "Analysis of qualitative data is best described as a progression, not a stage; an ongoing process, not a one time event" – Discuss.
2. Why and how are categorization and classification not seen separately in qualitative data analysis?

3. How are content analysis and triangulation analysis techniques different from research technique?

18.11 SUGGESTED READINGS

Blease, D. and Cohen, L. (1990): *Coping with Computers: An Ethnographic Study in Primary Classroom*. London: Paul Champman.

Cohen, Louis, and Manion, Lawrence (1994): *Research Methods in Education*. 4th edn., London: Routledge.

Dane, Francis, C. (1990): *Research Methods*. California: Brooks / Cole Publishing Company, Wadsworth, Inc.

Denzin, Norman, K. (1978): *The Research Act: A Theoretical Introduction to Sociological Methods*. New York : McGraw-Hill.

Douglas, Jack, D. (1976): *Investigative Social Research: Individual and Team Field Research*. Beverly Hills, CA: Sage.

Erlandson, David, A., Harris, Edward. L., Skipper, Barbara, L. and Allen, Steve, D. (1993): *Doing Naturalistic Inquiry: A Guide to Methods*. New Delhi: Sage Publications.

Glaser, B. G. and Strauss, A. L. (1967): *The Discovery of Grounded Theory: Strategies for Qualitative Research*. Chicago: Aldine, USA.

Marshall, C. and Rossman, G. B. (1989): *Designing Qualitative Research*. Newbury: Sage.

Mayring, P. (1983): *Qualitative Content Analysis*. As quoted in Uwe Flick (1998): *An Introduction to Qualitative Research*. New Delhi: Sage Publications.

Miles, Mathew B. and Huberman, A. Michael (1994): *Qualitative Data Analysis*. 2nd edn., New Delhi: Sage Publications.

Panigrahi, S. C., Menon, Geeta, S. and Joshi, Vibha (1987): *Education of Tribal Women: A Socio-ecological Perspective*. The Indian Journal of Social Work, Vol. XLVII, No. 4, Jan. 1987.

Patton, Michael, Quinn (1986): *Utilization Focused Evaluation*. 2nd edn., Beverly Hills, CA: Sage.

Patton, Michael, Quinn (1990): *Qualitative Evaluation and Research Methods*. 2nd edn., New Delhi: Sage Publications.

Peter, K. Manning and Betsy, Cullum – Swan (1994): *Narrative, Content, and Semantic Analysis*. in Norman K. Denzin and Yvonna S. Lincoln (eds.) (1994): *Handbook of Qualitative Research*. New Delhi: Sage Publications.

Renata Tesch (1990): *Qualitative Research: Analysis Types and Software Tools*. Hampshire: The Falmer Press.

Rossman, Gretchen. B. and Rallis, Sharon. F. (1998): *Learning in the Field – An Introduction to Qualitative Research*. New Delhi: Sage Publications.

Singh, Jaspal (1991): *Introduction to Methods of Social Research*. New Delhi: Sterling Publishers Pvt. Ltd.

18.12 ANSWERS TO CHECK YOUR PROGRESS

1. Codes are tags or labels for assigning units of meaning to the descriptive or inferential information compiled during a study. Coding is the formal representation of analytical thinking.

2. There are two types of codes, namely, Descriptive codes and Pattern codes. Descriptive codes involve no interpretation but attribute a class of phenomena to a segment of text. For Example, one may use ACH-MOT for achievement motivation. Pattern codes are inferential in nature and identify an emergent theme or explanation. For example, PATT is used for (Pattern) of TH for (Theme) etc.

3. i) It is a process of identifying patterns in the data; recurring ideas, themes, perspectives and descriptions that depict the social world we are studying.

   ii) Indigenous typologies (classification schemes) are expressed by participants and are generated through their analysis of how they use language and what they express.

   iii) These typologies are created by the researcher and do not necessarily correspond directly to the categories of meaning used by the participants.

   iv) Classification is a way of knowing; we have to be cognizant of the attributes of things to be able to group them.

4. Content analysis is concerned with the classification, organization and comparison of content of the document or communication. A central idea in the content analysis is that the many words of the text are classified into much fewer content categories.

5. a) Characteristics of content.

   b) Procedure or causes of content.

   c) Audience or effects of content.

6. a) Summarizing content analysis.

   b) Explicative content analysis.

   c) Structuring content analysis.

7. Triangulation may be defined as the use of two or more methods of data collection in the study of some aspect of human behaviour.

8. a) Data triangulation

   b) Investigator triangulation

   c) Theory triangulation

   d) Methodological triangulation

9. a) Methods triangulation

   b) Triangulation of sources

   c) Analyst triangulation

   d) Perspective triangulation

UNIT 19 COMPUTER DATA ANALYSIS

Structure

19.1 Introduction

19.2. Objectives

19.3 What is SPSS?

19.4 Get Yourself Acquainted with SPSS

19.5 Menu Commands and Sub-commands

19.6 Basic Steps in Data Analysis

19.7 Defining, Editing and Entering Data

19.8 Data File Management Functions

- 19.8.1 Merging Data Files
- 19.8.2 Aggregate Data
- 19.8.3 Split File
- 19.8.4 Select Cases

19.9 Running a Preliminary Analysis

- 19.9.1 Six Characteristics of a Dataset
- 19.9.2 Data Transformation
- 19.9.3 Exploring Data
- 19.9.4 Graphical Presentation of Data
- 19.9.5 Scatterplots and Histograms

19.10 Understanding Relationship Between Variables

- 19.10.1 The Mean Procedure
- 19.10.2 Linear Regression
- 19.10.3 Curve Estimation

19.11 Non-parametric Tests

19.12 SPSS Production Facility

19.13 Satistical Analysis System (SAS)

19.14 NUDIST

19.15 Let Us Sum Up

19.16 Unit-end Activities

19.17 Suggested Readings

19.18 Answers to Check Your Progress

Appendix

19.1 INTRODUCTION

In earlier units, we provided you with a detailed understanding of how quantitative and qualitative data are analysed manually. Although some of us still carry out analysis of data manually, the advent of sophisticated computer software has made data analysis more convenient and easier. Earlier, the software which could only be run on large mainframe computers can now be run with considerable ease on the PCs. SPSS is one such software which is used in educational research. You can analyze large and computer data files with thousands of variables on your PC without compromising the quality and the precision of analysis.

In this unit, we will introduce you to the software for quantitative and qualitative data analysis. We will provide in this details of SPSS package which is comparatively more popular among research students for quantitative data analysis. We will also introduce in this unit the Statistical Analysis System (SAS), another software for quantitative data analysis. We will introduce a software called NUDIST for qualitative data analysis.

19.2 OBJECTIVES

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

- explain describe the main features of the SPSS;
- write about as well as use the data management operations and techniques of analysis using SPSS;
- acquire skills in the use of SPSS for basic statistical analysis with a special focus on the measures of central tendency, dispersion, correlation and regression; and
- present the data and the SPSS results graphically.

19.3 WHAT IS SPSS ?

SPSS* is one of the leading desktop statistical packages. It is an ideal companion to the database and spreadsheet, combining many of their features as well as adding its own specialized functions. SPSS for windows is available, as a base module and a number of optional add-on enhancements are also available. Some versions present SPSS as an integrated package including the base and some important add-on-modules.

SPSS Professional Statistics provides techniques to examine similarities and dissimilarities in data and to classify data, identify underlying dimensions in a data set. It includes procedures for cluster, k-cluster, discriminating factor, multi-dimensional scaling, and proximity and reliability analysis.

SPSS Advanced Statistics includes procedures for logistic regression, log-linear analysis, multivariate analysis and analysis of variance. This module also includes procedures for constrained non-linear regression, probit, Cox and actuarial survival analysis.

SPSS Tables creates a high quality presentation—quality tabular reports including stub and banner tables and display of multiple response data sets. The new features include pivot tables, a valuable tool for presentation of selected analytical output tables.

SPSS Trends performs comprehensive forecasting and time series analysis with multiple curve fitting models, smoothing models and methods for estimation of autoregressive functions.

SPSS categories performs conjoint analysis and optional scaling procedures, including correspondence analysis.

SPSS Chaid provides simplified tabular analysis of categories data, develops predictive models, screens out extraneous predictor variables, and produces easy to

* SPSS is registered trademark of the SPSS Corporation, USA.

read tree diagrams that segment a population into sub-groups that share similar characteristics.

Recently, the SPSS Corporation announced the release of SPSS version 8.0. Many new add-on products have also been launched in the recent months. You can consult the SPSS World Wide Web site for the latest developments and additions to the computing power SPSS. Technical support is also available to the registered user at the SPSS site. The SPSS Web site is http://www.spss.com. Select white papers on SPSS applications in major disciplines are also available on this site.

SPSS version 7.5 for Windows is now available with most users across the globe. The present unit discusses some of the commonly used data management techniques and statistical procedures using SPSS 7.5. Since new features are added almost daily, you are advised to check for these details on the currently installed version of SPSS on your computer and also consult the user manuals before undertaking complex type of data analysis. The on-line help is also available. There may be some procedures and syntax related changes from one version to another. We will attempt to provide you with procedures that are most commonly used with SPSS Release 7.5. In case these are not available on your version of SPSS, please consult the relevant SPSS authorized representative or the WWW site of the SPSS corporation.

19.4 GET YOURSELF ACQUAINTED WITH SPSS

The SPSS for Windows can be run from Windows 98 through Windows XP operating systems. Unix, Mac and mainframe versions of the SPSS software are also available. The illustrations in this Unit are based on SPSS version for Window 95/98/NT operating systems.

Starting SPSS

The SPSS for Windows uses graphical environment, descriptive menus and simple dialog boxes to do most of the work. It produces three type of files, namely data files, chart files and text files.

To start SPSS, click the start button on your Computer[2]. On the start menu that appears, click *Program.* Another menu appears on the right of the *start* menu. If there is an entry marked SPSS, that's the one you want to click. If there isn't, click the program group where SPSS was installed and an entry marked SPSS will appear. Click the SPSS 7.5 entry. You will know when the SPSS has started and SPSS Data Editor window appears. To begin with, the SPSS data editor window will be empty and a number of menus will appear on the top of the window. You will start the operations by loading a data set or by creating a new file for which data is to be entered from the data editor window. The data can also be imported from other programs like Dbase, ASCII, Excel and Lotus.

Existing SPSS

Make sure that all SPSS and other files are saved before quitting the program. You should exit the software by shutting off the program by selecting Exit SPSS command from the file menu of the SPSS Data Editor window. In case of unsaved files, the SPSS will prompt you to save or discard the changes in the file.

2 It is assumed that a proper licensed and valid version of SPSS is already installed on the computer you are working with.

Saving data and other files

Many types of file can be saved using 'save' or 'save as' command. Various types of files used in SPSS are: Data, Syntax, Chart or Output. Files from spreadsheets or other databases can also be imported by following the appropriate procedure. Similarly, an SPSS file can be saved as a spreadsheet or in dbase format. Select the appropriate save type command and save the file. The SPSS data files are saved with .sav as the secondary name. Though SPSS files could be given any name, the use of reserved words and symbols is to be avoided in all types of file names.

Printing of data and output files

The contents of SPSS data files, Output Navigator files and Syntax Files can be printed using the standard 'Print' command. The SPSS uses the default printer for printing. In the case of network printers, an appropriate printer should be selected for printing the output. It is suggested that Ink jet or Laser jet printers should be used for printing graphs and charts. Tabular data can be easily printed using a Dot matrix Printer.

Operating Windows in SPSS

There are seven types of Windows in SPSS which are frequently referred to during the data management and analysis stages. These are:

Data Editor

As mentioned earlier, the data editor window opens automatically as soon the SPSS gets loaded. To begin with, the data editor does not contain any data. The file containing the data for analysis has to be loaded with the help of the 'file' menu sub-commands by using various options available for this purpose. The contents of the active data file are displayed in the data editor window. Only one data editor window will be active at a time. No statistical operations can be performed until some data is loaded into data editor.

Output Navigator

All SPSS messages, statistical results, tables and charts are displayed in the output navigator. The output in the navigator Window can be edited and saved for future reference. The Output Navigator opens automatically, the first time some output is generated. The user can customize the presentation of reports and tables displayed in the Output Navigator. The output can be directly imported into reports prepared under word processing packages, and the output files are saved with an extension xxxx.spo.

Pivot Tables

The output shown in the Output Navigator can be modified in many ways using the Edit and Pivot Table Option, which can be used to edit text, swap rows and column, add colour, prepare custom made reports/output, create and display selectively multi-dimensional tables. The results can be selectively hidden and shown using features available in Pivot Tables.

Graphics

The Chart Editor helps in switching between various types of charts, swapping of X-Y axis, changing colour and providing facilities for presenting data and results through various type of graphical presentations.

It is useful for customizing the charts to highlight specific features of the charts and map.

The text output not displayed in the Pivot Tables can be modified with the help of Text Editor. It works like an ordinary Text Editor. The output can be saved for future reference or sharing purposes

Syntax Editor

The Syntax Editor can be opened and closed like any other file using the File Open/ New command. The use of Syntax file is recommended when the same type of analysis is to be performed at frequent intervals of time or on a large number of data files. Using Syntax File for such purposes automates complex analysis and also avoids errors due to frequent typing of the same command. The commands can be pasted on the Syntax files using a particular command and pastes buttons from the menu. Experienced user can directly type the commands in the Syntax window. To run the syntax, select the commands to be executed and click on the run button at the top of the syntax window. All or some selected commands from the Syntax File will be executed. The Syntax File is saved as xx.sps.

Script Editor

This facility is normally used by the advanced users. It offers fully featured programming environment that uses the Sax BASIC language and includes a Script Editor, Object Browser, Debugging features and context sensitive help. Scripting allows you to automate tasks in SPSS including:

- Automatically customizing output
- Open and save data files
- Display and manipulate SPSS dialog boxes
- Run data transformation and statistical procedures using SPSS command Syntax.
- Export charts as graphics files in a variety of formats.

The present module will not go into the details of the advanced features of SPSS including scripting.

Check Your Progress

Notes : a) Space is given below for your answer.

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

1. Write any three features of SPSS package.

.....

.....

.....

2. How many types of windows are there in SPSS and what are those windows?

.....

.....

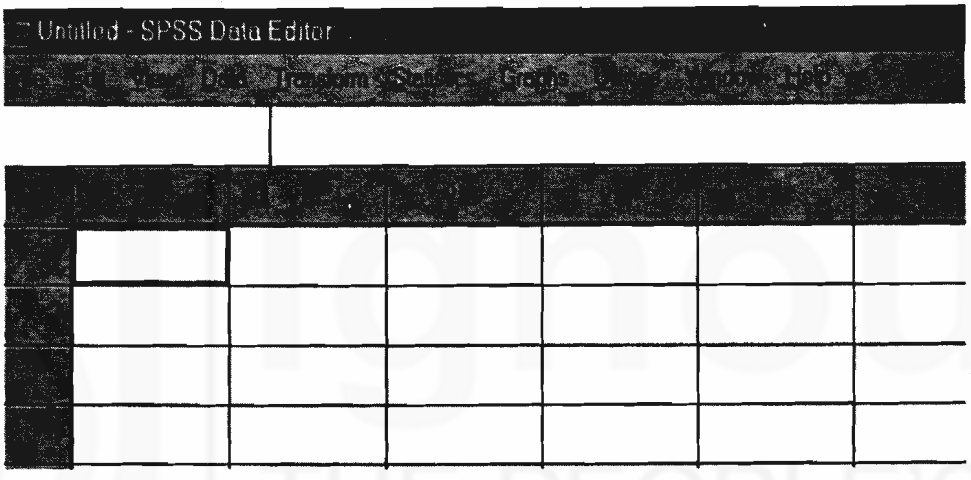
.....

.....

19.5 MENU COMMANDS AND SUB-COMMANDS

Most of the commands can be executed by making appropriate selections from the menu bar. Some additional commands and procedures are available only through the Syntax Windows. The SPSS user manuals provide a comprehensive list of commands, which are not available through menu driven options. If you want a comprehensive overview of the basics of SPSS, there is an on-line tutorial, as extensive help of SPSS is available by using the 'Help' menu command. The CD version of the software contains an additional demo module.

Since SPSS is menu driven, each Window has its own menu bar. While some of the menu bars are common, the others are specific to a particular type of Window. We will present below the menu and sub-menus of the Data Editor window. You may consult the SPSS manuals for other types of menu and sub-menu commands.



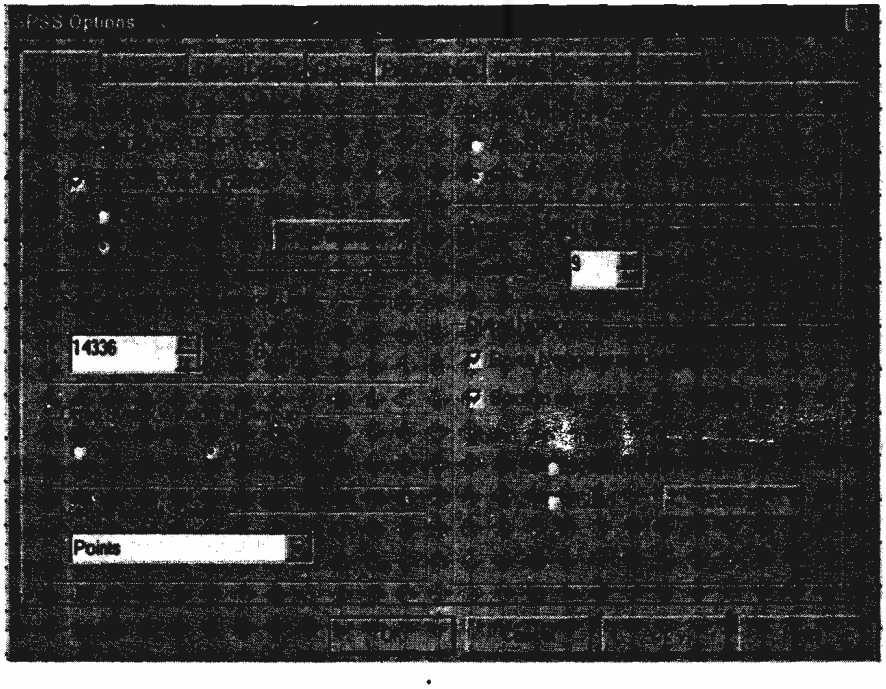
The following table shows the Data Editor Menus. Each commands in the main menu has a number of sub-command.

Menu		Function/sub-commands
File	:	Open and save data file, to import data created in other formats like Lotus, Excel, Dbf etc. Print control functions like page setup, printer setup and associated functions. ASCII data can also be read into SPSS. Data capture command is used to import data from RDMS structures.
Edit	:	These functions are similar to those available in general packages. These include undo, redo cut, copy, paste, paste special, find, find and replace. Option setting for the SPSS are controlled through Edit menu.
View	:	Customize tool bars, Fonts grid and display of data, displays option for showing value labels.
Data	:	This is very important menu as far as management of data is concerned. Variable definition, inserting new variables, transposing templates, aggregating and merging of data files, splitting data files for specific analysis are some important commands in Data Menu.

Transform	:	Compute new variables, recede, random number generation, ranking, time series data transformation, count and missing value analysis are undertaken using Transform Command.
Statistics	:	As the name implies, Statistics Menu incorporates statistical procedures. Frequency distribution, cross-tabulations, comparison of means, correlation, simple and multiple regression, ANOVA, Log linear regression, discriminate analysis, canonical analysis, factor analysis, non-parametric tests and time series analysis are undertaken using Statistics menu.
Graphs	:	Includes options for generating various type of custom made graphics like bar, pie, area, X-Y and high-low charts, pareto, control charts, box-plots, histograms, P-P and Q-Q charts and time series representation of data.
Utilities	:	Information about variables, information on working a data file, run scripts and define sets are some of the important functions carried out through Utilities command.
Window	:	Windows menu are used to switch between SPSS windows.
Help	:	Context specific help through dialog boxes, demo of the software, and information about the software are some of the important options under Help command. It provides a connection for the SPSS home page. The statistical coach included in the help module is very useful in understanding various stages of executing a procedure.

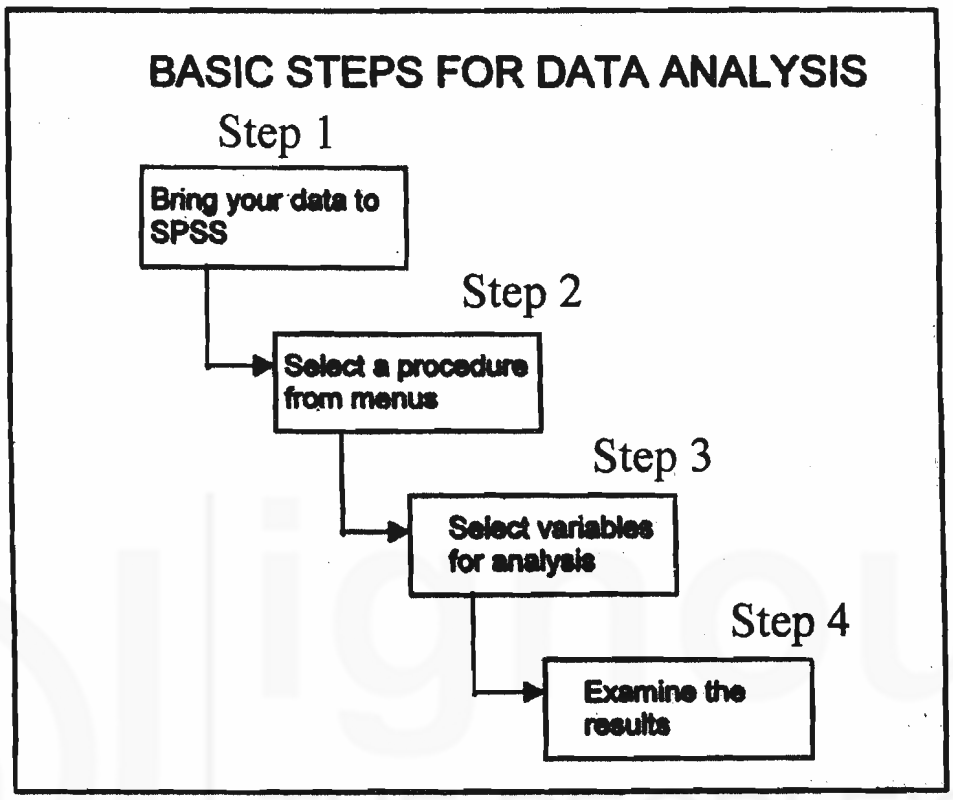
Setting the Options

The SPSS provides a facility for setting up of the user-defined options. Use the Edit menu and then select Options. The following types of optional setting are allowed in SPSS. Make the appropriate changes to set the options according to your choice.



19.6 BASIC STEPS IN DATA ANALYSIS

There are four basic steps involved in data analysis using SPSS. These are shown in the following figure.



Bring your data into SPSS: You can bring your data into SPSS in the following ways:

— Enter data directly into SPSS Data Editor.

— Open previously saved SPSS data file

— Read a spreadsheet data into SPSS data editor.

— Import data from DBF files

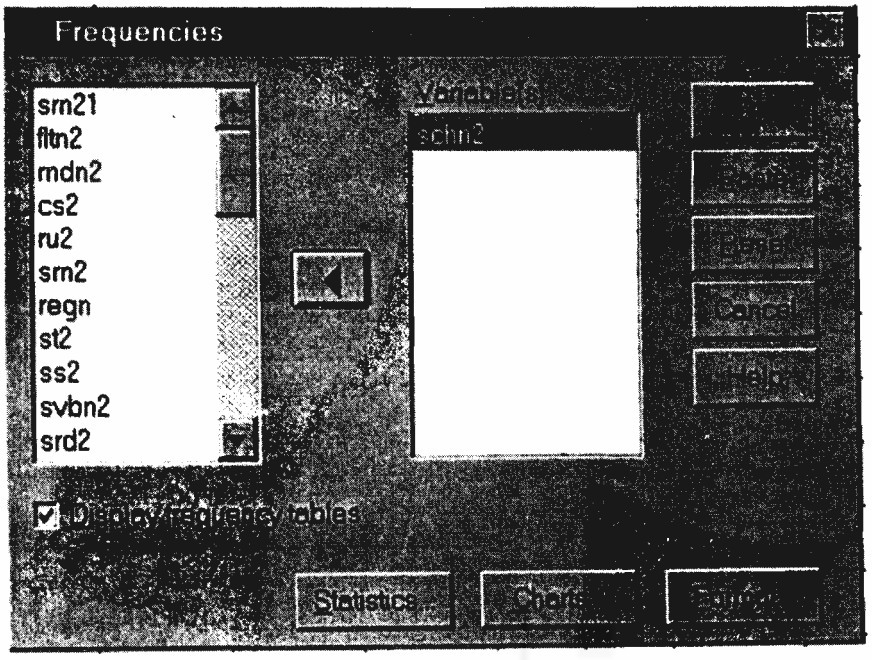
— Import data from RDBMS packages like Access, Oracle, Power, Builder, etc.

Select a Procedure from Menu: Before embarking on a statistical analysis, it is advised that you are clear as to what analysis is to be performed. Select the corresponding procedure to work on the data or create charts or tables using the selected procedure.

The command could either be directly executed or pasted on a Syntax Window. As mentioned earlier, pasting the command on the Syntax Window will be useful for undertaking batch processing or for subsequent use, especially where the same type of repetitive analysis required. Pasting the command will not lead to its execution. The command has to be selected and executed using the run command.

Select the variables: All the variables in the active file are listed each time a dialog box is opened. Select the appropriate variables for the selected procedure. Selection of at least one variable is necessary to run a statistical procedure. The

variables may be numeric, string, date or logical. You should be aware that string variables cannot be manipulated to the same extent as the numeric variables.



Run the Procedure and Examine the Output: After completing the selection process for the procedure and the variables, execute the SPSS command. Most of the commands are executed by clicking OK on the dialog box. The processor will execute the procedures and produce a report in the Output Navigator.

19.7 DEFINING, EDITING AND ENTERING DATA

As mention earlier, there are many options for creating SPSS data files. The data can either be directly entered through Data Editor or imported from spreadsheets, ASCII file and other RDBMS packages like Oracle and Access. The data is arranged in the form of rows and column in the Editor Window. The rows refer to the observations or cases and the columns to the variables. Each cell is defined as the intersection of a row and a column and refers to the value of a particular variable for a specific case/observation. While defining data, it is important to identify a primary key which is unique for each observation/case.

Variable Definition

Before entering the data into SPSS, it is advised that you define your variables. Such a definition will be very helpful for data entry and analysis stages. The following information about each variable is provided to define it:

— A name for the variable (upto 8 characters only)

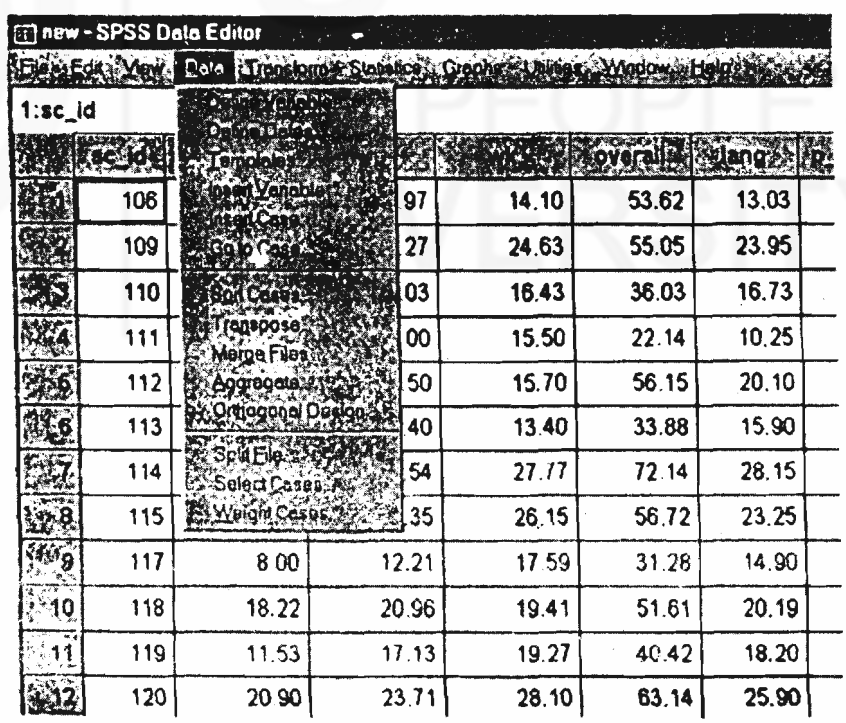
— A description (label)

— A series of labels which explain the value entered (value labels)

— A declaration as to which values are non-valid and should be excluded from the statistical analysis and other operations (missing values). The information is important to understand the response pattern and also to specify the observations which should be excluded from the analysis.

The following table provides an example of the above description.

Variable name	Variable labels	Value labels	Missing value	Variable type
STID	Student number	None	None	Number 6 Digit no decimal place
Name	None	None	None	String, 24 character long
Gender	Sex of respondent	M male F female X Unknown	X	String, 1 character long
MTL	Marital Status	1. Married 2. Widowed 3. Divorced 4. Separated 5. Never Married 9. Missing	9	Number, 1 character
DOB	Date of Birth	None	None	Date, dd/mm/yy

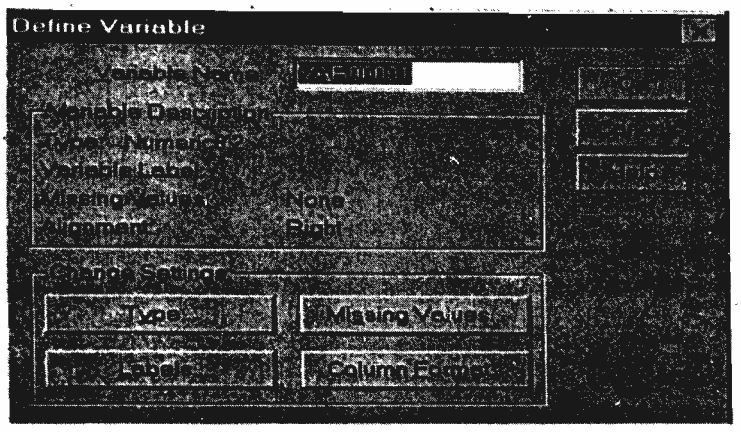


	sc_id				overall	lang	
1	106			97	14.10	53.62	13.03	
2	109			27	24.63	55.05	23.95	
3	110			03	16.43	36.03	16.73	
4	111			00	15.50	22.14	10.25	
5	112			50	15.70	56.15	20.10	
6	113			40	13.40	33.88	15.90	
7	114			54	27.77	72.14	28.15	
8	115			35	26.15	56.72	23.25	
9	117	8.00	12.21	17.59	31.28	14.90		
10	118	18.22	20.96	19.41	51.61	20.19		
11	119	11.53	17.13	19.27	40.42	18.20		
12	120	20.90	23.71	28.10	63.14	25.90		

Defining variables is easy in SPSS. The variable names can be changed and altered with ease even during analysis. Any change made to the working files will be permanently changed only when the data file is saved using 'save 'or 'save as' command. To start the procedure for defining variables, place the cursor in a particular column and from the menu click:

The following dialog box will appear.

Define variable (provide relevant information asked in the dialog box).



Data can be entered directly using SPSS Data Editor window. However, if the data is large, you are advised to use a data entry package. The data can also be edited/ changed in the data editor window. To change the value in any cell, bring the cursor to the particular cell, enter the new value and press enter. New variables can also be added and the existing variables can be deleted in the Data Editor Window.

19.8 DATA FILE MANAGEMENT FUNCTIONS

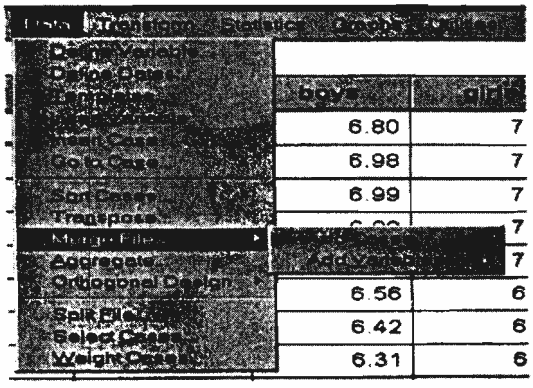
SPSS is very flexible as far as management of data files is concerned. While only one file can be opened for analysis at a time, the SPSS provides flexibility in merging multiple data files with the same structure into one single data file, merging files to add new variables, partially select the cases for analysis, make group of data based on certain characteristics and use different weights for different variables. Some of these functions are discussed below. Groups of data can also be defined to facilitate the analysis of the most commonly referred variables (see utilities and data commands).

19.8.1 Merging Data Files

Researchers are often faced with a situation where data from different files are to be merged or a limited number of variables from large complex data files are required. The following types of facility are available for merging files using SPSS.

Adding variables: Adding variables is useful when two data files contain the information about the same case but on different variables. For example, the teachers database may contain two files, one having the educational qualifications and the other having the names of the courses taught. Both the files could be combined to analyze the variables available in them. The data on a key and unique variable from both the files can be combined easily. The key variables must have the same name in both the data files. Both the data files should be sorted on the common key variable.

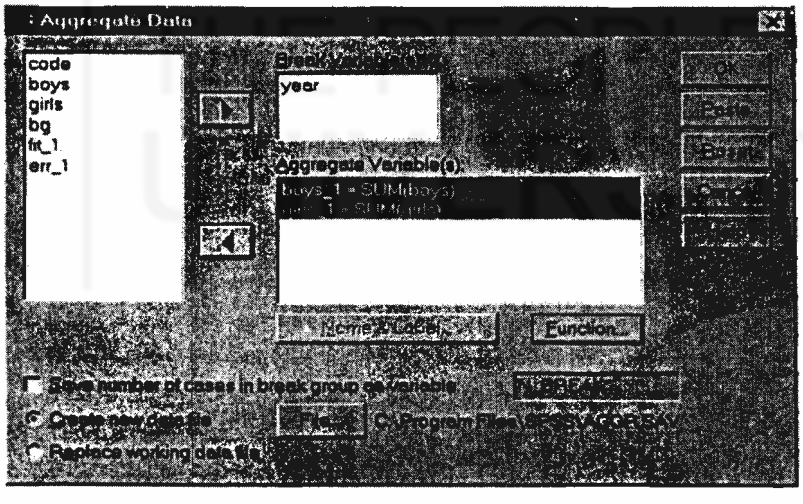
Adding cases: This option is used when the data from two files having the same variables are to be combined. For example, you may record the same information for students in different study centers in India abroad. The data can be merged to create a centralized database by using Add cases command.



19.8.2 Aggregate Data

Aggregate Data command combines groups of cases into a single summary case and creates a new aggregate data file. Cases are aggregated, based on the value of one or more grouping variables. The new (aggregated) file contains one record for each group. The aggregate file could be saved with a specific name to be provided by you. Otherwise, the default name is aggregate.sav. For example, the data on learners, achievement could be aggregated by sex, state and region.

A number of aggregate functions are available in the SPSS. These include sum, mean, number of cases, maximum value, minimum value, standard deviation and first and the last value. Other summary functions include percentage and fractions below and above a particular cut-off user-defined value.



19.8.3 Split File

The researcher is often interested in the comparison of a summary and other statistics based on certain group behaviour. For example, in a study of learning achievement, the researcher may be interested in comparing the mean scores for students belonging to different sex groups. The sex is taken as a grouping variable. Multiple grouping variables can also be selected. A maximum of eight grouping variables can be defined. Cases need to be sorted out by grouping variables. Two options are available for comparative analysis. These are: compare groups and organize output by groups. The split file is available under Data menu for making such comparisons.

19.8.4 Select Cases

Select case command can be used for selecting a random sub-sample or sub-group of cases based on a specified criteria that includes variables and complex expression. The following criteria are used for Select Case command.

Select if (condition is satisfied)

Variable values and their range

Date and time range

Arithmetic expressions

Logical expressions

Functions

Row numbers

Following the Select Case command, the unselected cases can either be deleted or temporarily filtered. Deleted cases are removed from the active file and cannot be recovered. You should be careful while selecting Delete option. Filtered option will be deleted temporarily. When the Select Case option is on, it is indicated in the Data Editor window.

Check Your Progress

Notes : a) Space is given below for your answer.

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

3. What are the basic steps in data analysis?

.....

.....

.....

.....

.....

.....

4. What is a split file?

.....

.....

.....

.....

19.9 RUNNING A PRELIMINARY ANALYSIS

Before running advanced statistical analysis it is important that you understand the salient features of your data. Use of statistical applications on a data set, the behaviour of which is not known, can give misleading conclusions. The following section explains the six characteristics which must be examined for a given data set before attempting an advanced analysis.

19.9.1 Six Characteristics of a Dataset

One strong argument for using computers and graphical presentation of the data is the advantage of viewing the data in a variety of ways. Preliminary exploration of data and its graphical presentation helps attain these objectives. The following characteristics will help you in deciding on the best plan for data management, analysis and presentation. SPSS includes commands for analyzing of data along the following lines.

Shape: The shape of the data will be the main factor in determining what set of summary statistics best explains the data. Shape is commonly categorised as symmetric, left-skewed or right-skewed, and an uni-modal, bi-modal or multi-modal. Frequency distribution, plots and graphical presentation of data histogram, P-P, Q-Q, scatter, Box-Plot are illustrative of the techniques that can be used for determining the shape of a data set. It is important that the user should have enough knowledge of the properties of various statistical distributions, their graphical presentations, characteristics and limitations.

Location: Location is simpler and more descriptive than measures of central tendency. Common measures of location are the mean and the median. Measures of central tendency also can be calculated for various sub-groups of a data set.

Spread: This measure describes the amount of variation in the data. Again approximate value is sufficient initially, with the measure of spread being informed by the shape of the data, and its intended use. Common measures of spread are variance, standard deviation and inner-quartile range. Percentile range is another measure which is used for measurement of dispersion.

Outliers: Outliers are data values that lie away from the general cluster of values. Each outlier needs to be examined to determine if it represents a possible value from the population being studied, in which case it should be retained, or if it is non representative (or an error) in which case it should be excluded. You should properly weigh and carefully examine the behaviour of outliers before accepting or rejecting of an observation/case. The best choice to display when looking for outlier is Box-plot. Range, i.e., maximum and minimum values can also be used to examine the behaviour of outliers.

Clustering: Clustering implies that data tend to bunch around certain values. Clustering shows most clearly on a dot-plot. Histogram, stem and leaf analysis are also important procedures to examine the clustering pattern of a data set.

Association and relationship: Researchers often look for associative characteristics or similarities and dissimilarities in the behaviour of some variables. For example, achievement scores and hours of study may be positively correlated whereas the teacher motivation and drop-out rate may be negatively associated with each other. Correlation coefficient is the most commonly used measure for understanding the nature and magnitude of association between two variables.

You should be clear that association does not imply relationship. A relationship is defined by the cause and effect type of link. Normally, there is one dependent variable and one or more than one independent variable in the cause and effect relationship. Cause and effect relationship is captured through regression analysis.

The analysis of data along the above lines provides considerable insight into the nature of data and also helps researchers in understanding key relationships between variables. It is assumed that the relationships are of linear type. Non-linear relationships can also be examined using non-linear techniques of analysis and also using data transformation techniques.

19.9.2 Data Transformation

Data transformation is a very useful aspect of SPSS. Using data transformation, you can collapse categories, record the data and create new variables based on complex equations and conditional statements. Some of the functions are detailed below:

Compute variable

- Compute value for numeric and string variables
- Create new variables or replace the value of existing variables. For the new variable, you can specify the variable type and label.
- Compute value selectively for sub-sets of data based on logical conditions.
- Use built-in functions, statistical functions, distribution functions and string functions.

Recode variables

Recoding of variables is an important characteristic of data management using SPSS. Many continuous and discrete variables need to be recoded for meaningful analysis. Recoding can be done either within the same variable or a new variable can be generated. Recoding in the same variable will replace and original values for this purpose. Recoding in a new variable will replace the old values with new values. The following example illustrates the need and use of recoding variables.

A survey of the primary school was conducted in Delhi. Along with other variables, information on the type of management was also collected. The management code was designed as follows:

1. Government
2. Local bodies
3. Private aided
4. Private unaided
5. Others

Lets us assume that a comparative analysis of the government and the private management schools is to be undertaken. This will be done by combining categories 1 and 2 and also 3 and 4. This can be achieved by recoding the management code as 1 (for 1 and 2 categories) and 2 for 3 and 4 categories into a new variable.

Assuming that a database on primary schools in Delhi is available, the enrolment analysis could be attempted by making suitable categories, i.e. schools with less than 50 students, 51-150,15-250 and more than 250 students. This could be achieved by recoding the enrolment variable into a new variable 'category'. If at a later stage in the analysis, it is found that a new category is to be introduced, it can again be achieved by recoding the enrolment data.

Count

Count is an important command available in SPSS and is used for counting occurrences of the same value(s) in a list if variables within the same case. For example, a survey might contain a list of books purchased (yes/no) by the students. You could count the number of 'yes' response, or a new variable can be generated which gives the value of count indicating the number of books bought.

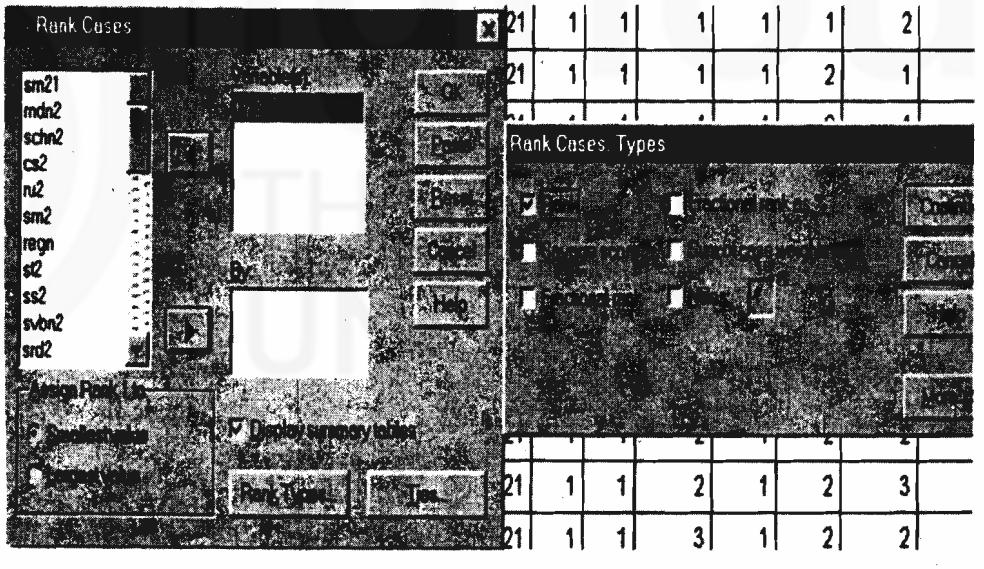
Procedure to run count command

- Choose Transform from the main menu
- Chose count
- Enter the name of a target variable (variable where the count value will be stored).
- Select two or more variables of the same type (numeric or string)
- Click define variable and specify which value(s) to be counted.
- Click OK after the selection has been made.

In survey on learners' achievement, the answer code to each question in language and mathematics could be recorded for each student. The codes could be '1' for the correct answer '2' for the wrong answer and '3' for no reply. Count command can then be used to count the number of correct answers.

Rank Cases

Rank Cases command can be used to rank observation in an ascending or a descending order. Other options available for ranking cases are shown in the right hand panel of the following figure.



19.9.3 Exploring Data

The Frequencies procedure provides statistics and graphic displays that are useful for describing many type of variables. Frequency counts, simple and cumulative percentages, mean, median and mode, sum, standard deviation, range minimum and maximum values, standard error of the mean, skewness and kurtosis, bar charts and pie charts, and histograms are some of the methods used to explore the data before a sophisticated and advanced analysis is undertaken.

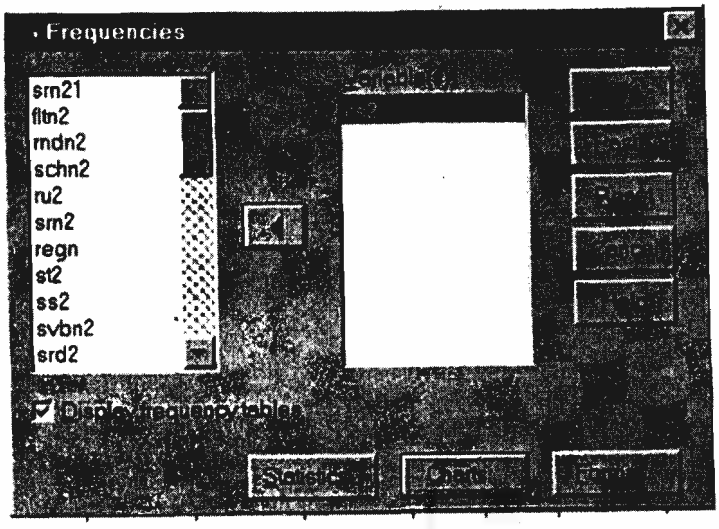
If you want to compare summary statistics for each of the several groups of cases, use split file on the 'Data' menu. Use of Explore, Summarize or means procedure is recommended for initial exploration of data. Use the following commands to obtain frequencies:

From the menu choose:

Statistics

Summarize

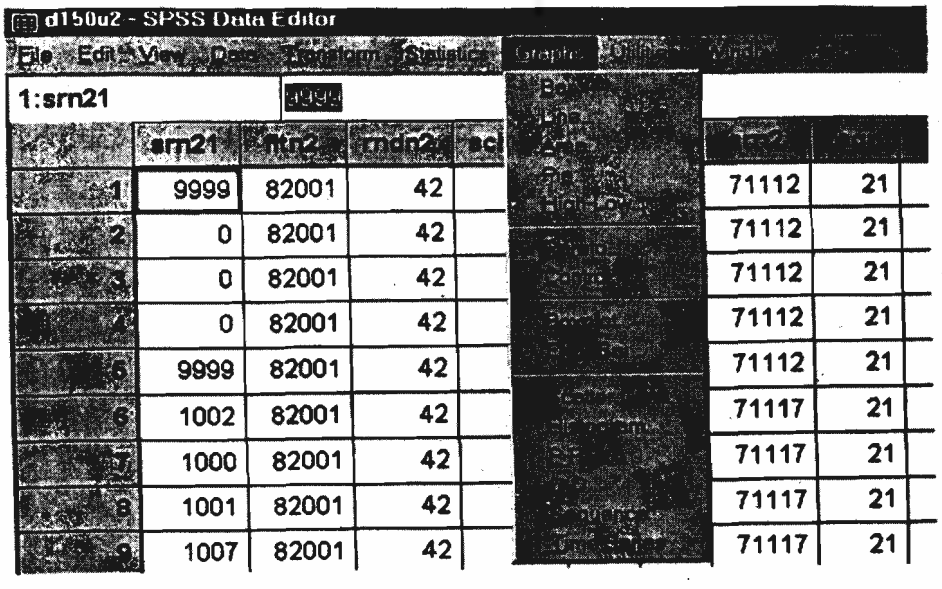
Frequencies



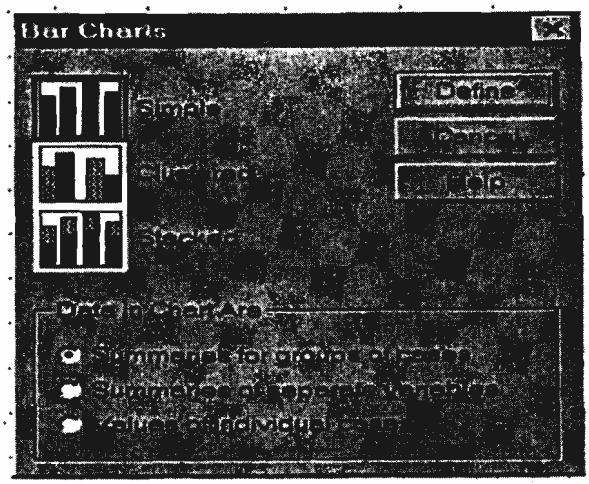
Use the Statistics and Charts sub-commands (as shown in the above figure) to select the desired features. More than one variable could be selected for frequency distribution. You must remember, that before attempting frequency distribution, recoding of continuous type of variables will be necessary.

19.9.4 Graphical Presentation of Data

SPSS offers extensive facilities for viewing the data and its key features in high-resolution charts and plots. From the main menu, select Graphs and the following screen appears. Various types of Graph that can be drawn using SPSS are indicated in the sub-commands.



Select a chart type from the Graph menu. This opens a chart dialog box as shown below:

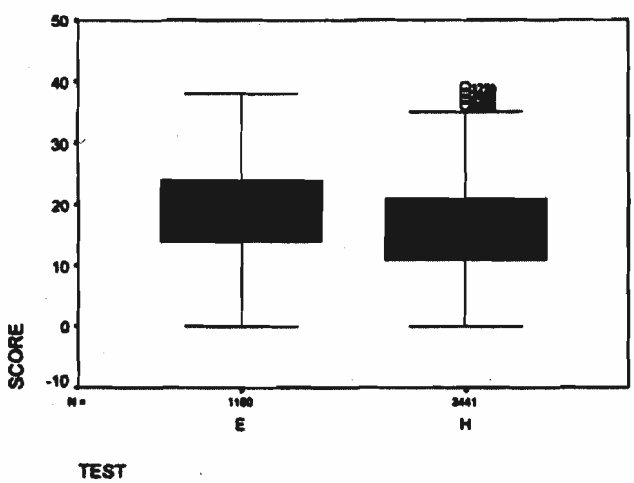


After the appropriate selections have been made, the output is displayed in Output Navigator window. The chart can be modified by a double click on any part of the chart. Some typical modifications include the following:

- Edit axis titles and labels and footnotes
- Change Scale (X–Y)
- Edit the legend
- Add or modify a title
- Add annotation
- Add an outer frame

Another important category of charts is High –Low which are often used to represent variables like maximum and minimum temperature in a day, stock market behaviour or other similar variables.

Box-plot and Error Bar charts helps you to visualize distribution and dispersion. Box-plot displays the median and quartiles and special symbols are used to identify outliers, if any. Error Bar charts displays the mean and confidence intervals or standard errors. To obtain a box-plot, choose Box-plot from the Graphs menu. The simple box-plot for mean scores obtained in English and Hindi is shown in the following diagram:



The above figure shows that there were a large number of outliers in the case of Hindi scores as compared to English. The outliers were along the higher side. This shows that many students were scoring very high marks. The size (numbers) of cases are shown along the X-axis. The boxes show the median and the quartile values for both the tests.

19.9.5 Scatterplots and Histograms

Scatterplots highlight the relationship between two quantitative variables by plotting the actual values along X-Y axis. The scatterplots are useful to examine the actual nature of relationship between these variables. This could be either linear or non-linear in form. To help visualize the relationship, you can add a simple linear or a quadratic regression line. A 3-D scatter plot adds a third variable in the relationship. You can rotate the two dimensional projection of the three dimensions to delineate the underlying patterns. In order to obtain a scatter plot, select Scatter from the Graphs option.

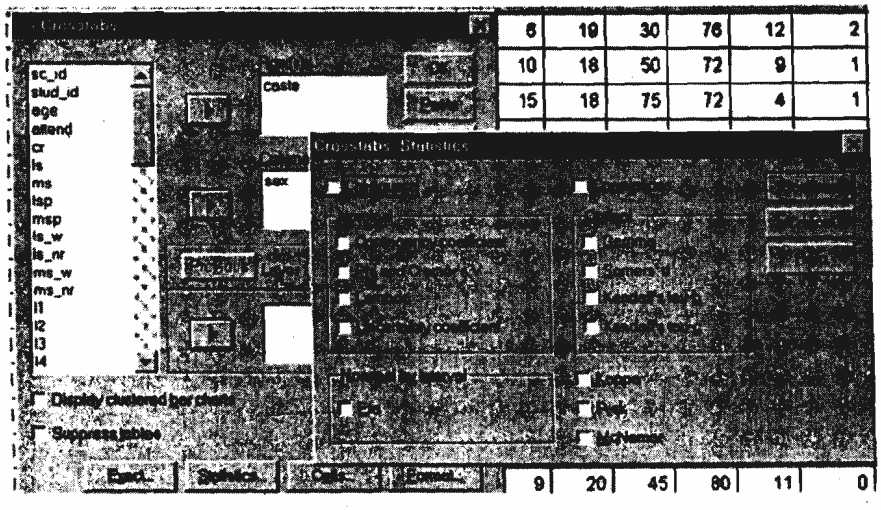
A histogram will be obtained by selecting histogram option from the Graphs menu. The variable for which a histogram is to be obtained should be selected from the dialog box. The normal curve can also be displayed along with the histogram to visually see the extent of similarity between the actual distribution of values and the normal curve.

Pareto and control charts are used to analyze and improve the quality of an ongoing process. You may refer to the SPSS manuals for use of these techniques.

19.10 UNDERSTANDING RELATIONSHIP BETWEEN VARIABLES

The foregoing details focused on the techniques of analysis describing the behaviour of individual variables. However, most of the research studies require relationship between two or more variables to be examined. For example, one may be interested in questions like, "do the achievement scores of boys and girls in the same class differ?"

Cross-tabulation is the simplest procedure to describe a relationship between two or more categories of variables. Cross-tabulation is useful for any type of categorical variable, especially, when the categories are small and mutually exclusive. Some variables could be aggregated into convenient categories by using the Recode command.



The cells in the standard two-way frequency table display the counts or the number of cases falling into the categories distinguished by the row and column variables. There is no category showing the missing data in a two-way classification.

The SPSS also provides for a number of options while displaying the results of cross-tabulations. These relate to percentage distribution of frequencies/cases in terms of row total, column total and grand total. Any or all of these options can be selected. Each of the options can be selected depending upon the objective of analysis. The following table shows the distribution of students by their sex in a simple study.

The simple output of the cross-tabulation procedure is shown below.

Cast * Sex Cross Tabulation

Count		Sex		
		Male	Female	Total
Caste	SC/ST	204	322	528
	OBC	86	136	222
	General	1256	1274	2530
Total		1546	1732	3278

19.10.1 The Mean Procedure

The Mean procedure is very useful tool when the average value of a variable is to be computed based on sub-division of the data into groups based on the value of some other variable. For example, you may be required to compute average achievement score of children based on their age. Or you may be required to compute average monthly income of the respondents by their occupations and experience. While the mean procedure has immense use in understanding the sub-group behaviour, it also suffers from certain limitations. It cannot be used in the case of categorical variables. Moreover, the sub-groups should have a reasonably large number of value for the mean value to be representative. The specifications for a subgroup average are:

Place the continuous variables in the 'Dependent list'

Place the categorical variables in the 'Independent list'.

The mean scores in Mathematics by caste groups as obtained using the Mean procedure are given below:

Report

Score Math		
Mean	SC/ST	17.76
	OBC	17.51
	General	20.09
	Total	19.54
N	SC/ST	526
	OBC	222
	General	2530
	Total	3278
Std. Deviation	SC/ST	6.40
	OBC	6.20
	General	5.11
	Total	5.50

19.10.2 Linear Regression

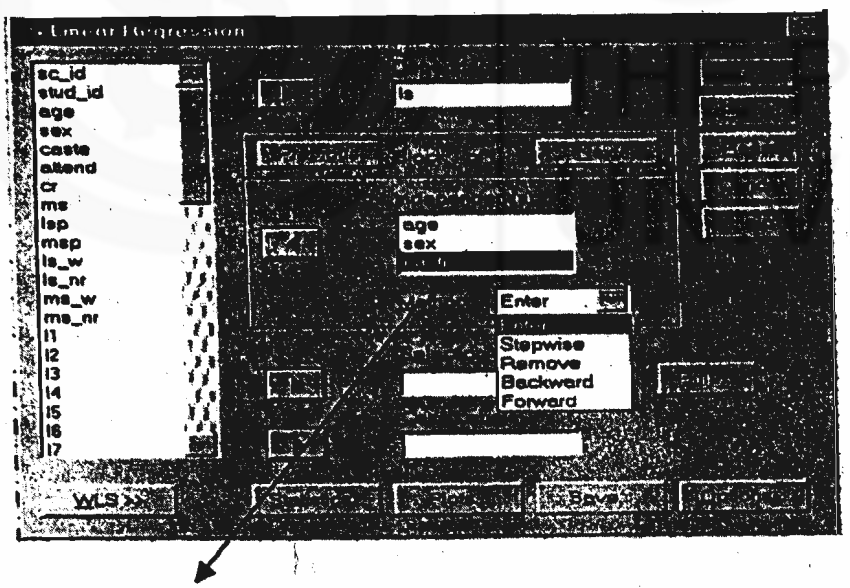
How do you predict the sales of ice cream in the coming summer season? What are the important determinants of achievement in government schools? Is there any relationship between educational attainment and per capita income of a household? These are the types of question which are often asked by development planners and policy analysts. Regression analysis is a technique to address some of these questions.

Linear regression is the most commonly used procedure for the analysis of a cause and effect relationship between one dependent variable and a number of independent variables. The dependent and independent variables should be quantitative. Categorial variables like sex and religion should be recorded to dummy (binary) variables and other types of contrast variables. An important assumption of the regression analysis is that the distribution of the dependent variable is normal. Moreover, the relationship between the dependent and all the independent variables should be linear and all observations should be independent of each other.

SPSS provides extensive scope for regression analysis using various types of selection processes.

The method of selecting of independent variables for linear regression analysis is an important choice which the researcher should consider before running the analysis. You can construct a variety of regression models from the same set of variables by using different methods.

You can enter all the variables in a single step or enter the independent variables selectively.



Variable selection method:

It allows you to specify how independent variables are entered into the regression analysis. The following options are available:

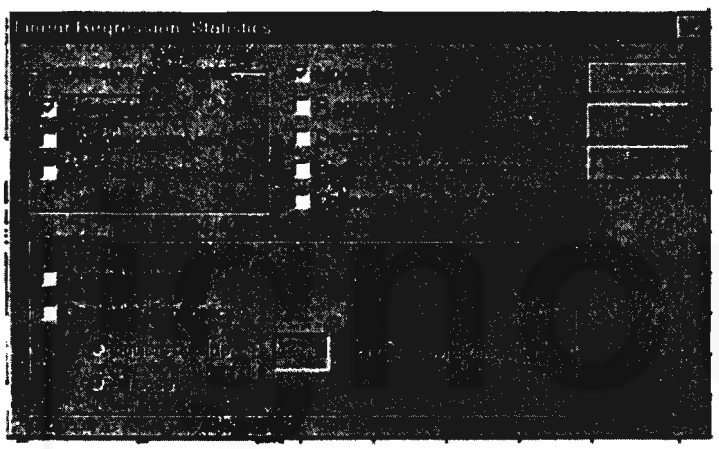
- Enter: To enter all the variables in a single step, select Enter option.
- Remove: To remove the variables in a block in a single step.
- Forward: It enters one variable at a time based on the selected criterion.

- Backward: All variables are entered in the first instance and then one variable is removed at a time on the selected criterion.
- Stepwise: Stepwise variable entry and removal examines the variables in the block at each step for entry and removal. This is a forward step procedure.

All the variables must pass the tolerance criterion to be entered in the equation, regardless of the entry method specified. The default tolerance limit is 0.0001. A new variable will not be entered if it is causes the tolerance of another variable already entered to be dropped below the tolerance limit.

Linear Regression Statistics

The following statistics are available on linear regression models. Estimates and Model Fit are the two options which are selected by default.



Regression coefficient: The Estimates option displays regression coefficient, *B*, standard error, standard coefficient beta, t-value, and two tailed significance level of *t*. Covariance matrix displays a variance-covariance matrix of regression coefficients with covariance off the diagonal and variance off the diagonal. A correlation matrix will also be displayed.

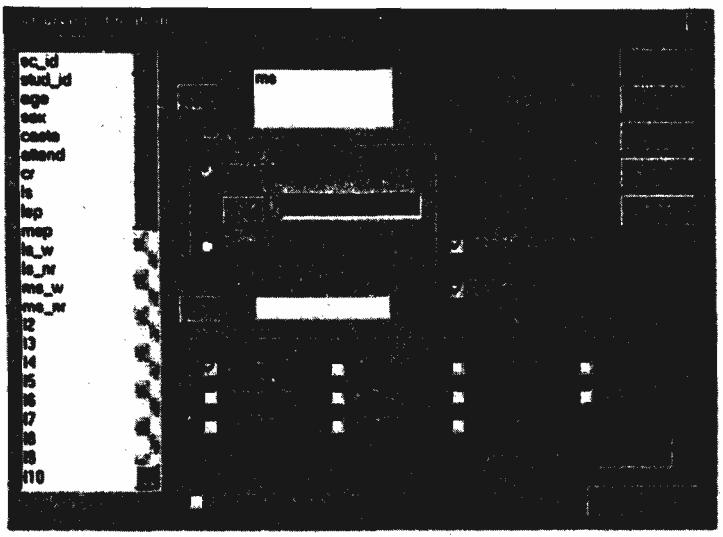
Model fit: The variables entered and removed from the model are displayed. Goodness of fit statistics, R-square, multiple R, and adjusted R square, standard error of the estimate and an analysis of variance table is displayed.

If other options are ticked, the statistics corresponding to each of the options are also displayed in the output Navigator.

If the data does not show linear relationship and the transformation procedure does not help, try using Curve Estimation procedure

19.10.3 Curve Estimation

There are many situations when the researcher is not sure about the nature of the curve that fits in a given data set. In such cases, Curve Estimation command is used to fit various types of curve on a given data. After examining the results can decide on the best fit equation. The SPSS includes 11curve estimation regression models. A separate model is produced for each dependent variable. It is recommended that before running the curve estimation procedure, you should examine the graphical output to ascertain how the independent and dependent variables appear to be related to each other. The linear relationship assumes that the dependent variable will be normally distributed.



A scatter plot of learning achievement may reveal that the relationship between the mean score and the time spent on a task is linearly related. You might like to fit a linear model to the data and check the validity of the assumption of the model. It is quite possible that a non-linear model may give the best fit.

19.11 NON-PARAMETERIC TESTS

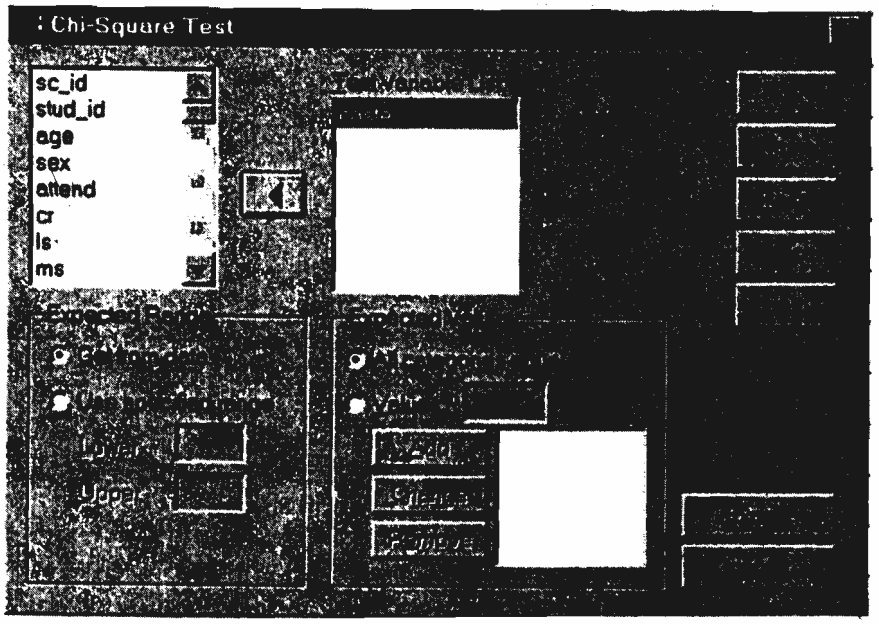
The non-parametric test procedure provides several tests that do not require assumptions about the shape of the underlying distribution. These include the following most commonly used test:

- Chi-square test
- Binomial test
- Run test
- One sample kolmogorov Seminov test
- Two independent Sample tests
- Tests for several independent samples
- Two related sample tests
- Tests for several related samples.

Here, we shall discuss the procedure for Chi-square test only. You are advised to consult the SPSS users' manual and other statistical books for detailed discussion on the other tests.

Chi-Square

Chi-square test is the most commonly used test in educational research. This test compares the observed and the expected frequencies in each cell/category to test either that all categories contain the same proportion of values or that each category contains a user specified proportion of values.



Consider that a bag contains red, white and yellow balls. You want to test the hypothesis that the bag contains all types of balls in equal proportion. To obtain Chi-square test, choose Chi-square from Non-parametric tests in the Statistics command. Select one or more variables. Each variable produces a separate output.

By default, all categories have equal expected values as shown in the above figure. Categories can have user specified proportion also. In order to provide user specific expected values, select the values option and add the user expected values. The sequence in which the values are entered is very important in this case. It corresponds to the ascending order of the category values of the test variable.

19.12 SPSS PRODUCTION FACILITY

The SPSS Production facility provides the ability to run SPSS in an automated mode. SPSS runs unattended and uninterrupted and terminates after executing the last command. Production mode is useful if you run the same set of time-consuming analysis periodically.

The SPSS Production facility uses command syntax file to tell SPSS about the commands to be executed. We have already discussed the important features of the command syntax. The command syntax file can be edited in a standard text editor.

To run the SPSS Production facility, quit the SPSS if it is already running. SPSS Production facility cannot be run when SPSS is running. Start SPSS Production program from the start window of window 2000/XP. Specify the syntax file that you want to use in the production job. Click Browse to select the Syntax File. Save the production file job. Run the production file job at any time.

19.13 STATISTICAL ANALYSIS SYSTEM (SAS)

Like the SPSS, the Statistical Analysis System (SAS) package calculate descriptive statistic of your choice e.g., Mean, Standard Deviation etc. SAS is available for both mainframe and personal computers. It is strong in its treatment of data, in clarity of its graphics and in certain business applications. The various statistical procedures carried out by SAS are always preceded by the word PROC which

stand for procedure. The most commonly used SAS statistical procedures are as follows: (Sprinthall et.al, 1991).

- PROC MEANS: Descriptive statistics (mean, standard deviation, maximum and minimum values and so on).
- PROC CORR: Pearson correlation between two or more variables.
- PROC t-TEST: t-test for significant difference between the means of two groups.
- PROC ANOVA: Analysis of variance for all types of designs (one way, two-way and other).
- PROC FREQ: Frequency distribution for one or more variables.

As pointed out by Klieger (1984) SAS package is comparatively more difficult to use due to its procedural complexities. For greater details on SAS package you are advised to consult the books by Klieger and Sprinthall.

19.14 NUDIST

Computer programs help in the analysis of qualitative data, especially in understanding a large (say 500 or more pages) text database, Studies using large databases such as ethnographies with extensive interviews, computer programs provide an invaluable aid in research.

NUDIST (Non-numerical unstructured data indexing, searching and theorizing) program was developed in Australia in1991. This package is used for qualitative analysis of data. Here we present briefly the main features of this package. This software requires, 4 megabytes of RAM and atleast 2 megabytes space for data files in your PC or MAC. In your PC it operates under windows (Creswell 1998).

As a researcher this software will help you to provide the following.

1. Storing and organizing files: First establish document files and store information with the NUDIST programme. Document files consist of transcript from an interview, notes of observation or any article scanned from a newspaper.
2. Searching for themes: Tag segments of text from all the documents that relate to a single idea or theme. For example, distance learners, in a study on effectiveness of distance education talk about the role of academic counsellors. The researcher can create a node in NUDIST as 'Role of Academic Counsellors'. Researcher will select text in the transcript where learners have talked about this role and merge it into role of Academic Counsellors. Information can be retained in this node and researcher can take print in different ways in which learners talk about the role of Academic Counsellors.
3. Crossing themes: Taking the same example of roll of counsellors, the researcher can relate this node to other nodes. Suppose the other node is qualifications of counsellor. There are two categories like Graduate and Post Graduate. The researcher will ask NUDIST to cross the two categories, role of counsellors and qualification of counsellors to see for example whether there is any relation between graduate counsellor and their roll than the post graduate counsellor and their roll. NUDIST software generates information for a matrix with information in the cells reflecting different perspectives.
4. Diagramming: In this package, once the information is categorized, categories are identified. These categories are developed into nine visual picture of the

categories that display their inter connectedness. This is called a tree diagram in NUDIST software. Tree diagram is a hierarchical tree of categories where root node is the top and parents and siblings in the tree. This tree diagram is a useful device for discussing the data analysis of quantitative research in conference.

5. Creating a template: In a qualitative research, at the beginning of data analysis, the researcher will create a template which is apriori code book for organizing information.

For further details on NUDIST software you may like to consult the following.

Kelle, E.(ed.), *Computer-aided Qualitative Data Analysis,* Thousand Oaks, CA: Sage, 1995.

Tesch, R., *Qualitative Research: Analysis Types and Software Tools,* Bristol, PA: Falmer, 1990.

Check Your Progress

Notes : a) Space is given below for your answer.

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

5. Name the six characteristics of data sheet.

.....

.....

.....

.....

.....

6. What is NUDIST?

.....

.....

.....

.....

19.15 LET US SUM UP

The foregoing details examined the various types of statistical application of the SPSS in data management, presentation and analysis. The discussion was based on the assumption that you have a basic understanding of the statistical methods. It was highlighted that the researchers must try to explore the data using various simple but powerful statistical techniques. It this connection, six characteristics of the data were examined for exploring it fully. The procedures involved in the use of various statistics were also discussed in detail. Procedures for running regression analysis to understand the relationship between variables were also discussed. Those of you who are comfortable with the basic statistical procedures in SPSS can explore the advanced features, including those aimed at automating the statistical analysis using the SPSS Production facility and also the use of scripting in data analyzing.

19.16 UNIT-END ACTIVITIES

1. Obtain quantitative data for any kinds of research work with the help of research tools and analyse the data using SPSS package.

2. Analyse the data collected for your disseration work using SPSS package.

19.17 SUGGESTED READINGS

SPSS Base 7.5 for Windows, User's Guide, SPSS Inc.

SPSS Base 7.5 Application Guide, SPSS Inc.

SPSS Advanced Statistics 7.5, SPSS Inc.

A number of white papers dealing with various applications are available on the SPSS website: *www.spss.com*. This site is updated regularly with new materials. Advanced user may like to obtain/download the relevant materials for their use.

Creswell, John, W. (1998): *Qualitative Inquiry and Research Design: Choosing Among Five Traditions.* Sage Publications, Inc., International Educational and Professional Publisher, USA.

Klieger, D.M.(1984): *Computer Usage for Social Scientists.* Newton, Mass: Allyn and Bacon.

Sprinthall, Richard C., (et.al.) (1991): *Understanding Educational Research.* New Jersey: Prentice Hall.

19.18 ANSWERS TO CHECK YOUR PROGRESS

1. a) SPSS table creates a high quality presentation in tabular form.

   b) SPSS Trends performs comprehensive forecasting and time series analysis.

   c) SPSS categories performs conjoint analysis and optional scaling procedures.

2. There are seven types of windows in SPSS. These are data editor, output navigator, pivot tables, graphics, text editor, syntax editor and script editor.

3. i) Bring your data into SPSS

   ii) Select a procedure from Menu

   iii) Select the variables

   iv) Run the procedure and examine the output

4. Split file is meant for comparing a summary and other statistics based on certain group behaviour.

5. i) shape, ii) Location, iii) Spread, iv) Outliers, v) Clustering vi) Association and relationship.

6. NUDIST (Non-numerical Unstructured Data Indexing, Searching and Theorizing) is a software package for qualitative data analysis.

APPENDIX

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

x	*.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	9675	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	2764	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	3749	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	4920	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	4983	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	4999.9971333									

Table B : Ordinates of the normal probability curve expressed as fractional parts of the mean ordinate, Y.

x	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
0.0	00000	99995	99980	99955	99920	99875	99820	99755	99685	99596
0.1	99501	99396	99283	99158	99025	98881	98728	98565	98393	98211
0.2	98020	97819	97609	97390	97161	96923	96676	96420	96156	95882
0.3	95600	95309	95010	94702	94387	94055	93723	93382	93024	92677
0.4	92312	91399	91558	91169	90774	90371	89961	89543	89119	88688
0.5	88250	87805	87353	86896	86432	85962	85488	85006	84519	84060
0.6	83527	83023	82514	82010	81481	80957	80429	79896	79359	78817
0.7	78270	77721	77167	76610	76048	75484	74916	74342	73769	73193
0.8	72615	72033	71448	70861	70272	69681	69087	68493	67896	67298
0.9	66689	66097	65494	64891	64287	63683	63077	62472	61865	61259
1.0	60653	60047	59440	58834	58228	57623	57017	56414	55810	55209
1.1	54607	54007	53409	52812	52214	51629	51027	50437	49848	49260
1.2	48675	48092	47511	46933	46357	45783	45212	44644	44078	43516
1.3	42956	42399	41845	41294	40747	40202	39661	39123	38569	38058
1.4	37531	37007	36487	35971	35459	34950	34445	33944	33447	32954
1.5	32465	31980	31500	31023	30550	30082	29618	29158	28702	28251
1.6	27804	27361	26923	26489	26059	25634	25213	24797	24385	23978
1.7	23575	23176	22782	22392	22008	21627	21251	20879	20511	20148
1.8	19790	19436	19086	18741	18400	18064	17732	17404	17081	16762
1.9	16448	16137	15831	15530	15232	14939	14650	14364	14083	13806
2.0	13534	13265	13000	12749	12483	12230	11981	11737	11496	11259
2.1	11025	10795	10570	10347	10129	09914	09702	09495	09290	09090
2.2	08892	08698	08507	08320	08136	07956	07778	07604	07433	07265
2.3	07199	06939	06780	06624	06471	06321	06174	06029	05888	05750
2.4	05614	05481	05350	05222	05096	04973	04852	04734	04618	04505
2.5	04394	04285	04179	04074	03972	03873	03775	03680	03586	03494
2.6	03405	03317	03232	03148	03066	02986	02908	02831	02757	02684
2.7	02612	02542	02474	02408	02343	02280	02218	02157	02098	02040
2.8	01984	01929	01876	01823	01772	01723	01674	01627	01581	01536
2.9	01492	01449	01408	01367	01328	01288	01252	01215	01179	01145
3.0	01111	00819	00598	00432	00309	00219	00153	00106	00073	00050
4.0	00034	00022	00015	00010	00006	00004	00003	00002	00001	00001
5.0	00000									

Table C : Table of Critical Values of t.

Degrees of Freedom	*Level of Significance*			
	0.10	0.05	0.02	0.01
1	6.34	12.71	31.82	63.66
2	2.92	4.30	6 .96	9.62
3	2.35	3.18	4.54	5.84
4	2.13	2.78	3.75	4.60
5	2.02	2.57	3.36	4.03
6	1.94	2.45	3.14	3.71
7	1.90	2.36	3.00	3.50
8	1.86	2.31	2.90	3.36
9	1.83	2.26	2.82	3.25
10	1.81	2.23	2.76	3.17
11	1.80	2.20	2.72	3.11
12	1.78	2.18	2.68	3.06
13	1.77	2.16	2.65	3.01
14	1.76	2.14	2.62	2.98
15	1.75	2.13	2.60	2.95
16	1.75	2.12	2.58	2.92
17	1.74	2.11	2.57	2.90
18	1.73	2.10	2.55	2.88
19	1.73	2.09	2.54	2.86
20	1.72	2.09	2.53	2.84
21	1.72	2.08	2.52	2.83
22	1.72	2.07	2.51	2.82
23	1.71	2.07	2.50	2.81
24	1.71	2.06	2.49	2.80
25	1.71	2.06	2.48	2.79
26	1.71	2.06	2.48	2.78
27	1.70	2.05	2.47	2.77
28	1.70	2.05	2.47	2.76
29	1.70	2.04	2.46	2.76
30	1.70	2.04	2.46	2.75
35	1.69	2.03	2.44	2.72
40	1.68	2.02	2.42	2.71
45	1.68	2.02	2.41	2.69
50	1.68	2.01	2.40	2.68
60	1.67	2.00	2.39	2.66
70	1.67	2.00	2.38	2.65
80	1.66	1.99	2.38	2.64
90	1.66	1.99	2.37	2.63
100	1.66	1.98	2.36	2.63
125	1.66	1.98	2.36	2.62
150	1.66	1.98	2.35	2.61
200	1.65	1.97	2.35	2.60
300	1.65	1.97	2.34	2.59
400	1.65	1.97	2.34	2.59
500	1.65	1.96	2.33	2.59
1000	1.65	1.96	2.33	2.58
	1.65	1.96	2.33	2.58

Table D : F-ratios for .05 (roman) and .01 (bold face) Levels of Significance

	Degrees of freedom for greater mean square									
	1	*2*	*3*	*4*	*5*	*6*	*8*	*12*	*24*	
1	161.45	199.50	215.72	224.57	230.17	233.97	238.89	243.91	249.04	254.32
	4052.10	**4999.03**	**5403.49**	**5625.14**	**5764.08**	**5859.39**	**5981.34**	**6105.83**	**6234.16**	**6366.48**
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
	98.49	**99.01**	**99.17**	**99.25**	**99.30**	**99.33**	**99.36**	**99.42**	**99.46**	**99.50**
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.64	8.53
	34.12	**30.81**	**29.46**	**28.71**	**28.24**	**27.91**	**27.49**	**27.05**	**26.60**	**26.12**
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.77	5.63
	21.20	**18.00**	**16.69**	**15.98**	**15.52**	**15.21**	**14.80**	**14.37**	**13.93**	**13.46**
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68	4.53	4.36
	12.26	**13.27**	**12.06**	**11.39**	**10.97**	**10.67**	**10.27**	**9.89**	**9.47**	**9.02**
6	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00	3.84	3.67
	13.74	**10.92**	**9.78**	**9.15**	**8.75**	**8.47**	**8.10**	**7.72**	**7.31**	**6.88**
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
	12.25	**9.55**	**8.45**	**7.85**	**7.46**	**7.19**	**6.84**	**6.47**	**6.07**	**5.65**
8	5.32	4.46	4.07	3.84	3.69	3.58	3.41	3.28	3.12	2.93
	11.26	**8.65**	**7.59**	**7.01**	**6.63**	**6.37**	**6.03**	**5.67**	**5.28**	**4.86**
9	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07	2.90	2.71
	10.56	**8.02**	**6.99**	**6.42**	**6.06**	**5.80**	**5.47**	**5.11**	**4.73**	**4.31**
10	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91	2.74	2.54
	10.04	**7.56**	**6.55**	**5.99**	**5.64**	**5.39**	**5.06**	**4.71**	**4.33**	**3.91**
11	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79	2.61	2. 40
	9.65	**7.20**	**6.22**	**5.67**	**5.32**	**5.07**	**4.74**	**4.40**	**4.02**	**3.60**

12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.50	2.30
	9.33	**6.93**	**5.95**	**5.41**	**5.06**	**4.82**	**4.50**	**4.16**	**3.78**	**3.36**
13	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60	2.42	2.21
	9.07	**6.70**	**5.74**	**5.20**	**4.86**	**4.62**	**4.30**	**3.96**	**3.59**	**3.16**
14	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53	2.35	2.13
	8.86	**6.51**	**5.56**	**5.03**	**4.69**	**4.46**	**4.14**	**3.80**	**3.43**	**3.00**
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48	2.29	2.07
	8.68	**6.36**	**5.42**	**4.89**	**4.56**	**4.32**	**4.00**	**3.67**	**3.29**	**2.87**
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42	2.24	2.01
	8.53	**6.23**	**5.29**	**4.77**	**4.44**	**4.20**	**3.89**	**3.55**	**3.18**	**2.75**
17	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38	2.19	1.96
	8.40	**6.11**	**5.18**	**4.67**	**4.34**	**4.10**	**3.79**	**3.45**	**3.08**	**2.65**
18	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34	2.15	1.92
	8.28	**6.01**	**5.09**	**4.58**	**4.25**	**4.01**	**3.71**	**3.37**	**3.01**	**2.57**
19	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31	2.11	1.88
	8.18	**5.93**	**5.01**	**4.50**	**4.17**	**3.94**	**3.63**	**3.30**	**2.92**	**2.49**
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28	2.08	1.84
	8.10	**5.85**	**4.94**	**4.43**	**4.10**	**3.87**	**3.56**	**3.23**	**2.86**	**2.42**
21	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25	2.05	1.81
	8.02	**5.78**	**4.87**	**4.37**	**4.04**	**3.81**	**3.51**	**3.17**	**2.80**	**2.36**
22	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23	2.03	1.78
	7.94	**5.72**	**4.82**	**4.31**	**3.99**	**3.75**	**3.45**	**3.12**	**2.75**	**2.30**
23	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20	2.00	1.76
	7.88	**5.66**	**4.76**	**4.46**	**3.94**	**3.71**	**3.41**	**3.07**	**2.70**	**2.26**

24	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18	1.98	1.73
	7.82	**5.61**	**4.72**	**4.22**	**3.90**	**3.67**	**3.36**	**3.93**	**2.66**	**4.21**
25	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16	1.96	1.71
	7.77	**5.57**	**4.68**	**4.18**	**3.86**	**3.63**	**3.32**	**2.99**	**2.62**	**2.17**
26	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15	1.95	1.69
	7.72	**5.53**	**4.64**	**4.14**	**3.82**	**3.59**	**3.29**	**2.96**	**2.58**	**2.13**
27	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13	1.93	1.67
	7.68	**5.49**	**4.60**	**4.11**	**3.78**	**3.56**	**3.26**	**2.93**	**2.55**	**2.10**
28	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12	1.91	1.65
	7.64	**5.45**	**4.57**	**4.07**	**3.75**	**3.53**	**3.23**	**2.90**	**2.52**	**2.06**
29	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10	1.90	1.64
	7.60	**5.42**	**4.54**	**4.04**	**3.73**	**3.50**	**3.20**	**2.87**	**2.49**	**2.08**
30	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2 .09	1.89	1.62
	7.56	**5.39**	**4.51**	**4.02**	**3.70**	**3.47**	**3.17**	**2.84**	**2.47**	**2.01**
35	4.12	3.26	2.87	2.64	2.48	2.37	2.22	2.04	1.83	1.57
	7.42	**5.27**	**4.40**	**3.91**	**3.59**	**3.37**	**3.07**	**2.74**	**2.. /**	**1.90**
40	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00	1.79	1.52
	7.31	**5.18**	**4.31**	**3.83**	**3.51**	**3.29**	**2.99**	**2.66**	**2.29**	**1.82**
45	4.06	3.21	2.81	2.58	2.42	2.31	2.15	1.97	1.76	1.48
	7.23	**5.11**	**4.25**	**3.77**	**3.45**	**3.23**	**2.94**	**2.61**	**2.23**	**1.75**
50	4.03	3.18	2.79	2.56	2.40	2.29	2.13	1.95	1.74	1.44
	7.17	**5.06**	**4.20**	**3.72**	**3.41**	**3.19**	**2.89**	**2.56**	**2.18**	**1.68**
60	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92	1.70	1.39
	7.08	**4.98**	**4.13**	**3.65**	**3.34**	**3.12**	**2.82**	**2.50**	**2.12**	**1.60**
70	3.98	3.13	2.74	2.50	2.35	2.23	2.07	1.89	1.67	1.35
	7.01	**4.92**	**4.07**	**3.60**	**3.29**	**3.07**	**2.78**	**2.45**	**2.07**	**1.53**

80	3.96	3.11	2.72	2.49	2.33	2.21	2.06	1.88	1.65	1.31
	6.96	**4.88**	**4.04**	**3.56**	**3.26**	**3.04**	**2.74**	**2.42**	**2.03**	**1.47**
90	3.95	3.10	2.71	2.47	2.32	2.20	2.04	1.86	1.64	1.28
	6.92	**4.85**	**4.01**	**3.53**	**3.23**	**3.01**	**2.72**	**2.39**	**2.00**	**1.43**
100	3.94	3.09	2.70	2.46	2.30	2.19	2.03	1.85	1.63	1.26
	6.90	**4.82**	**3.98**	**3.51**	**3.21**	**2.99**	**2.69**	**2.37**	**1.98**	**1.39**
125	3.92	3.07	2.68	2.44	2.29	2.17	2.01	1.83	1.60	1.21
	6.84	**4.78**	**3.94**	**3.47**	**3.17**	**2.95**	**2.66**	**2.33**	**1.94**	**1.32**
150	3.90	3.06	2.66	2.43	2.27	2.16	2.00	1.82	1.59	1.18
	6.81	**4.75**	**3.91**	**3.45**	**3.14**	**2.92**	**2.63**	**2.31**	**1.92**	**1.27**
200	3.89	3.04	2.65	2.42	2.26	2.14	1.98	1.80	1.57	1.14
	6.76	**4.71**	**3.88**	**3.41**	**3.11**	**2.89**	**2.60**	**2.28**	**1.88**	**1.21**
300	3.87	3.03	2.64	2.41	2.25	2.13	1.97	1.79	1.55	1.10
	6.72	**4.68**	**3.85**	**3.38**	**3.08**	**2.86**	**2.57**	**2.24**	**1.85**	**1.14**
400	3.86	3.02	2.63	2.40	2.24	2.12	1.96	1.78	1.54	1.07
	6.70	**4.66**	**3.83**	**3.37**	**3.06**	**2.85**	**2.56**	**2.23**	**1.84**	**1.11**
500	3.86	3.01	2.62	2.39	2.23	2.11	1.96	1.77	1.54	1.06
	6.69	**4.65**	**3.82**	**3.36**	**3.05**	**2.84**	**2.55**	**2.22**	**1.83**	**1.08**
1000	3.85	3.00	2.61	2.38	2.22	2.10	1.95	2.76	1.53	1.03
	6.66	**4.63**	**3.80**	**3.34**	**3.04**	**2.82**	**2.53**	**2.20**	**1.81**	**1.04**
	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	
	6.64	**4.60**	**3.78**	**3.32**	**3.02**	**2.80**	**2.51**	**2.18**	**1.79**	

Degree of freedom for smaller mean square

Table E: X^2 Table. P gives the probability of exceeding the tabulated value of x^2 for the specified number of degrees of freedom (df). The values of x^2 are printed in the body of the table.

df.	*0.95*	*0.90*	*0.80*	*0.70*	*0.50*	*0.30*	*0.20*	*0.10*	*0.05*	*0.02*	*0.01*
1	0.00393	0.0158	0.0642	0.148	0.455	1.074	1.642	2.706	3.841	5.412	6.635
2	0.103	0.211	0.446	0.713	1.386	2.408	3.219	4.605	5.991	7.824	9.210
3	0.352	0.584	1.005	1.424	2.366	3.665	4.642	6.251	7.815	9.837	11.345
4	0.711	1.064	1.649	2.195	3.357	4.878	5.989	7.779	9.488	11.668	13.277
5	1.145	1.610	2.343	3.000	4.351	0.064	7.289	9.236	11.070	13.388	15.086
6	1.635	2.204	3.070	3.828	5.348	7.231	8.558	10.645	12.592	15.083	16.812
7	2.167	2.833	3.822	4.671	6.346	8.383	9.803	12.017	14.067	16.622	18.475
8	2.733	3.490	4.594	5.527	7.344	9.524	11.030	13.362	15.507	18.168	20.090
9	3.325	4.168	5.380	6.393	8.343	10.656	12.242	14.684	16.919	19.679	21.666
10	3.940	4.865	6.179	7.267	9.342	11.781	13.442	15.987	18.307	21.161	23 .209
11	4.575	5.578	6.989	8.148	10.341	12.899	14.631	17.275	19.675	22.618	24.725
12	5.226	6.304	7.807	9.034	11.340	14.011	15.812	18.549	21.026	24.054	26.217
13	5.892	7.042	8.634	9.926	12.340	15.119	16.985	19.812	22.362	25.472	27.688
14	6.571	7.790	9.467	10.821	13.339	16.222	18.151	21.064	23.685	26.873	29.141
15	7.261	8.547	10.307	11.721	14.339	17.322	19.311	22.307	24.996	28.259	30.578

16	7.962	9.312	11.152	12.624	15.338	18.418	20.465	23.542	26.296	26.633	32.000
17	8.672	10.085	12.002	13.531	16.338	19.511	21.615	24.760	27.587	30.995	33.409
18	9.390	10.865	12.857	14.440	17.338	20.601	22.760	25.869	28.869	32.346	34.805
19	10.117	11.651	13.716	15.352	18.338	21.689	23.900	27.204	30.144	33.687	36.191
20	10.851	12.443	14.578	16.266	19.337	22.775	25.038	28.412	31.410	35.020	37.566
21	11.591	13.240	15.445	17.182	20.337	23.858	26.171	29.615	32.671	36.343	38.932
22	12.338	14.041	16.314	18.101	21.337	24.939	27.301	30.813	33.924	37.659	40.289
23	13.091	14.848	17.187	σ19.021	22.337	26.018	28.429	32.007	35.172	38.968	41.638
24	13.848	15.659	18.062	19.943	23.337	27.096	29.553	33.196	36.415	40.270	42.980
25	14.611	16.473	18.940	20.867	24.337	28.172	30.675	34.382	37.652	41.566	44.314
26	15.379	17.292	19.820	21.792	25.336	29.246	31.795	35.563	28.885	42.856	45.642
27	16.151	18.114	20.703	22.719	26.336	30.319	32.912	36.741	40.113	44.140	46.963
28	19.928	18.939	21.588	23.647	27.366	31.391	34.027	37.916	41.337	45.9	48.278
29	17.708	19.768	22.475	24.577	28.336	32.461	35.139	39.087	42.557	46.69	49.588
30	18.493	20.599	23.364	25.508	29.386	33.530	36.250	40.256	43.773	47.962	50.892

Table F: Conversion of a Pearson r into a corresponding Fisher's z coefficient.*

r	*z*	*r*	*z*	*r*	*z*	*r*	*z*	*r*	*z*	*r*	*z*
.25	.26	.40	.42	.55	.62	.70	.87	.85	1.26	.950	1.83
.26	.27	.41	.44	.56	.63	.71	.89	.86	1.29	.955	1.89
.27	.28	.42	.45	.57	.65	.72	.91	.87	1.33	.960	1.95
.28	.29	.43	.46	.59	.68	.74	.95	.89	1.42	.970	2.09
.30	.31	.45	.48	.60	.69	.75	.97	.90	1.47	.975	2.18
.31	.32	.46	.50	.61	.71	.76	1.00	.905	1.50	.980	2.30
.32	.33	.47	.51	.62	.73	.77	1.02	.910	1.53	.985	2.44
.33	.34	.48	.52	.63	.74	.78	1.05	.915	1.56	.990	2.65
.34	.35	.49	.54	.64	.76	.79	1.07	.920	1.59	.995	2.99
.35	.37	.50	.55	.65	.78	.80	1.10	.925	1.62		
.36	.38	.51	.56	.66	.79	.81	1.13	.930	1.66		
.37	.39	.52	.58	.67	.81	.82	1.16	.935	1.70		
.38	.40	.53	.59	.68	.83	.83	1.19	.940	1.74		
.39	.41	.54	.60	.69	.85	.84	1.22	.945	1.78		

*r's under 25 may be taken as equivalent to z' s.

Table G: Table of Critical Values of K_D in the Kolmogorov-Smirnov Two-Sample Test. (Small Samples)

	One-tailed test		*Two-tailed test*	
N	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
3	3	-	-	-
4	4	-	4	-
5	4	5	5	5
6	5	6	5	6
7	5	6	6	6
8	5	6	6	7
9	6	7	6	7
10	6	7	7	8
11	6	8	7	8
12	6	8	7	8
13	7	8	7	9
14	7	8	8	9
15	7	9	8	9
16	7	9	8	10
17	8	9	8	10
18	8	10	9	10

	One- tailed test		*Two-tailed test*	
N	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
19	8	10	9	10
20	8	10	9	11
21	8	10	9	11
22	9	11	9	11
23	9	11	10	11
24	9	11	10	12
25	9	11	10	12
26	9	11	10	12
27	9	12	10	12
28	10	12	11	13
29	10	12	11	13
30	10	12	11	13
35	10	12	11	13
40	11	14	13	

Table H: Table of Critical Values of *D* in the Kolmogorov-Smirnov Two-Sample Test.
Large samples: two-tailed test

Level of significance	*Value of D so large as to call for reflection of null hypothesis at the indicated level of significance where D - maximum* $S_{n_1}(X) - S_{n_2}(X)$
10	$1.22\sqrt{\frac{n_1+n_2}{n_1 n_1}}$
.05	$1.36\sqrt{\frac{n_1+n_2}{n_1 n_2}}$
.025	$1.48\sqrt{\frac{n_1+n_2}{n_1 n_2}}$
.01	$1.63\sqrt{\frac{n_1+n_2}{n_1 n_2}}$
.005	$1.73\sqrt{\frac{n_1+n_2}{n_1 n_2}}$
.001	$1.95\sqrt{\frac{n_1+n_2}{n_1 n_2}}$

Table I : Table of Probabilities Associated with Values as Small as Observed Values of x in the Binomial Test.

Given in the body of this table are one-tailed probabilities under null hypothesis for the binomial test when $P = Q = 1/2$.

x / N	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
5	.031	.188	.500	.812	.969											
6	.016	.109	.344	.656	.891	.984										
7	.008	.062	.227	.500	.773	.938	.992									
8	.004	.035	.145	.363	.637	.855	.965	.996								
9	.002	.020	.090	.254	.500	.746	.910	.980	.998							
10	.001	.011	.055	.172	.377	.623	.828	.945	.989	.999						
11		.006	.033	.113	.274	.500	.726	.887	.967	.994	†					
12		.003	.019	.073	.194	.387	.613	.806	.927	.981	.997	†				
13		.002	.011	.046	.133	.291	.500	.709	.867	.954	.989	.998	†			
14		.001	.006	.029	.090	.212	.395	.605	.788	.910	.971	.994	.999			
15			.004	.018	.059	.151	.304	.500	.696	.849	.941	.982	.996	†		
16			.002	.011	.038	.105	.227	.402	.598	.773	.895	.962	.989	.998	†	
17			.001	.006	.025	.072	.166	.315	.500	.685	.834	.928	.975	.994	.999	†
18			.001	.004	.015	.048	.119	.240	.407	.593	.760	.881	.952	.985	.996	.999
19				.002	.010	.032	.084	.180	.324	.500	.676	.820	.916	.968	.990	.988
20				.001	.006	.021	.058	.132	.252	.412	.588	.748	.868	.942	.979	.994
21				.001	.004	.013	.039	.095	.192	.332	.500	.668	.808	.905	.961	.987
22					.002	.008	.026	.067	.143	.262	.416	.584	.738	.857	.933	.974
23					.001	.005	.017	.047	.105	.202	.339	.500	.661	.798	.895	.953
24					.001	.003	.011	.032	.076	.154	.271	.419	.581	.729	.846	.924
25						.002	.007	.022	.054	.115	.212	.345	.500	.655	.788	.855

† 1.0 or approximately 1.0.

Table J: Table of Critical Values of T in the Wilcoxon Matched-Pairs Signed-Ranks Test.

N	*Level of Significance for one-tailed test*		
	.025	*.01*	*.005*
	Level of Significance for two-tailed test		
	.05	*.02*	*.01*
6	0	–	–
7	2	0	–
8	4	2	0
9	6	3	2
10	8	5	3
11	11	7	5
12	14	10	7
13	17	13	10
14	21	16	13
15	25	20	16
16	30	24	20
17	35	28	23
18	40	33	28
19	46	38	32
20	52	43	38
21	59	49	43
22	66	56	49
23	73	62	55
24	81	69	61
25	89	77	68

Table K: Table of Critical Values of Pearson Product Correlation at the .05 and .01 Levels of Significance.

Degrees of freedom (N-2)	*.05*	*.01*	*Degrees of Freedom (N-2)*	*.05*	*.01*
1	.997	1.000	24	.388	.496
2	.950	.990	25	.381	.487
3	.878	.959	26	.374	.478
4	.811	.917	27	.367	.470
5	.754	.874	28	.361	.463
6	.707	.834	29	.355	.456
7	.666	.798	30	.349	.449
8	.632	.765	35	.325	.418
9	.602	.735	40	.304	.393
10	.576	.708	45	.288	.372
11	.553	.684	50	.273	.354
12	.532	.661	60	.250	.325
13	.514	.641	70	.233	.302
14	.497	.623	80	.217	.283
15	.482	.606	90	.205	.267
16	.468	.590	100	.195	.254
17	.456	.575	125	.174	.228
18	.444	.561	150	.159	.208
19	.433	.549	200	.138	.181
20	.423	.537	300	.113	.148
21	.413	.526	400	.098	.128
22	.404	.515	500	.088	.115
23	.396	.505	1000	.062	.081

Table L: Table of Critical Values of Spearman Rank Order Correlation at .05 and .01 Levels of Significance (One-tailed test).

N	*.05*	*.01*	*N*	*.05*	*.01*
5	.900	1.000	16	.425	.601
6	.829	.943	18	.399	.564
7	.714	.893	20	.377	.534
8	.643	.833	22	.359	.508
9	.600	.783	24	.343	.485
10	.564	.764	26	.329	.465
12	.506	.712	28	.317	.448
14	.456	.645	30	.306	.432

Table M: Values of r_t taken as the Cosine of an Angle.

Angle	*Cosine*	*Angle*	*Cosine*	*Angle*	*Cosine*
0°	1.000	41°	.755	73°	.292
		42	.743	74	.276
5	.996	43	.731	75	.259
		44	.719	76	.242
10	.985	45	.707	77	.225
		46	.695	78	.208
		47	.682	79	.191
15	.966	48	.669	80	.174
16	.961	49	.656		
17	.956	50	.643	81	.156
18	.951			82	.139
19	.946	51	.629	83	.122
20	.940	52	.616	84	.105
		53	.602	85	.087
21	.934	54	.588		
22	.927	55	.574		
23	.921	56	.559	90	.000
24	.914	57	.545		
56	.906	58	.530		
26	.899	59	.515		
27	.891	60	.500		
28	.883				
29	.875	61	.485		
30	.866	62	.469		
		63	.454		
31	.857	64	.438		
32	.848	65	.423		
33	.839	66	.407		
34	.829	67	.391		
35	.819	68	.375		
36	.809	69	.358		
37	.799	70	.342		
38	.788				
39	.777	71	.326		
40	.766	72	.309		

Table N: A few .05 Level of Significance Values from the F_{max} Distribution Table

		2	3	4
Degrees of Freedom for Largest Group (N–1)	9	4.03	5.34	6.31
	10	3.72	4.85	5.67
	12	3.28	4.16	4.79
	15	2.86	3.54	4.01