Unit 1
Introduction to Statistics

1. Introduction

Numbers play an essential role in statistics. They provide the raw material of statistics. These materials must be processed to be useful. The study of statistics involves methods of refining numerical information into useful forms. This unit introduces the concept of statistics, functions and use of statistics in education and research, and limitations of statistics.

2. Specific objectives and contents

Specific Objectives	Contents
• Define statistics • Identify the use of statistics in education and research • Point out the limitation of statistics	**Unit I: Introduction to statistics (4)** 1.1 Concept of statistics 1.2 Function of statistics 1.3 Use of statistics in education and research 1.4 Limitations of statistics

3. Content Presentation

3.1 Concept of Statistics

The word 'statistics' seems to have been derived from the Latin word 'Status' or Italian word 'Statista' or German word 'Statistik' all of which mean the political state. In those days statistics was used only in collecting the information relating to the population, military strength, incomes, etc. The government needed such information for framing the military and the economic policy. With the passage of time, statistics has been applied very widely. It is used not only to the state administration but it is used in economics, science, business, research, etc. It plays an import role in various aspects of human activities.

Different writers have defined 'Statistics' differently. Some of them have used in singular sense and some of them in plural sense. In the plural sense, it means the quantitative information or numerical facts collected systematically. For example, population, national incomes, unemployment, exports, imports, etc. In the singular sense, it means the various methods and techniques adopted for the collection, presentation, analysis and the interpretation of the figures.

Definition in Singular Sense

According to A. L. Bowley, 'Statistics may be as the science of counting'. Again he defined 'Statistics as the science of averages'. These are the narrow definitions because they cover one aspect of statistics. The first definition is limited to counting only neglecting other important aspects of statistics. The second definition includes only the averages which is related to the measures of central tendency, which is a small part of statistics.

Prof. Boddington has defined 'Statistics as the science of estimates and probabilities'. This definition is also not complete as it covers only estimates and probabilities, which are only the part of statistical methods.

Croxton and Cowden have given a very comprehensive definition of statistics. According to them, 'Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data'. This definition clearly indicates the following five features of statistics.

1. Collection of data

The first step for a statistical investigation is the collection of data. Utmost care must be exercised in collecting data because they form the foundation of statistical analysis. The findings and interpretation depends on the data obtained. If the collected data is faulty, then the results obtained will not be reliable.

2. Organization of data

The data may be obtained from different sources. After the data have been collected, they are organized in a systematic way. A large mass of figures that are collected from a survey frequently needs organization. The data are organized by editing, classifying, and tabulating them.

3. Presentation.

After collecting and organizing the data, they are arranged in systematically. It can be presented in table form, diagrammatical form and graphical form. Tabulation of data is one way of presenting data in organized form so that the data are condensed and readily comprehensible.

4. Analysis of data

After the data have been organized and presented in a concise form, the next step is to analyze them. There are various well defined statistical tools which are employed to analyze the data. Some of the commonly employed tools and techniques are the measures of central tendency, dispersion, correlation, regression techniques, etc.

5. Interpretation of data

The last stage in statistical enquiry is interpretation of data. In this step the investigator has to draw conclusions and logical inferences from the organized and analysis of data. If the analyzed data are not properly interpreted, the whole object of the investigation may be defeated and fallacious conclusions be drawn. Hence, It is indeed a challenging task, which requires a great deal of skill and experience.

Definition in Plural Sense

Numerical data may be found almost everywhere in business, education, economics and many other areas. In the plural sense, i.e. in the sense of numerical data, there are many definitions.

Webster defined statistics as 'the classified fats representing the conditions of the people in a State….especially those facts which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement. This definition is too narrow as it confines the scope of statistics to only such facts and figures which are related to the conditions of the people in a State.

The most comprehensive definition of statistics has been given by Prof. Horace Secrist. He defines 'By statistics, we mean aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other.' This definition includes all aspects of Statistics. This definition has following features:

1. Statistics are aggregate of facts

Single or isolated figures cannot constitute statistics. To constitute statistics, there should be an aggregate of facts.

2. They are affected by multiplicity of causes

Statistics are usually affected not just by a single factory but they are affected by a number of factors. For example, the production statistics of a certain crop depends upon a number of factors such as quality of seed, weather condition, irrigation facility, fertility of soil, method of cultivation and so on.

3. They are numerically expressed

Statistical methods are applicable to only those information or facts which are expressed numerically. For example, Nepal is one of the least developed countries in the world, is not statistics. But the statement that 'the production of wheat in creased by 6.5% in 2012 against 5.6% in 2011' is a statistical statement.

4. **They are enumerated or estimated according to reasonable standards of survey**

Statistical data can be collected either by enumeration or by estimation. If the data is collected through enumeration, the data will be exact and accurate. Whenever, enumeration is not possible, we collect the data by estimation in which case data may not be accurate. 100% accuracy in the statistical work is rare. The degree of accuracy desired depends upon the nature and object of the enquiry. However, certain standards of accuracy must be maintained for drawing valid conclusion.

5. **They are collected in a systematic manner**

The data should be collected in a very systematic manner with suitable plans. Haphazard collection of data could lead to fallacious conclusions.

6. **They are collected for a predetermined purpose**

The purpose of investigation and data should be clearly specified before collecting data. This will facilitate the collection of proper and relevant data.

7. **They are placed in relation to each other**

The numerical data collected constitutes statistics if they are comparable. Hence, data should be related to leach other so that they are comparable to each other. Figures which cannot be compared to each other are not statistics. For example, the data relating to the production of wheat for different years constitutes statistics because they can be compared. But the weight of a student and mark obtained by him in an examination do not constitute statistics because the data cannot be compared.

Activity:

1. Define statistics in singular sense and explain its features.
2. Define statistics in plural sense and explain its features.

3.2 Functions of Statistics

The important functions of statistics are as follows:

1. **Statistics express facts in numbers**

The first and important function of statistics is to express facts in numbers so that they are suitable for analysis and interpretation.

2. **It presents facts in a definite form**

Statistics expresses facts in numbers. Thus, the statements which can not be precisely understood are brought into its definite forms by expressing them in numbers by statistics. Hence, statistics presents facts in a precise and definite form. It is possible to understand properly what is stated.

3. It simplifies complexities

A huge mass of data is difficult to understand and remember. The complex mass of figures can be made simple and understandable with the help of statistical methods. Various statistical techniques condense the huge mass of data and make them easily comprehensive and ready for analysis.

4. It facilitates comparison

Unless the figures are compared with other figures with the same kind, they are usually meaningless. Statistical methods such as classification, averages, ratios, percentages, correlations, etc. can be used for comparison between two or more groups of data. For example: the per capita income of Nepal is $560 is not so clear, whether it is low, moderate or high unless it is compared with the income of other countries.

5. It helps in formulating and testing hypothesis

Statistical methods are extremely helpful in formulating and testing hypothesis and to develop new theories. For example, hypothesis like whether a particular coin is fir or not can be tested by appropriate statistical tools.

6. It helps in forecasting

While preparing suitable policies and plans, it is necessary to have the knowledge of future tendency. Statistical methods provide helpful means of forecasting future events. Statistics enables us to predict how much of an event will happen under conditions we know, e.g. we can predict the achievement of a student on the basis of his intelligence.

3.3 Use of Statistics in Education and Research

Statistics is regarded as an indispensable instrument in the fields of education, and research, especially where any sort of measurement or evaluation is involved. There are many uses of statistics in education and research. Some of the common uses are as follow:

1. Measurement and evaluation

In the field of education and research various tests and measures for carrying out the task are to be constructed and standardized. These may include intelligence tests, achievement tests, interest inventories and similar other measures of education or psychological interest. The knowledge of statistical methods helps not only in carrying out the construction and standardization of these tests and measures but also in using them properly by presenting, comparing, analyzing and interpreting the results of these tests and measures.

2. Day-to-day tasks

Statistics and its methods help the people belonging to the fields of education, psychology and research in carrying out their day-to-day tasks and activities. For example, a teacher may utilize statistics for:

a. knowing individual differences of his students,

b. comparing the suitability of one method or technique with another,

c. comparing the function and working of one institution with another,

d. making prediction regarding the future progress of the students, and

e. making, selection, classification, and promotion of the students.

3. Research

Research and innovations are essential in any field of knowledge for enrichment, progress and development. Research and statistics are two areas that must not be separated. As for a study to be empirical there must be the use of quantitative measurements, that can support and prove the conclusion of a given research. Guilford has summarized the advantages of statistical methods in research. According to him, they are

a. permit the most exact kind of description,

b. force us to be definite and exact in our procedures and in our thinking,

c. enable us to summarize our results in a meaningful and convenient form,

d. enable us to predict and

e. enable us to analyze some of the causal factors underlying complex and otherwise bewildering events.

4. Understanding and Using the Products of Research

Statistics and its methods help the practitioners-students, teachers, educationists, guidance personnel, etc. to keep abreast of the latest developments and research. The knowledge of these methods helps them in reading and understanding the reports of applied and theoretical research. The proper use of the results of these researches also becomes possible with proper knowledge of statistics.

Statistics carry wide application for the persons working in the fields of education, psychology and research not only in discharging their roles and duties effectively but also enriching and contributing significantly to their respective fields by bringing useful research and innovations for the welfare of the society and humanity at large.

3.4 Limitations of Statistics

The limitations of statistics are as follows:

1. Statistics does not deal with individual measurements

Statistics deal with aggregates of facts, not with individual figure. A single item or the isolated figure cannot be regarded as statistics. For example, the marks obtained by a student is 60 does not constitute statistics but the average mark of a group of students in English is 60 forms statistics.

2. Statistics deals with quantitative phenomena

Statistics are numerical statements of facts. Thus, the study of qualitative phenomena lies outside the scope of statistics. Qualitative phenomena such as honesty, beauty, intelligence, poverty etc. cannot be directly studied. However, it is possible to analyze such problems statistically by expressing them in numbers.

3. Statistical results are true only on an average

The conclusions obtained statistically are not universally true—they are true only under certain conditions. Bowley has said statistics is the science of average.

4. Statistics is only one of the methods of studying problem

There are various methods of studying the various problems we are facing. Statistics is one of those methods of studying problems. Statistical methods should be supplemented by other techniques to arrive at a conclusions.

5. Statistics can be misused

The most important limitation of statistics is that it must be handled by experts. If statistics is applied by those persons who are not expert in this field, then the conclusions drawn may be misleading. Similarly, statistical conclusions based on incomplete information may not be true as well. It cannot be used to full advantage in the absence of proper understanding of the subject to which it is applied.

4. Summary

Different writers have defined 'Statistics' differently. Croxton and Cowden have given a very comprehensive definition of statistics. According to them, 'Statistics may be defined as the collection, presentation, analysis and interpretation of numerical data'. This definition clearly indicates the following five features of statistics.

1. Collection of data
2. Organization of data
3. Presentation.
4. Analysis of data
5. Interpretation of data

In the plural sense, i.e. in the sense of numerical data, there are many definitions. Prof. Horace Secrist defines 'By statistics, we mean aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a pre-determined purpose and placed in relation to each other.' This definition includes all aspects of Statistics. This definition has following features:

1. Statistics are aggregate of facts
2. They are affected by multiplicity of causes
3. They are numerically expressed
4. They are enumerated or estimated according to reasonable standards of survey
5. They are collected in a systematic manner
6. They are collected for a predetermined purpose
7. They are placed in relation to each other

The important functions of statistics are as follows:

1. Statistics express facts in numbers
2. It presents facts in a definite form
3. It simplifies complexities
4. It facilitates comparison
5. It helps in formulating and testing hypothesis
6. It helps in forecasting

Statistics is regarded as an indispensable instrument in the fields of education, and research, especially where any sort of measurement or evaluation is involved. There are many uses of statistics in education and research. Some of the common uses are as follow:

1. Measurement and evaluation
2. Day-to-day tasks
3. Research
4. Understanding and Using the Products of Research

The limitations of statistics are as follows:

1. Statistics does not deal with individual measurements
2. Statistics deals with quantitative phenomena
3. Statistical results are true only on an average
4. Statistics is only one of the methods of studying problem
5. Statistics can be misused

5. Exercise

Objective Questions

1. Which one of the following is the feature of statistics in singular form ?
   a. Collection of data
   b. Simplifies complexities
   c. Affected by multiplicity of causes
   d. Collected for predetermined purpose
2. What is the limitation of statistics?
   a. Deals with individual measurement
   b. Deals with group measurement
   c. Results are 100 percent true
   d. Results are not true

Short Answer Questions

1. Define statistics in singular sense and explain its features.
2. Define statistics in plural sense and explain its features.
3. Describe the functions of statistics.
4. Describe the uses of statistics in education and research.

Long Answer Question

1. Define statistics and describe the uses of statistics in education and research.

Unit 2

Measures of Central Tendency, Dispersion and Relative Position

1. Introduction

This unit consists of central tendency, dispersion and measures of relative position. There are many measures of central tendency. We will consider only the three most commonly used in education: arithmetic mean, median and mode. Similarly, we will consider range, inter-quartile range, standard deviation and variance towards dispersion. Regarding measures of relative position, we will consider percentile, percentile rank and standard score. The purpose of these statistics is to summarize sets of numbers, so that interesting features may be seen and understand more easily.

2. Specific Objectives and Contents

Specific Objectives	Contents
- Explain the meaning and use of central tendency and dispersion. - Compute mean, median and mode. - Describe the use and limitations of different measures of central tendency. - Compute region, inter-quartile range, standard deviation and variance. - State the use and limitations of different measures of dispersion. - Compute and interpret percentile rank, percentile, stanine and standard score.	**Unit II: Measures of Central Tendency, Dispersion and Relative Position (10)** **2.1 Central tendency** 2.1.1 Concept 2.1.2 Computation and arithmetic mean, median and mode with their use and limitations. **2.2 Dispersion** 2.2.1 Concept 2.2.2 Computation of range, inter-quartile range, standard deviation and variance with their use and limitations. **2.3 Measures of relative position** 2.3.1 Computation of percentile rank, percentile, stanine, standard score and their use

3. Content Presentation

3.1 Central Tendency

One of the most important objectives of statistical analysis is to get one single value that describes the characteristics of the entire mass of data. Such a value is called average or central value which can represent the entire mass of data. Averages are the typical values around which most of the data tend to cluster. Such a value lies somewhere at the center of distribution of data. For this reason, it is also known as measure of central tendency.

Central tendency is an average which represents all the scores made by the group and as such gives concise description of the performance of the group as a whole. It is the value around which most of the data tend to concentrate. The objectives of central tendency are as follows:

(1) to get single value that describes the characteristics of the entire group.

(2) to facilitate comparison

(3) to help in decision making

The most commonly used measures of central tendency are as follows:

(1) Arithmetic mean

(2) Median

(3) Mode

3.1.1 Arithmetic Mean

The arithmetic mean, commonly called the mean or average, is the most often used measure of central tendency. It is defined as the sum of measures divided by the number of measures. For brevity, the arithmetic mean is usually called the mean and denoted by $\overline{X}$.

Computation of Mean (Ungrouped data)

The arithmetic mean or simply mean is the sum of separate scores or measures divided by the number of scores. If x_1, x_2, x_3, ..., x_n represent the values of the respective N items, then the arithmetic mean denoted by $\overline{X}$, of the given values is defined as,

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \ldots x_n}{N}$$

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

Example 2.1

Compute mean from the following data:

Student	Ram	Hari	Sita	Raj	Raju	Pun	Ila	Babu	Rita	Ritu
Marks in Science	30	35	40	45	45	50	55	60	65	70

Solution:

Here, $\sum X = 30 + 35 + 40 + 45 + 45 + 50 + 55 + 60 + 65 + 70$

$= 495$

N = No. of students = 10

We have, $\overline{X} = \frac{\Sigma X}{N} = \frac{495}{10} = 49.5$

Therefore mean = 49.5

Note: For discrete frequency distributions, the following steps are adopted.

1. Multiply each value of the variable by its corresponding frequency and obtain the total of those values $\sum fx$.

2. Divide the total obtained in step 1 by the total number of frequencies, i.e. $\sum f = N$, $\overline{X} = \frac{\Sigma fx}{N}$.

Example 2.2

100 students of a class obtained the following marks. Compute mean for this data:

Marks	5	15	25	35	45
No. of students	12	15	28	25	20

Solution:

Marks (X)	No. of Students (f)	fx
5	12	60
15	15	225
25	28	700
35	25	875
45	20	900
	N = ∑f = 100	∑fX = 2,760

$$\overline{X} = \frac{\Sigma fX}{N} = \frac{2,760}{100} = 27.60$$

Therefore mean = 27.60.

Mean in Grouped Data (Continuous Series)

In case of a continuous series, then are two methods of computing the mean:

(1) Direct Method

(2) Short-cut Method

Direct Method

Under the direct method the following steps will be used to calculate the value of mean:

(1) Find the mid-point (x) of each class interval by taking the average of upper and lower of all class intervals (column 2 in the following example).

(2) Multiply each mid-point (x) by its corresponding frequency to get the fx (column 4 in the following example).

(3) Add the values of fx to get ∑fx.

(4) Apply the following formula to get $\overline{X}$.

$$\overline{X} = \frac{\Sigma fx}{N}$$

Example 2.3

Calculate the mean for the following frequency distribution.

Class interval	23 – 25	21 – 23	19 – 21	17 – 19	15 – 17	13 – 15	11 – 13
Frequency	3	4	5	6	5	4	3

Solution:

Calculation of Arithmetic Mean

Class Interval	Mid-point (x)	f	fx
23 - 25	24	3	72
20 – 23	22	4	88
19 – 21	20	5	100
17 – 19	18	6	108
15 – 17	16	5	80
13 – 15	14	4	56
11 – 13	12	3	36
		N = 30	∑fx = 540

Here, ∑fx = 540 and N = 30

$$\therefore \quad \overline{X} = \frac{\Sigma fx}{N} = \frac{540}{30} = 18$$

So, the mean = 18

Short-cut Method

In this method arithmetic mean is calculated by applying the following steps:

(1) Find out the mid-point of each class interval by taking the average of the lower and upper scores of the class (column 2 of the following solution).

(2) Near the centre of the distribution, choose an estimated mean class. (We chose 17-19 class. The mid-point of this is 18 which will be the value of assumed mean A.)

(3) d can be found for each class by subtracting A from x and then dividing by i (as in column 4 in the following example).

(4) The value of fd can be found by multiplying f and d values (as in column 5 in the following example).

(5) Find ∑fd by first adding fd values and negative fd values, thereby taking the numerical differences of these sub-totals.

(6) Apply the following formula to get $\overline{X}$.

$$\overline{X} = A + \frac{\Sigma fd}{N} \times i.$$

Where, A = assumed mean, f is the respective frequency, i is the class interval, d is the deviation of x from assumed mean A and divide by i, the class interval.

Solution:

Example 2.3 by short-cut method.

Class Interval	Mid-point (x)	f	d	fd
23 - 25	24	3	+3	9
21 – 23	22	4	+2	8
19 – 21	20	5	+1	5
17 – 19	18	6	0	0
15 – 17	16	5	–1	–5
13 – 15	14	4	–2	–8
11 – 13	12	3	–3	–9
			N = 30	∑fd = 0

Here, A = 18, ∑fd = 0, N = 30, i = 2

$$\therefore \quad \overline{X} = A + \frac{\Sigma fd}{N} \times i = 18 + \frac{0}{30} \times 2 = 18 + 0$$

So, the mean = 18

Mean from Combined Groups

If there are K groups whose means ($\overline{X}$) and number of cases in each group (N_1) are known. We can compute their combined mean by using the following formula:

$$\overline{X}_{comb} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2 + ...N_x\overline{X}_x}{N_1 + N_2.....N_x}$$

When only two groups are combined, arithmetic mean of the two groups can be obtained by using the following formula:

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

Example 2.4

Twenty five boys and 15 girls in a class appeared in an examination. The average grade for boys and girls is 5 and 6 respectively. Find the average grade of all the 40 students in a class.

Solution:

$N_1 = 25$, $\overline{X}_1 = 5$

$N_2 = 15$, $\overline{X}_2 = 6$

Using the formula, the combined average grade of all the 40 students will be:

$$\overline{X}_{12} = \frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}$$

$$= \frac{25\times5+15\times6}{25+15} = \frac{125+90}{40} = \frac{215}{40} = 5.375$$

$\therefore$ $\overline{X}_{12}$ = 5.375 grade.

Activities

1. What is the mean for the following set of scores:

   17, 1, 12, 10, 8, 8, 5

2. Find the mean of the following data:

Scores (x)	30	35	40	45	50	55
Frequency (f)	6	4	13	25	9	7

3. Find the mean from the following series:

Marks:	10-19	20-29	30-39	40-49	50-59	60-69	70-79
Students	4	5	1	7	3	4	2

Uses of Mean

Arithmetic mean is most widely in practice. Some of the uses of mean are as follows:

1. It is the simplest average to understand and easiest to compute. Neither the arraying of data as required for calculating median nor grouping of data as required for calculating mode is needed while calculating mean.
2. Mean is the centre of gravity in the distribution and each score contributes to the determination of it when the spread of the scores are symmetrically around a central point.
3. It is affected by the value of every item in the series.
4. Mean is more stable than the median and mode. When the measures of central tendency having greatest stability is wanted, mean is used.
5. Mean is used to calculate other statistics like standard deviation, coefficient of correlation, etc.
6. Mean is defined by a rigid mathematical formula so that there is no question of misunderstanding about its meaning and nature.

Limitations of Mean

1. Since the value of mean depends upon each and every item of the series, extreme items i.e. very small and very large items, unduly affect the value of the mean.
2. In case of open ended class intervals, it cannot be calculated without assuming the size of the open end classes.
3. The arithmetic mean is not always a good measure of central tendency. It indicates there most of the values lie, only when the distribution of the score is reasonably normal (bell shaped). In case of a U-shaped distribution, the mean is not likely to serve a useful purpose.
4. Sometimes mean is a value which is not present in the series and gives absurd values. For example, there are 41, 42 and 44 students in Class VIII, IX and X of a school. The average students per class is 42.3. It is never possible.

3.1.2 Median

The value lying at the middle of a series when the items in the series are arranged in an increasing or decreasing order of magnitudes is known as median. Hence, the median is the mid-point in a distribution. It divides the distribution into two haves with respect to frequencies. For example, the median of the series 1, 4, 6, 7, 9 is 6 as it lies exactly at the middle position of the series.

Computation of Median for Individual Series (Ungrouped Data)

The calculation of the median for an individual series is done in the following manner:

1. Arrange all the items in an ascending or descending order.
2. Then the formula given below is applied to obtain the median: Median = Value of the $\left(\frac{N+1}{2}\right)^{th}$ item, where N = No. of observations of scores.

Example 2.5

Computer the median score from the following scores obtained by score students:

7, 10, 6, 8, 9, 10, 11.

Solution:

First of all we arrange the given data in an ascending order as below:

Serial No.	Scores
1	6
2	7
3	8
4	**9**
5	10
6	10
7	11

Median = $\frac{N+1}{2} = \frac{7+1}{2}$ = 4th item

Median = Value of the 4th item = 9.

Example 2.6

Find the median height of 8 students with the following heights (In inches) :

53, 58, 50, 54, 52, 56, 55, 57.

Serial No.	Heights of Ascending Order
1	50
2	52
3	53
4	**54**
5	**55**
6	56
7	57
8	58

∴ Median = Value of the $\frac{N+1}{2} = \frac{8+1}{2}$ = 4.5th item

Now value of the 4th item and 5th item will be taken

∴ Median = $\frac{\text{Value of the } 4^{th} \text{ item} + \text{Value of the } 5^{th} \text{ item}}{2}$

$= \frac{54+55}{2} = \frac{109}{2}$ = 54.5 inches

Note: In cases like this where we do not get the value in a whole number and get the value in decimal number then we take the items just below and just greater than this decimal number.

Discrete Series

The following steps are adopted to calculate the median in use of discrete series.

(1) Arrange the data in an ascending (or descending) order of magnitudes.

(2) Find the cumulative frequency.

(3) Use the formulas, Median = $\left(\frac{N+1}{2}\right)^{th}$ item, where N = total frequency.

(4) Now see the cumulative frequency column and note the value corresponding to the cumulative frequency either equal to or greater than $\left(\frac{N+1}{2}\right)$. This gives the value of the median.

Example 2.7

Find the median marks from the following data:

Marks	20	30	50	60	70	80	90
No. of students	2	3	11	10	6	6	2

Solution:

Calculation of Median

Marks in Ascending Order	**Frequency**	**Cumulative Frequency (c.f.)**
20	2	2
30	3	2 + 3 = 5
50	11	5 + 11 = 16
60	10	16 + 10 = **26**
70	6	26 + 3 = 32
80	6	32 + 6 = 38
90	2	38 + 2 = 40

Here, N = total frequencies = 40

$$\text{Median} = \text{Value of } \left(\frac{N+1}{2}\right)^{th} \text{ item} = \frac{40+1}{2}$$

$$= \text{Value of the } 20.5^{th} \text{ item}$$

Now, the c.f. value just greater than 20.5 is 26 and its corresponding value = 60.

Therefore, the required median marks = 60.

Continuous Series

The median in a continuous series is calculated following the steps given below:

(1) First find the cumulative frequencies.

(2) Applying the formula median = value of $\frac{N}{2}^{th}$ item, find the class interval in which the median lies. This class is called the median class. (The class interval corresponding to the c.f. value equal to or just greater than $\frac{N}{2}$ gives the median class.)

(3) The exact value of the median is calculated using the following formula:

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

Example 2.8

Compute median score for the following frequency distribution.

Scores	30 – 39	40 – 49	50 – 59	60 – 69	70 – 79
No. of students	5	8	12	9	6

Solution:

Scores	**No. of Students (f)**	**Cumulative Frequency (c.f.)**
30 – 39	5	5
40 – 49	8	5 + 8 = 13
50 – 59	**12**	13 + 12 = **25**
60 – 69	9	25 + 9 = 35
70 – 79	6	34 + 6 = 40
	N = 40	

Here, $\frac{N}{2} = \frac{40}{2} = 20$

The c.f. value just greater than 20 is 25 and its corresponding class interval is 50 – 59. Hence, 50 – 59 is the median class.

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

$$= 49.5 + \frac{20-13}{12} \times 10 = 49.5 + \frac{7}{12} \times 10$$

$$= 49.5 + 5.83 = 55.33$$

Hence, median score = 55.33.

Uses of Median

1. Median is used when the mid-point of the distribution is needed or the 50% point is wanted.
2. When extreme scores affect the mean at that time median is the best measure of central tendency. Extreme values do not affect the median as strongly as they do the mean.
3. It is the most appropriate average in dealing with qualitative data i.e. where ranks are given or there are other types of items that are not counted or measured but are scored.
4. Median is used when the distribution has unequal class interval or open ended classes.

Limitations of Median

1. For calculating medina, it is necessary to arrange data in ascending or descending order. Other averages do not need any arrangement.
2. Median is the positional average which value is not determined by each and every item.
3. It cannot be further treated algebraically like mean. For example, median cannot be used for determining the combined median of two or more groups.
4. In some cases, median cannot be computed exactly. When the number of items included in a series of data is even, the median is determined approximately as the mid-point of the two middle items.

3.1.3 Mode

Mode is that value which repeats maximum number of times. It is that value which occurs with the highest frequency in the set of observation. For example, the mode of the series 1, 4, 5, 7, 5, 9, 5 is 5 because this value occurs the maximum number of times. It is never affected by extreme scores but by extreme frequencies of the values. To determine mode different methods are used. Some of the important methods are as follows:

(1) Inspection method

(2) Grouping method

(3) Empirical relation method

(1) Inspection Method

In this method mode is determined just be observation. Incase of individual and discrete series the mode can be found out by inspection. Here, mode is determined by observing the most frequently occurring score or the class interval against which the maximum frequency stands is taken as the model class. But in case of continuous series, we use the following formula to calculate the mode.

Formula : Mode (M_0) = $L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i$

Where, L = lower limit of the modal class.

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

i = size (width) of the modal class.

This formula may also be written as:

Mode (M_0) = $L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times i$

Where, $\Delta_1 = f_1 - f_0$

$\Delta_2 = f_1 - f_2$.

If the value of f_0 or f_2 are greater than f_1, instead of applying the formula given above we should apply the following alternative formula:

$$\text{Mode} = L + \frac{f_2}{f_0 + f_2} \times i$$

Individual Series

Example 2.9

Find the mode from the following data:

Role No	1	2	3	4	5	6	7	8	9	10
Marks	10	25	40	33	25	40	40	35	33	45

Solution:

Calculation of Mode

Marks	No. of Repetition
10	1
25	2
33	2
35	1
40	**3**
45	1

The required mode is 40 as it has appeared the maximum number of times.

Discrete Series

In most of the cases mode is obtained by applying the method of inspection.

Example: 2.10

Calculation the mode from the following data:

Shoe size	5	6	7	8	9	10
No. of persons	19	40	66	48	15	9

Solution:

By inspection mode of the series is (shoe size) 7 as it corresponds to the highest frequency.

Continuous Series

Example 2.11

Find the mode from the following distribution:

Score	15-19	20-24	25-29	30-34	35-39	40-44	45-49
No. of students	1	4	6	9	5	3	2

Solution:

Calculation of Mode

Scores Class Interval	**Frequency (f)**
15 – 19	1
20 – 24	4
25 – 29	6 (f_0)
30 – 34	9 (modal class) f_1
35 – 39	5 (f_2)
40 – 44	3
45 – 49	2

Here, 30 – 34 is the modal class having maximum frequency $f_1 = 9$. But the lower limit (L) of this class interval is 29.5. The frequency of preceding and succeeding the modal class are 6 and 5 respectively and class interval is 5.

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times i = 29.5 + \frac{9-6}{2\times 9 - 6 - 5} \times 5$$

$$29.5 + \frac{3}{18-11} \times 5 = 29.5 + \frac{3}{7} \times 5 = 29.5 + \frac{15}{7} = 29.5 + 2.14 = 31.64$$

Mode = 31.64

(2) Grouping Method

In discrete and continuous series quite often mode can be determined just by inspection i.e. by looking to that value of the variable around which the items are most heavily concentrated. However, where the mode is determined just by inspection, an error of judgement is possible in those cases where the difference between the maximum frequency and the frequency preceding it or succeeding it is very small and the items are heavily concentrated on either side. Hence, in the following situations it is appropriate to use grouping method.

(a) If the difference between the highest frequency and second highest frequency is very small.

(b) If the highest frequency occurs either in the very beginning or at the very end of the distribution.

(c) If the given distribution is found to be as irregular one i.e. the frequencies of the variable increase and decrease in a haphazard way.

There are two steps to be employed for the method of grouping. First we have to form a grouping table and then an analysis table.

The following steps are to be carried out while implementing the method of grouping:

(a) First of all, we prepare a grouping table consisting of six columns.

(b) In column I, given frequencies are placed against the corresponding value and the highest frequency is highlighted by any means.

(c) In column II, we add frequencies in two by making several groups of two frequencies starting from the top and highlight the resulting highest frequency.

(d) In column III, leaving the frequency at the top, we add frequencies pairwise as in step (c) and highlight the maximum one.

(e) In column IV, frequencies are added in three's by making several groups of three frequencies starting with the top and again highlight the highest frequency in this column.

(f) In column V, we repeat the same process as in step (e) by leaving the first frequency at the top and highlight the resulting maximum frequency.

(g) In column VI, leaving the first two frequencies at the top we add three frequencies as in the previous step and highlight the resulting highest frequency in this column.

(h) In analysis table, we assign frequencies to the respective values with the help of the highlighted frequencies in various columns of the grouping table.

The process of implementing the grouping table will be more clear from the following example.

Example 2.12

Calculate the value of mode for the following data:

Marks	10	15	20	25	30	35	40	45
No. of students	8	12	36	35	28	18	5	4

Solution:

When we observe the distribution, we see that the highest frequency is 36 and difference between the highest frequency and second highest frequency is very small. Thus, we cannot immediately obtain the mode in this case. We apply grouping method to calculate mode value.

Grouping Table

	Frequency					
Marks	I	II	III	IV	V	VI
10	8	20		56		
15	12		46		83	
20	36	71				99
25	35		63	81		
30	28	46			51	
35	18		23			27
40	5	9				
45	4					

Analysis Table

Frequency Column	Probable Values							
	10	15	20	25	30	35	40	45
I			1					
II			1	1				
III				1	1			
IV				1	1	1		
V		1	1	1				
VI			1	1	1			
		1	4	5	3	1		

Here, as the value 25 has got the maximum frequency of 5, the required mode is 25 marks.

Example 2.13

Find the mode form the following data:

Class	Frequency	Class	Frequency
0 – 10	1	50 – 60	20
10 – 20	5	60 – 70	9
20 – 30	17	70 – 80	3
30 – 40	22	80 – 90	2
40 – 50	21		

Solution:

Here, the highest frequency is 22 but the frequency following it is 21, where there is high concentration of frequencies, so modal class cannot be identified by inspection. We shall form a grouping table and analyzed table to identify the modal class.

Grouping Table

Class Interval	1	2	3	4	5	6
0 – 10	1	8		23		
10 – 20	5		22		44	
20 – 30	17	39				60
30 – 40	22		45	63		
40 – 50	21	41				
50 – 60	20		29		50	32
60 – 70	9	12		14		
70 – 80	3		5			
80 – 90	2					

Analysis Table

Col. No	Class Interval								
	0 – 10	**10 – 20**	**20 – 30**	**30 – 40**	**40 – 50**	**50 – 60**	**60 – 70**	**70 – 80**	**80 – 90**
1				1					
2					1	1			
3				1	1				
4				1	1	1			
5					1	1	1		
6			1	1	1				
			1	4	5	3	1		

So, 40 – 50 becomes the modal class. Here, since $f_0 > f_1$ we have to apply the following formula:

$$\text{Mode } (M_0) = L + \frac{f_2}{f_0 + f_2} \times i$$

Here, $L = 40$, $f = 20$, $f_0 = 22$, $i = 10$

$$\therefore \quad M_0 = 40 + \frac{20}{22 + 20} \times 10 = 40 + \frac{100}{21}$$

$$= 40 + 4.76 = 44.76$$

(3) Empirical Relation Method

This is the most effective method of determining mode. Prof. Karl Pearson has envisaged this method. Prof. Pearson has found that in a moderately asymmetric or skewed series a pertinent relationship exists among the mean, median and mode. In such series the distance between mean and median is 1/3rd of the distance between the mean and mode.

Therefore 1/3 (mean – mode) = Mean – Mode

From this relationship Prof. Pearson has observed that

Mode = Mean – 3 (Mean – Median)

= 3 Median – 2 Mean.

A distribution with only mode is said to be unimodal. If a distribution has two equal maximum frequencies, then the distribution is said to be bio-modal distribution. If there are 3 or more equal maximum frequencies, then the distribution is said to be multimodal distribution. If there are two or more values having the same highest (maximum) frequencies, then the mode is said to be ill-defined. In such a circumstance mode cannot be obtained directly. Then, mode is calculated by the empirical relation method mode = 3 median – 2 man.

Use of Mode

1. When we want a quick and approximate measure of central tendency, we use mode.
2. Mode is the most typical or representative value of a distribution. Hence, when we talk of modal wage, modal size of shoe, modal score in the assessment, it is this average. The mode is the most frequently occurring value.
3. Like mean, mode is not affected by extreme values. Mode is that value which has occurred most often in the distribution.
4. Its value can be determined in open and distribution without ascertaining the class limit.
5. It can be used to describe qualitative phenomenon. For example, if we want to compare the consumer preference for difference types of products, say, soap, toothpaste, etc., we should complete the modal preferences expressed by different groups of people.

Limitations of Mode

1. The value of mode cannot be determined sometimes. In some cases, we may have a bimodal series.
2. It cannot be further treated algebraically like mean. For example, mode cannot be used for determining the combined mode of two or more groups.
3. The value of the mode is not based on each and every item of the series.
4. It is not rigidly defined as mean. There are several formulas for calculating mode, which yields different results.
5. While dealing with quantitative data, the disadvantages of the mode outweigh its goods features. Hence, it is seldom used.

3.2 Dispersion

We discussed the commonly used measures of central frequency which are useful in providing descriptive information concerning a set of data. After all it is a single numerical value and may fail to reveal the data entirely. This leads to conclude that a central value alone cannot describe the distribution adequately. Moreover, two or more sets may have the same mean but they may be quite different. Let us consider the scores of two groups of students on the same test:

Scores in 1st group	45	57	24	41	68	84	55	72	90
Scores in 2nd group	57	60	63	68	70	52	40	59	61

Here, both the groups have the same mean score of 60. From first inspection, we might say that the two sets of scores are equal in nature. In the first group, the range of scores is from 24 to 90, while in the second, the range is from 50 to 70. The difference in range shows that the students in the second group are more homogenous in scoring than those in the first. So both the groups differ widely in the variability of scores.

Measures of dispersion helps us in studying the extent to which observation are scattered about mean or central tendency.

Measures of dispersion are the following:

1. Range
2. Inter quartile range and quartile deviation
3. Variance
4. Standard deviation

3.2.1 Range

The range is defined as the difference between the largest and the smallest value in a set of observations. The range is a measure of observations among themselves and does not give an ideal about the spread of the observations around some central value. The range is defined by

$$R = X_L - X_S.$$

Where X_L is the largest of the observed values and X_S is the smallest of the observed values.

Example 2.14

Calculate the range of the following set of scores:

28, 21, 41, 30. 33, 37, 19, 27

Solution:

Here, the largest observed score = $X_L = 41$

The smallest observed score = $X_S = 19$

Therefore, the range (R) = $S_L - X_S$

$= 41 - 19 = 21$

1. What are the different measures of dispersion?
2. Calculate the range for the following set of scores:

   18, 40, 42, 46, 356, 25, 23, 17

Use of Range

1. It is useful for the data having small variations like rate of exchange.
2. It is mainly used in education.
3. It is used in meteorological department for quality control.
4. It is simple to understand and easy to calculate.

Limitations of Range

1. It is not based on all items of distribution.
2. It is affected by fluctuations of sampling.
3. It is affected by two extreme values.
4. It cannot be computed from frequency distribution with open end classed.

3.2.2 Interquartile Range and Quartile Deviation

The range as a measure of dispersion has certain limitations. It is based on two extreme items and it fails to take account of scatter within the range. Hence, a measure called interquartile range has been developed which includes the middle 50 percent of the distribution. This is, one quarter of the observations at the lower end and another quartile at the upper end are excluded in computing the interquartile range. Symbolically,

Interquartile range = $Q_3 - Q_1$

Very often the interquartile range is reduced to the form of quartile deviation by dividing it by 2. Symbolically,

$$\text{Quartile deviation or Q.D.} = \frac{Q_3 - Q_1}{2}$$

Q.D. is an absolute measure of dispersion. It gives the average amount of which two quartiles differ from the median. When quartile deviation is very small, it describes high uniformity or small variation of the 50% items and a high quartile deviation means that the variation among the central items is large. The relative measure corresponding to this measure called the coefficient of quartile deviation is calculated as follows:

$$\text{Coefficient of Q.D.} = \frac{Q3 - Q1/2}{Q3 - Q1/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

For comparative studies of variability of two distributions, we need coefficient of quartile deviation.

Example 2.15

Calculate the quartile deviation and its coefficient from the following data:

Roll No.	1	2	3	4	5	6	7
Marks	20	28	40	12	30	15	50

Solution: Calculation of Quartile Deviation

Marks arranged in ascending order: 12 15 20 28 30 40 50

$$Q_1 = \text{Size of } \frac{N+1}{4}^{\text{th}} \text{ item} = \frac{7+1}{4} = 2^{\text{nd}} \text{ item}$$

Size of 2nd item is 15. Thus $Q_1 = 15$

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right)^{\text{th}} \text{ item} = \frac{3\times 8}{4} = 6^{\text{th}} \text{ item}$$

Size of 6th item is 40. Thus, $Q_3 = 40$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{40 - 15}{2} = \frac{25}{2} = 12.5$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40 - 15}{40 + 15} = \frac{25}{55} = 0.455$$

Example 2.16

Compute coefficient of quartile deviation from the following data:

Marks	10	20	30	40	50	80
No. of students	4	7	15	8	7	2

Solution:

Calculation of Coefficient of Quartile Deviation

Marks	Frequency	c.f.	Marks	Frequency	c.f.
10	4	4	40	8	34
20	7	11	50	7	41
30	15	26	80	2	43

$$Q_1 = \text{Size of } \frac{N+1}{4} \text{ th item} = \frac{43+1}{4}\ 11^{\text{th}} \text{ item.}$$

Size of 11th item is 20. Thus, $Q_1 = 20$

$$Q_3 = \text{Size of } 3\left(\frac{N+1}{4}\right)^{\text{th}} \text{item} = \frac{43+1}{4} = 33^{\text{th}} \text{ item}$$

Size of 33rd item is 40. Thus $Q_3 = 40$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{40 - 20}{2} = 10$$

Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{40 - 20}{40 + 20} = 0.333$.

Example 2.17

Calculate the quartile deviation and its related coefficient from the following data:

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
No. of students	5	3	7	5	10	3	2

Solution:

Calculation of Q.D. and Coefficient of Q.D.

Marks	f	c.f.
0 – 10	5	5
10 – 20	3	8
20 – 30	7	15
30 – 40	5	20
40 – 50	10	30
50 – 60	3	33
60 – 70	2	35
	N = 35	

For Q_1, $\frac{N}{4} = \frac{35}{4} = 8.75$. The c.f. just greater than 8.75 is 15.

So, corresponding quartile class is 20 – 30.

Here, $l = 20$, h = 10, c.f. = 8, f = 7

$$Q_1 = l + \frac{\frac{N}{4} - \text{c.f.}}{f} \times h$$

$$= 20 + \frac{8.75 - 8}{7} \times 10$$

$$= 20 + \frac{0.75}{7} \times 10 = 20 + 1.07$$

$$= 21.07$$

For Q_3, $\frac{3N}{4} = \frac{3 \times 35}{4} = 26.25$. The c.f. just greater than 26.25 is 30. So, corresponding quartile class is 40 – 50.

Her $l = 40$, h = 10, c.f. = 20, f = 10

$$Q_3 = l + \frac{\frac{3N}{4} - \text{c.f.}}{f} \times h$$

$$40 + \frac{26.25 - 20}{10} \times 10 = 40 + 6.25 = 46.25$$

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = \frac{46.25 - 21.07}{2} = \frac{25.18}{2} = 12.59$$

and coefficient of Q.D. $= \dfrac{Q_3 - Q_1}{Q_3 + Q_1}$

$$= \frac{46.25 - 21.07}{46.25 + 21.07} = \frac{25.18}{67.32}$$

$$= 0.374.$$

Use of Interquartile Range (Quartile Deviation)

1. It is easy to calculate and simple to understand.
2. It considers 50% of the data, which is better than range.
3. When the data is in open-end classes, it is the only measure of dispersion.
4. When the data is skewed, containing a few very extreme scores.
5. When the measure of central tendency is available in the form of median.

Limitations

1. It ignores completely 50% of the observations.
2. It is very much affected by extreme values.
3. It is affected by fluctuations of sampling.
4. It is not capable for further mathematical treatment.

2.2.3 Standard Deviation

The standard deviation concept was introduced by Karl Pearson in 1823. It is the best and widely used measure of dispersion. It is free from those defects from which the earlier methods suffer and satisfies most of the properties of a good measure of dispersion.

Standard deviation is defined as the positive square root of the mean of the square of the deviations taken from the arithmetic mean. It is also known as root mean square deviation. It is denoted by the Greek letter σ (read as sigma).

The standard deviation measures the absolute dispersion or variability of a distribution. The greater the standard deviation the greater will be the magnitude of the deviations of the values from their means.

Calculation of Standard Deviation : Individual Series

(1) Direct Method: When actual data are used, the standard deviation is calculated by following formula:

$$\sigma = \sqrt{\frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2}$$

(2) Actual mean method: When deviations of items are taken from actual mean, then standard deviation is defined by the following formula.

$$\sigma = \sqrt{\frac{\Sigma x^2}{n}} \text{, where } x = X - \overline{X}$$

(3) Short cut method or Assumed mean method: When deviations of items are taken from assumed mean, then standard deviation is defined by following formula:

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2} \text{, where } d = X - A \text{ (Assumed mean)}$$

Note: If value of actual mean be in fraction (i.e. decimal), in that case short-cut method is more suitable method to calculate standard deviation than actual mean method.

Discrete and Continuous Series

(1) Direct method: When actual data are used, the standard deviation is calculated by following formula:

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

(2) Actual mean method: When deviations are taken from actual mean, then S.D. is defined by following formula:

$$\sigma = \sqrt{\frac{\Sigma fx^2}{N}} \text{, where } x = X - \overline{X} \text{, N= total frequency}$$

(3) Assumed mean method or Short-cut method: When deviations are taken from assumed mean, then S.D. is defined by following formula:

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \text{, where } d = X - A \text{, A = assumed mean}$$

(4) Step deviation method: When step-deviations are used, the S.D. is defined by following formula:

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i \text{, where } d' = \frac{X - A}{i} \text{, i = class size or common factor}$$

Note: X is the mid-value of corresponding classes for continuous series. For discrete series, i is taken as common factor from each given item, if possible.

Example 2.18

Find the standard deviation by (i) actual mean method and (ii) short-cut method from the following data:

30, 40, 42, 44, 46, 48, 58.

Solution:

Calculation of Standard Deviation by (i) Actual Mean Method (ii) Short-cut Method

X	$x = X - 44$	x^2	$d = X - 46$	d^2
30	-14	196	-16	256
40	-4	16	-6	36
42	-2	4	-4	16
44	0	0	-2	4
46	2	4	0	0
48	4	16	2	4
58	14	196	12	144
$N = 7, \Sigma X = 308$	$\Sigma x = 0$	$\Sigma x^2 = 432$	$\Sigma d = -14$	$\Sigma d^2 = 460$

For Actual Mean Method

Here, $\overline{X} = \frac{\Sigma X}{n} = \frac{308}{7} = 44$

Now, $\sigma = \sqrt{\frac{\Sigma x^2}{n}} = \frac{432}{7} = \sqrt{61.71} = 7.8$

For Short-cut Method

$n = 7$, $A = 46$, $\sum d = -14$, $\sum d^2 = 460$

$$\therefore \quad \sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{460}{7} - \left(\frac{-14}{7}\right)^2}$$

$$= \sqrt{65.71 - 4} = \sqrt{61.71} = 7.8$$

Example 2.19

Calculate standard deviation from the following data:

Marks	10	20	30	40	50	60
No. of students	8	12	20	10	7	3

Solution:

Calculation of Mean and Standard Deviation

Marks (X)	f	$d' = \frac{X - 30}{10}$	fd'	fd'^2
10	8	-2	-16	32
20	12	-1	-12	12
30	20	0	0	0
40	10	1	10	10
50	7	2	14	28
60	3	3	9	27
	N = 60		Σfd' = -5	$\Sigma fd'^2 = 109$

Here, A = 30, i = 10 (common factor), Σfd' = 5, $\Sigma fd'^2 = 109$

$$\overline{x} = A + \frac{\Sigma fd'}{N} \times i = 30 + \frac{5}{60} \times 10$$

$$= 30 + 0.83 = 30.83$$

$$\text{and } \sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i$$

$$= \sqrt{1.817 \times 0.0069} \times 10 = 1.345 \times 10 = 13.45.$$

Standard deviation = 13.45.

Example 2.20

Find the mean and standard deviation from the following data:

Marks	No. of Students	Marks	No. of Students
Upto 10	12	Upto 50	157
Upto 20	30	Upto 60	202
Upto 30	65	Upto 70	222
Upto 40	107	Upto 80	230

Solution:

Calculation of Mean and Standard Deviation

Marks	Mid-value (X)	f	$d' = \frac{X-45}{10}$	fd'	fd'^2
0 – 10	5	12	-4	-48	192
10 – 20	15	18	-3	-54	162
20 – 30	25	35	-2	-70	140
30 – 40	35	42	-1	-42	42
40 – 50	45	50	0	0	0
50 – 60	55	45	1	45	45
60 – 70	65	20	2	40	80
70 - 80	75	8	3	24	72
		N = 230		Σfd' = 105	$\Sigma fd'^2 = 733$

Here, N = 230, A = 45, Σfd' = -105, i = 10, $\Sigma fd'^2 = 733$

$$\bar{X} = A + \frac{\Sigma fd'}{N} \times i$$

$$= 45 + \frac{-105}{230} \times i$$

$$= 45 - 4.6 = 40.4$$

Standard deviation, $\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times i$

$$= \sqrt{\frac{733}{230} - \left(\frac{-105}{230}\right)^2} \times 10$$

$$= \sqrt{3.2 - 0.21} \times 10 = \sqrt{2.99} \times 10$$

Standard deviation = 1.729 × 10 = 17.29

3.2.4 Variance

The term variance was used to describe the square of the standard deviation by R.A. Fisher in 1913. The concept of variance is highly important in advanced work. It is defined as:

$$\text{Variance} = \frac{\Sigma(X - \bar{X})^2}{N}$$

Thus, variance is nothing but the square of the standard deviation i.e. variance = σ^2 or $\sigma = \sqrt{\text{Variance}}$

Example 2.21

The following table gives the marks obtained by a group of 80 students in an examination. Calculate the variance and standard deviation.

Marks Obtained	No. of Students	Marks Obtained	No. of Students
10 – 14	2	34 – 38	10
14 – 18	4	38 – 42	8
18 – 22	4	42 – 46	4
22 - 26	8	46 - 50	6
26 – 30	12	50 – 54	2
30 – 34	16	54 – 58	4

Solution:

Calculation of Variance and Standard Deviation

Marks	Mid-value (X)	f	$d' = \frac{X - 32}{4}$	fd'	fd'
10 – 14	12	2	-5	-10	50
14 – 18	16	4	-4	-16	64
18 – 22	20	4	-3	-12	36
22 – 26	24	8	-2	-16	32
26 – 30	28	12	-1	-12	12
30 – 34	32	16	0	0	0
34 – 38	36	10	+1	+10	10
38 – 42	40	8	+2	+16	32
42 – 46	44	4	+3	+12	36
46 – 50	48	6	+4	+24	96
50 – 54	52	2	+5	+10	50
54 – 58	56	4	+6	+24	144
		N = 80		Σfd' = 30	$\Sigma fd'^2 = 562$

$$\text{Variance} = \left(\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2\right) \times i^2$$

$\sum fd'^2 = 562, \sum fd' = 30, N = 80$

$$\text{Hence, variance} = \left(\frac{562}{80} - \left(\frac{30}{80}\right)^2\right) \times 4^2$$

$= (7.025 - 0.141) \times 6 = 6.884 \times 16 = 110.144$

Variance = 110.144.

Uses of Standard Deviation

1. It is based on all observations.
2. It is free from those defects with which other methods like range and inter quartile range suffers.
3. It is the best measure of variation possessing excellent sampling properties.
4. It is used when we need most reliable measure of variability.
5. When there is a need of computation of correlation coefficients and significance of difference between means.
6. When the distribution is normal or near to normal.

Limitations of Standard Deviation

1. As compared to other measures, it is somewhat more difficult to understand and compute.
2. Its value is unduly affected by extreme observations.

2.3 Measures of Relative Position

In a class of 50 students if one student has obtained a score 60 on a certain achievement, then in order to gather some idea about his performance, we must have clear information regarding his position in the class. Has the student got the highest or lowest score in the class? How many students have got more than his score? How many students have secured less than him? All such questions are concerned with the relative position or performance of an individual in relation to a reference group may be answered through the concepts of percentile, percentile ranks and standard scores.

2.3.1 Percentile

Percentile is nothing but a sort of measure used to indicate the relative position of a single item or individual with reference to the group to which the item or individual belongs. In other words, it is used to tell the relative position of a given score among other scores. A percentile refers to a point in a distribution of scores below which a give percentage of the cases occur.

The percentile is named for the percentage of cases below it. Hence, let percentile (written as P_1) will mean "a score point in the given distribution below which one percent cases lie and above which 99% cases lie. Similarly, 15th percentile (P_{15}) will indicate the score point below which 15% and above 85% members of the group fall.

The middle of a distribution or a point below which 50 percent of cases lies in the fiftieth percentile P_{50}, which is the median, which has been discussed in the previous section. Similarly, P_{25} and P_{75} are the first quartile Q_1 and third quartile Q_3 respectively.

Computation of Percentile

When we wish to compute percentile, we will determine the score below which a gives percent cases will fall. First the class in which the p^{th} percentile lies may be identified. This is the class in which $PN/100^{th}$ frequency falls. The formula for computing percentile is as follows:

$$P_P = L + \frac{\left(\frac{PN}{100} - c.f.\right) \times i}{f}$$

Where L is the lower limit of P^{th} percentile class, N is the total number of cases, c.f. is the less than cumulative frequency of class preceding to percentile class, f is the frequency of the percentile class and i is the class interval.

Example 2.22

Find the 25th percentile (P_{25}) and 60th percentile (P_{60}) for the following distribution.

Scores	f	c.f. (less than cumulative frequency)
70 – 79	3	24
60 – 69	2	21
50 – 59	2	19
40 – 49	3	17
30 – 39	5	14
20 – 29	4	9
10 – 19	5	5
	N = 24	

Solution: (For P_{25})

Since, $\frac{PN}{100} = \frac{25 \times 24}{100}$ = 6th frequency lies in 20 – 29 class P_{25} class is 20 – 29. Therefore, the actual lower limit of the class (L) = 19.5.

c.f. = 5, f = 4 and i = 10

Substituting the values in the formula,

$$P_{25} = L + \frac{\frac{24N}{100} - c.f.}{f} \times i$$

$$= 19.5 + \frac{\left(\frac{25 \times 24}{100}\right) - 5}{4} \times 100$$

$$= 19.5 + 2.5 = 22.$$

In this frequency distribution, 22 is the point which 25 percent of cases will fall.

For P_{60}

Since, $\frac{PN}{100} = \frac{60 \times 24}{100}$ = 14.4th frequency lies in 40 – 49 class, hence, P_{60} class is 40 – 49. Therefore, the actual lower limit of the class (L) is 39.5.

c.f. = 14, f = 3, i = 10

Substituting the values in the formula,

$$P_{60} = L + \frac{\frac{60N}{100} - c.f.}{f} \times i$$

$$= 39.5 + \frac{\left(\frac{60 \times 24}{100}\right) - 14}{3} \times 10$$

$$= 39.5 + \frac{4}{3} = 39.5 + 1.33$$

$$P_{60} = 40.83$$

3.4.2 Percentile Ranks

The percentile rank of a given score in a distribution is the percent of the total scores which fall below the given score. A percentile rank then indicates the position of a score in a distribution in percentile terms.

In a class there are 50 students. One student Ram obtains a score of 60 in an achievement test. There are 40 other students whose scores fall below the score 60. In other words, 40 out of 50 (40/50 × 100 = 80%) students lie below the score 60. Therefore, 60 will be termed as the 80th percentile of the given data. In conclusion, we can say that the achievement score of 60, obtained by Ram has a percentile rank of 80.

Computation of Percentile Rank

In case of ungrouped data, the data is first arranged in an order (descending order preferably) and each score is provided a rank or a relative position in order of merit. Then the following formula is employed.

Percentile rank, $PR = 100 - \frac{100R - 50}{N}$

N = Total number of individuals in a group

R = Rank position of the score whole percentile rank is to be determined.

Example 2.23

In an achievement test, 10 students of a class have scored as below:

44, 42, 28, 25, 40, 37, 29, 32, 35, 36

Find out the percentile rank of score 29.

Solution:

Scores	Rank Order	Scores	Rank Order
44	1	35	6
42	2	32	7
40	3	29	8
37	4	28	9
36	5	25	10

Here, rank R of the desired score 29 is 8 and N = 10.

Substituting these figures into the formula, we get,

$$\text{PR} = 100 - \frac{100 \times 8 - 50}{10} = 100 - \frac{750}{10}$$

$$= 100 - 75 = 25$$

The percentile rank of the score 29 is 25. This indicates 25 percent of the cases fall below the score 29.

Example 2.24

In an entrance examination, a candidate ranks 35 out of 150 candidates. Find out his percentile rank.

Solution:

The percentile rank is computed by the formula,

$$\text{PR} = 100 - \frac{100\text{R} - 50}{\text{N}}$$

In this case R = 35, and N = 150

$$\text{Hence, PR} = 100 - \frac{(100 \times 35) - 50}{150} = 100 - \frac{3450}{150}$$

$$= 100 - 23 = 77$$

The percentile rank is 77.

To compute the percentile rank for a score from the grouped data, the following formula is used:

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

Example 2.25

Find the percentile rank of the score 22 for the following distribution:

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

Solution:

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

$$= \frac{100}{24}\left(\frac{22-19.5}{10}\right)\times 4 + 5 = \frac{100}{24}\times\frac{2.5}{10}$$

$$= \frac{100}{24}\times 6 = 25.$$

The percentile rank of score 22 is 25.

Uses of Percentile and Percentile Ranks

1. Percentiles and percentile ranks can be used to indicate the relative position of an individual with respect to some attribute (achievement score, presence of some personality, etc.) in his own group.
2. For comparing two or more individual student belonging to two or more sections or schools.
3. For comparing the performance of two or more classes or schools.
4. For comparing the performance of an individual if tested under two or more different testing conditions.

3.4.3 Standard Score

Another method of indicating a student's relative position in a group is by showing how far the raw score is above or below the average. This is the approach used with standard scores. Basically standard scores express test performance in terms of standard deviation units from the mean.

The relationship between standard deviation units and percentile ranks enables us to interpret standard scores in simple and familiar terms. Standing from the left of the normal curve, each point on the base line of the curve can be equated to the following percentile rank.

– 2 SD = 2%

–1 SD = 16% (2 + 14)

0 SD = Mean = 50% (16 + 34)

+1 SD = 85% (50 + 34)

+ 2SD = 98% (84 + 14)

For example, –2 SD equals a percentile rank of 2 because 2 percent of the cases fall below that point.

Types of Standard Scores

There are numerous types of standard scores used in education and testing. Because they all are based on the same principle and interpreted in somewhat the same manner. Only the most common types of standard scores are discussed here.

3.4.3.1 Z-scores

The simplest of the standard scores is the Z-score. This score expresses test performance simply and directly as the number of standard deviation units a raw score is above or below the mean. For converting raw score into a Z-score, the steps to be taken as follows:

1. First compute mean ($\overline{X}$) and standard deviation (σ) of the distribution.
2. Then substitute the value of $\overline{X}$ and σ in the following formula for computing Z-score.

$Z = \frac{X - \overline{X}}{\sigma}$, here X stands for raw score.

Example 2.26

There are two sections – A and B in class IX of a school. To test their achievement in Math, two different question papers are prepared. Ram of section A got 80 marks, while Suresh, a student of section B got 60. Can you say which of these two students stand better in terms of achievement in Math? Mean and standard deviation of the distribution of scores for section and B are as follows:

Section A	Section B

Mean = 70
SD = 20
	Mean = 50
	SD = 20

Solution:

Here, we cannot conclude that Ram with 80 marks is a better student in Math as compared to Suresh who obtained only 60 marks. The paper set for section A might have quite easy or so many other variations for section B. As a result, the marks obtained by these two students do not belong to same scale of measuring. Their raw scores need to be transformed into Z-score or standard score for comparison.

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

Z score of Ram = $\frac{X-\overline{X}}{\sigma} = \frac{80-70}{20}$ = 0.5

Z score of Suresh = $\frac{X-\overline{X}}{\sigma} = \frac{60-50}{10}$ = 1.0

Thus, we can conclude that Suresh is placed better with 1σ score in terms of his achievement compared to Ram with 0.5 σ score.

Uses

1. Standard scores have essentially the same meaning for all tests.
2. Transformation of raw scores into Z-scores does not change the shape or characteristics of the distribution.
3. Z-scores can be safely used for inter-individual and intra-individual comparisons.

Limitations

1. In Z-scores, plus and minus signs are used. These can be overlooked or misunderstood.
2. Decimal points may create difficulty in interpretations.

3.4.3.2 T-scores

In the scale of measurement involving Z-scores, the starting point is the mean of the distribution, i.e. zero and the unit of measurement is 1 SD of the distribution. From zero on this scale to both sides may involve minus and plus signs and the use of unit of measurement, 1 SD may carry decimal points. To overcome such limitation of Z-scores, a more useful named T-scores (T-scale) may be used. This scale was derived and first used by Willian A. McCall and named T-scale in honour of Thorndike and Terman. In this scale, another type of score, slightly different from the Z-scores is used.

T-scores may be defined as normalized standard scores converted into a distribution with a mean of 50 and σ(SD) of 10. T-scores can be obtained by multiplying the Z-score by 10 and adding the product to 50. Hence,

T-score = 50 + 10 (Z)

In the above example, Ram's Z-score is 0.5 and Suresh Z-score is 1.0. Their T-score is computed as follows:

Ram's T-score = 50 + 10 × 0.5 = 55

Suresh's T-score = 50 + 10 × 1.0 = 60

3.4.3.3 Stanines

Stanines (pronounced stay-nine) are a simple type of normalized standard score. Stanines are single-digit score ranging from 1 to 9. This system of scores is so named because the distribution of raw scores is so divided into nine parts (standard nine).

Stanines 5 is precisely in the center of distribution and included all cases within one-fourth of a standard deviation on either side of the mean. Any score with a percentile rank of at least 40 but less than 60 is converted to a Stanine score of 5. Thus raw scores corresponding to percentile ranks between 40 and 59 all convert to Stanine score of 5. The remaining stanines are evenly distribution above and below Stanine 5. With the exception of Stanines 1 and 9, which cover the tails of the distribution, each Stanine includes a band of raw scores the width of one-half of a standard deviation unit. Stanines have a mean of 5 and a standard deviation of 2. The conversion from percentile ranks to Stanines is straight forward for a normal distribution as shown below.

Relationship of Percentile Ranks and Stanine

Range of Percentile Ranks	Stanine
96 or higher	9
89 to 95	8
77 to 88	7
60 to 76	6
40 to 59	5
23 to 39	4
11 to 22	3
4 to 10	2
Below 4	1

Uses

1. The Stanine system uses a nine point scale in which 9 is high, 1 is low and 5 is average. The system is easily understood by the students and parents.
2. Stanine are normalized standard scores that make it possible to compare student's performance on different tests.
3. The stanine system makes it easy to combine diverse types of data (e.g. test scores, ratings, ranked data) because Staynines are computed like percentile ranks but are expressed in standard score form.

Limitations

1. The main limitation of Stanine scores is that growth cannot be shown from one year to the next.
2. Stanines are criticized on the grounds that they are rather crude units because they divide a distribution of score into only nine parts.
3. Percentile ranks and other standard score indicate relative position in a particular group but Stanines do not.

4. Summary

Central tendency is an average which represents all the scores made by the group and as such gives concise description of the performance of the group. The most commonly used measures of central tendency are as follows: (1) arithmetic mean, (2) median and (3) mode.

The arithmetic mean, commonly called the mean or average. It is defined as the sum of measures divided by the number of measures. It is denoted by $\overline{X}$. The value lying at the middle of series in known as median. Mode is value which occurs maximum number of times.

Measures of dispersion describes the extent to which observations are scattered about mean. Measures of dispersion are as follows: (1) range, (2) inter-quartile range, (3) variance, and (4) standard deviation.

The range is defined as the difference between the largest and the smallest value in a set of observations. Inter-quartile range includes the middle 50 percent of the distribution. Standard deviation is defined as the positive square root of the square of the deviations taken from the arithmetic mean. Variance is nothing but the square of the standard deviation.

The relative position or performance of an individual in relation to a reference group may be answered through percentile, percentile ranks and standard scores.

The percentile is names for the percentage of cases below it. A percentile rank indicates the position of a score in a distribution in percentile terms.

There are numerous types of standard score used in education and testing. Most common types of standard scores are Z-scores, T-scores and Stanines. The simplest standard score is Z-score. Z-score expresses test performance simply and directly as the number of standard deviation units a raw score is above or below the mean. T-score may be defined as a normalized standard score converted into a distribution with a mean of 50 and SD of 10. Stanines are single digit, normalized standard scores ranging from 1 to 9. Stanines 5 is precisely in the center of distribution.

5. Exercise

Objective Questions

1. What is the median of the series:

   1, 4, 6, 7, 2, 8, 7?

   (a) 5 (b) 6 (c) 7 (d) 8

2. What is the mode of the series: 1, 4, 6, 7, 2, 8, 7?

   (a) 5 (b) 6 (c) 7 (d) 8

3. Median is

   a. Sum of measures divided by number of measures.

   b. Value lying at the middle of series.

   c. Value which occurs maximum number of times.

   d. Square of the standard deviation.

4. Which statement is true?

   a. Variance is the square root of the standard deviation.

   b. Variance is the square root of the range.

c. Standard deviation is the square of the variance.

d. Standard deviation is the square root of the variance.

5. The difference between the largest and the smallest value in the series is known as:

a. Range b. Percentile rank

c. Inter-quartile range d. Standard deviation.

6. Which one of the following includes middle 50 percent of the distribution in computing dispersion?

a. Percentile rank b. Standard deviation

c. Inter-quartile range d. Variance

7. What will be the Z-score of a student who scored 80 marks, if mean is 70 and standard deviation is 10.

a. 10 SD b. 1 SD c. 2 SD d. 0.5 SD

8. What is the T-score of Ram, if his Z-score is 0.5 SD?

a. 100 b. 55 c. 50 d. 25

9. Stanine 5 includes

a. 50 – 59 percentile rank b. 50 – 69 percentile rank

c. 40 – 49 percentile rank d. 40 – 59 percentile rank

Short Answer Questions

1. Describe the major uses and limitations of arithmetic mean.
2. Describe the uses and limitations of median.
3. Describe various tools of computing dispersion.
4. Describe inter-quartile range and explain its uses.
5. Describe various tools of measuring relative position.
6. Compute mean and median from the following list of test scores:

   50, 52, 50, 56, 68, 65, 62, 57, 70
7. Find the mean, median and mode for the following test scores:

   20, 14, 12, 14, 19, 14, 18, 12.

Long Answer Questions

1. Describe the various methods of computing mode with suitable examples.
2. What are the various methods of computing measures of dispersion? Describe with suitable example.
3. The following are the data given in the form of frequency distribution:

Scores	60 – 69	50 – 59	40 – 49	30 – 39	20 – 29	10 - 19
f	1	4	10	15	8	2

Calculate mean, median and mode.

4. From the above data Q.N. 3, compute standard deviation and interpret.

5. Compute the mean, median and mode for the following frequency distribution.

Scores	90 – 94	85 – 89	80 – 84	75 – 79	70 – 74	65 – 69
f	5	10	14	15	7	3

6. Compute the values of the following:

(a) P_{30}, P_{70} and P_{85}

(b) Percentile ranks of the scores 14, 20 and 26.

(i) Score	F	(ii) Score	f
34 – 36	12	40 – 44	6
31 – 33	15	35 – 39	5
28 – 30	19	30 – 35	10
25 – 27	16	25 – 29	9
22 – 24	8	20 – 24	11
19 – 21	9	15 – 29	10
16 – 18	8	10 – 14	9
13 – 15	3		

7. Compute inter-quartile range from the following data:

Marks	14	20	30	35	40	45	50
Frequency	4	7	8	10	7	5	3

8. Transforms the following raw scores to Z-score and T-score formats. The raw score distribution has a mean of 70 and standard deviation of 10.

(a) Raw score = 85 (b) Raw score = 55 (c) Raw score = 90

9. Convert the following Z-score to T-scores.

(a) Z-score = 1.6 (b) Z-score = –1.5 (c) Z-score = 2.5

10. Calculate mean, standard deviation and variance of the following score distribution:

10, 9, 8, 7, 6, 6, 6, 5, 5, 5, 4, 4, 3, 2, 2.

Chapter 3

The Normal Probability Curve

3.0 About the Chapter

This chapter is divided into five sections related to concept of normal probability curve, nature of normal probability curve, properties of normal probability curve, different forms of deviations of normality, and application of normal probability curve, The concept of normal probability curve describes the meaning and definition, the nature of normal probability curve states the characteristics of family of normal probability curve and their relation with measure of central tendencies (mean, median, and mode). The third section explains three different properties of normal probability curve. The divergence of normality explains the skewness and kurtosis with their application. The final section describes the use of normal probability curve for different purposes in education.

On the complete study of this chapter the following specific objectives are assumed to be achieved.

- To explain the nature of normal probability curve.
- To describe skewness and kurtosis as the deviations of a normal curve.
- To apply normal probability curve to interpret the given set of data.

3.1 Concept of Normal Probability Curve

The word 'normal' refers to the average of a thing/trait/attributes, etc. In education and social science we may encounter in such situations, in whi traits are distributed in average manner. For example, if w students of the same age level, many of their height fall (average). But the individual height of few students devi below of the average height. Similarly, if we take data, rando intelligence, wealth, achievement, etc. through test, survey plot the frequency of such distribution on graph, we would g This bell shaped curve representing the average characteristics of certain thing is called *normal curve.*

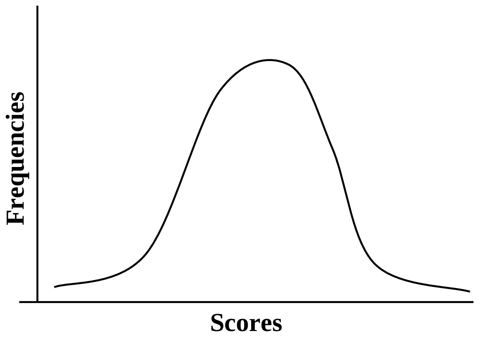


Figure 3.1: Normal Curve

Similarly, if we toss a coin a large number of times we will get certain number of head and tails. The occurrence of number of head or tail represents the chance success or probability. When plotted the frequencies of these numbers on graph paper it gives a frequency curve like normal curve.

Therefore, the curve representing chance success or probability is also known as normal probability curve. French mathematician Abraham de Moivre discovered the formula for normal curve in 1733. At the beginning of 19th century, mathematician Pierre - Simon (French) and Carl Friedrich Gauss (German) rediscovered the normal curve independently. Laplace and Gauss worked on experimental errors in physics and astronomy and found the errors to be distributed normally. Therefore, the curve is also called *'curve of error'* or Gaussian curve. Adolphe Quetlet, a Belgian statistician and sociologist, made use of normal curve wide in mid 19th century in meteorology, anthropology, and human affairs. He believed that mental and moral traits, when measured, would conform to the 'normal law'. In the latter part of the nineteenth century, the British genius Sir Francis Galton began the first serious study related to mental and physical traits and conformed reasonably well to the normal curve.

From above account, normal curve is a symmetrical bell shaped curve representing average mental and moral traits. But it is not essential for a normal distribution to be exactly perfect bell shaped curve. In our practice we cannot measure an entire population but we work with sample. Therefore, slightly deviated or distorted bell-shaped curve is also accepted as the normal curve on assumption of normal distribution of the particular characteristics measured in the entire population.

3.2 The Nature of Normal Probability Curve

The normal probability curve and normally (or approximately) distributed data are different things. The normal curve is a mathematical abstraction having a particular defining equation. As mathematical abstraction, it is not associated with any event or events with the real world. It is not, therefore, a law of nature contrary to the thought and terminology associated with it a century ago. For instance, the equation of circle (x^2 + a^2 = r^2) which represents circles with different radius (family of circles) is a mathematical abstraction. The equation circles with different radius gives the equation of normal curve and describes a family of normal curves. The family differs in their means and standard deviations. There are three possibilities: equal means and unequal standard deviation (*Fig. 3.2a*), unequal means and equal standard deviations (*Fig. 3.2b*), and unequal means and unequal standard deviations (*Fig. 3.2c*).

Therefore, the normal curve is a mathematical abstraction defined by an equation and represents different curves with similar nature. Though the means and standard deviations of normal curves are different, they have similar nature in terms of symmetry, mode, and continuity, i.e. the normal curves are symmetrical, unimodal, and continuous with tails that are asymptotic to horizontal axis. The area under all the normal curves is also distributed in similar pattern.

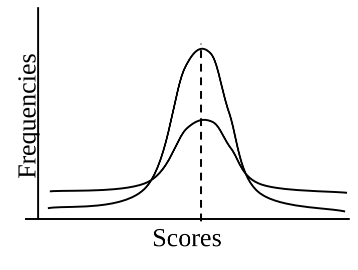


Figure 3.2a: Equal means and unequal σ

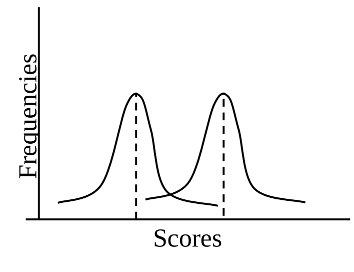


Figure 3.2b: Unequal means and equal σ

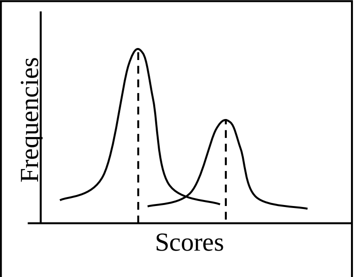


Figure 3.2c: Unequal means and unequal σ

As discussed in previous section, many mental and physical traits represent a normal curve. The distribution of data related to these traits in normal pattern is a law of nature or is a law of cosmic order. Therefore, it is useful to describe the data related to human traits. But many psychologist express doubt concerning the utility of normal probability curve as a model for the distribution of real world variables. For example, the test scores representing mental ability are seldom normally distributed since the tests are constructed with items positively interrelated. But the normal curve has been a useful mathematical model for statistician since they believe that the distribution of infinity number of sample means approaches normality as the sample size increases to population.

More clearly, the family of curves has some *common characteristics* describing the nature of normal curve.

1. All the normal curves have symmetrical distribution. That is, the left half of the normal curve is the image of right half.
2. Normal curves are unimodal (having single mode) distributions with the mode at the centre.
3. The measure of values of central tendencies coincide at a single point. That is, the mean, median, mode all have the same value.
4. Starting from centre of the curve, the height of the curve descends gradually at first, then faster, and finally slower.
5. At extreme ends of the curve, it never touches the horizontal axis. That is, it extends from minus infinity to plus infinity. Therefore, the nature is called asymptotic.
6. The curve covers certain area under it. That is, approximately two thirds (68.26%) of the area under a normal curve lies within the range of $\bar{X}$ ± 1σ, 95.44% area of the normal curve lies within the range $\bar{X}$ ± 2σ, and 99.74% area lies within the range of $\bar{X}$ ± 3σ.

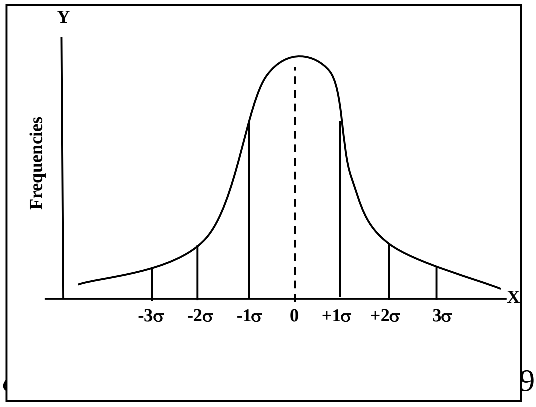


Figure 3.3: Areas within certain limits under normal curve.

7. The limit of the distances $\bar{X}$ ± 1.96 9% area under normal curve. That is, 5% and 1% area fall beyond this limit.
8. The normal curve has its maximum height (ordinate) at the mean of the distribution. The mean is starting point for working with normal curve. In unit normal curve the height of normal curve at mean equals 0.3989.
9. The curve changes from convex to concave (from outer side) at $\bar{X}$ ± 1σ. This point, therefore, is known as *point of inflection*.

10. The normal curve can be taken as a model for describing the flatness through kurtosis. For normal curve, the skewness is zero and kurtosis is 0.263. If the kurtosis is more than that value, the distribution is more flat at top and if the value of kurtosis is less than that value the distribution is more peaked at the centre.
11. The normal curve is an approximation of real world. Though the curve is continuous distribution, the values of observation form a discrete series.

3.3 Properties of Normal Probability Distribution.

1. *The normal curve as a representation of certain equation*.

Since the normal curve is a mathematical abstraction, the family of normal curve represents by the following equation.

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}} \text{ .... (3.1)}$$

In which, x = X- $\overline{X}$, x is a point at baseline, $\overline{X}$ is mean

y = the height of the curve about X axis.

N = number of cases

σ = standard deviations of the distribution.

π = ratio of circumference of circle to its diameter (3.1416)

e = base of the Napierean system of logarithm (2.7183) or exponential function.

In above equation when N and σ are known, we can compute the height (y) of the curve for a given value of x, and the percentage between two points, or above or below a given point in the distribution.

The frequency interpretable as the area under the normal curve that falls between two values of X. The change of distance between two values of X is approximately proportional to the height of the curve. The y is, therefore, an indication of frequency.

Note 3.1: Though we can compute the frequency for a given value of *X*, it is computed rarely since the tabulated value in statistics books is available. But the knowledge of computation makes the use of table more comprehensible.

2. *The normal curve covers certain area under it.*

If the scores related to individual traits or performance is distributed normally, then it can be expressed in standard score (or *z* -score) form. For this we apply,

$$z = \frac{X - \overline{X}}{\sigma} \text{ ....(3.2)}$$

In which, z = Standard score or sigma score

X = Row score obtained by individuals

$\overline{X}$ = Mean of the distribution

σ = Standard deviation of distribution.

It converts the raw score mean to 0 in standard score form. As already discussed in nature of normal curve, 34.13% of the area under normal curve lies between -1σ and mean and since the curve is symmetric about mean, the same amount of area lies between +1σ and mean under normal curve. Therefore, 68.26% of the area under normal curve lies within the range of $\overline{X}$ ± 1σ, 95.44% area of the normal curve lies within the range $\overline{X}$ ± 2σ and 99.74% area lies within the range $\overline{X}$ ± 3σ.

These areas and other fractional parts of the area can be obtained directly using normal curve table (Appendix A). The total area under the curve is taken arbitrarily to be 10000 (to make ease to calculate the fractional parts of the total area). The first columns (*x/σ*) of the table gives the one tenth of standard score unit and first row gives the fractional values for each σ measure (in hundredth of *σ*).

Now to find the number of cases in the normal distribution between mean and the ordinate (parallel line with *Y*-axis) erected at a distance of 1σ from the mean, see vertically downward in first column until you reached 1, and see horizontally right from the same point to find the required number of cases for fractional parts under normal curve for 1σ, if there is no fractional parts then see the value in the column under .00 then you will get 3413. This means that 3413 cases in 10000 or 34.13%.of the entire area of the curve lies between mean and 1σ. Since the curve is symmetric about mean, the area between mean and -1σ is also covers 34.13%. This concludes that the total area between ±1σ is 68.26%. In the similar way we can find the total area under normal curve for other specified limits.

3. *Relation of normal probability curve with constants*

Since the normal curve is bilaterally symmetrical, all the measures of central tendencies (mean, median, and mode) must coincide at the centre of the distribution. Similarly, the measure of variability includes certain constant fractions of the total area of the normal curve (which can be obtain using normal probability curve table in Appendix A). That is, middle 68.26% cases lies between mean and $\pm 1\sigma$, 95% of the distribution lie between mean and $\pm 2\sigma$, and 99.7% cases lie between mean and $\pm 3\sigma$. Therefore, about 68 chances in 100 that a case will lie within $\pm 1\sigma$ from the mean, 95 chances in 100 that it will lie within $\pm 2\sigma$ from mean, and 99.7 chances in 100 that it will lie within $\pm 3\sigma$ from mean.

We may use quartile $Q = \frac{Q_3 - Q_1}{2}$ as a unit of measurement instead of sigma (σ) to determine the area within given parts of normal curve. In case of normal curve Q is called the *probable error (PE)*. The relation between PE and σ are given below:

$$PE = 0.6745\sigma \Rightarrow \sigma = 1.4826\, PE \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3.3)$$

From above facts, it is clear that σ is almost 50% larger than PE. Therefore, from normal probability curve table, 50% include between mean and ±1 PE (25% just above and 25% just below the mean). The middle 50 defines the range of normal performance. The upper 25% is somewhat better and lowest 25% is somewhat poorer than the typical middle or average group. Similarly ± 2 PE (or $\pm 1.3490\sigma$) includes 82.26% ± 4PE (or $\pm 2.6980\sigma$) includes 99.30% of the measures in the normal curve.

3.4 Measuring Divergence from Normality

A normal curve is a perfectly symmetrical bell shaped curve. The three values of measure of central tendencies (mean, median, and mode) coincide at a single point (mean). When the curve deviates from normality these three values differ from one another. Therefore, the extent of divergence can be measured with the help of following measures.

3.4.1. Skewness

The term skewness refers to the lack of symmetry. A normal curve is symmetric around its mean and the values of median and mode also fall at the mean. As the curve is deviate from normality it also deviates from symmetry. The deviation of distributions from symmetry or lack of symmetry can be obtained with some measure called skewness. Therefore, the term skewness refers to the lack of symmetry and the distribution is said to be skewed when the mean and the median fall at different points in the distribution. In such situation the centre of gravity (mean) is shifted from one side to another (to the left or right). The distributions are said to be skewed negatively (or to the left) when scores are massed at the right and spread out more gradually toward the left end. The distributions are skewed positively when scores are massed at the left end of the scale, and spread out gradually toward the right end.

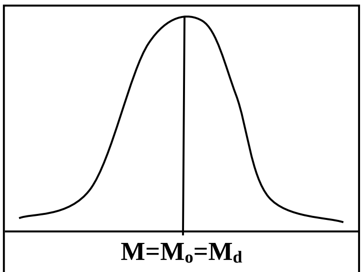


Figure 3.3a: Normal distribution

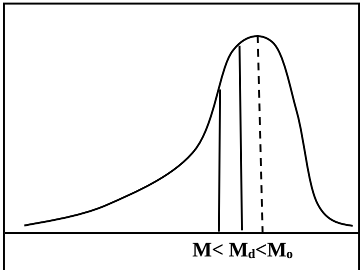


Figure 3.3b: Negatively skewed

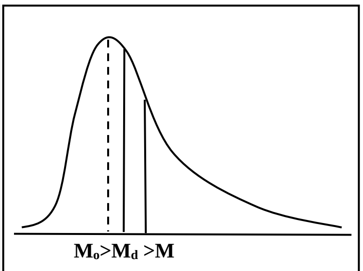


Figure 3.3c: Positively skewed

From above figures it is clear that the mean(M), median(M_d), and mode(M_o) lies at a single point when the curve is normal. Mean lies to the left of median when curve is negatively skewed, and mean lies to the right of median when curve is positively skewed. The more nearly the distribution approaches the normal form, the closer together are the mean and median, and less the skewness. When the distribution is normal, mean, and median coincides and therefore the skewness (S_k) is zero.

In sum, if $S_k = 0$, the distribution is symmetrical

If $S_k > 0$, the distribution is positively skewed

If $S_k < 0$, the distribution is negatively skewed.

In practical language, the distribution is said to normal when the individuals are scattered equally on both sides of mean. The distribution is said to be skewed negatively when there are many individuals in a groups whose scores are higher than the group average. Similarly, the distribution is said to be positive when there are more individuals whose scores are lower than the group average.

Measures of skewness

To find out the direction and the extent of symmetry in a given distribution, any of the following measures of skewness are employed. Since skewness is the extent of difference of median or mode from its centre of gravity (mean) then we can use following simple measures.

- Skewness = Mean - Median or Mean - Mode (In terms of measure of central tendency)
- Skewnewss = $(Q_3 - M_d) - (M_d - Q_1) = Q_3 + Q_1 - 2M_d$ (In terms of quartile)
- Skewness = $(P_{90}-P_{50}) - (P_{50} - P_{10}) = (P_{90}+P_{10}-2P_{50}$ (In terms of percentile)

The above measure are absolute and tell us the extent of asymmetry with the nature of positively or negatively skewed distribution. If distributions are available with two different units, then we cannot use the results of two distributions for comparison purpose. Thus, for comparing two or more distributions for skewness, we may use following relative measures called the coefficients of skewness.

I. Karl Pearson's measure of skewness: The relative measure of skewness based on mean, mode, and standard deviation called Karl Pearson's coefficient of skewness or Pearsonian coefficient of skewness. It is defined by following formula:

$$\text{Skewness } (S_k) = \frac{\text{Mean } (\overline{X}) \text{ - Mode } (M_o)}{\text{Standard Deviation } (\sigma)} = \frac{\overline{X} - M_o}{\sigma} \text{ .... } (3.4)$$

If mode is ill-defined, then it is difficult to find the accurate value of mode. In such case, we may replace mode with median formula (i.e., mode = 3 median - 2 mean)

$$\therefore S_k = \frac{\overline{X} - (3M_d - 2\overline{X})}{\sigma} = \frac{3\,(\overline{X} - M_d)}{\sigma} \text{.....................................}(3.5)$$

Theoretically, the value of skewness lies between the limits of ±3 but practically the value generally lies between the limits of ±1.

II. The Bowley's measure of skewness: The Bowley coefficient of skewness is based on the quartile of the given distribution. The absolute measure of skewness based on quartiles (called Bowley's measure of skewness) is given by the following formula.

Skewness = $(Q_3 - M_d) - (M_d - Q_1) = Q_3 + Q_1 - 2M_d$.........................(3.6)

In which, Q_3 = Third quartile

Q_1 = First quartile

M_d = Median

For comparison purpose, we use the relative measure of skewness based on quartiles called the Bowley's coefficient of skewness. It can be defined by the following formula;

$$S_k = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} \text{ .... } (3.7)$$

Note 3.2: The value of Bowley's coefficient of skewness lies between the limits ±1.

The Bowley's coefficient of skewness is useful especially when

- the mode is ill-defined and extreme observations are present in the data.
- the distribution is open ended or the class intervals are unequal.

III. Kelly's measure of skewness: The Bowley's measure of skewness is based only on central 50% data and ignores remaining data in extreme ends. To remove the weakness on Bowley's measure of skewness, Kelly took the two percentiles from the median and developed new formula as given below.

Skewness = $(P_{40} - P_{50}) - (P_{50} - P_{10}) = P_{90} + P_{10} - 2P_{50}$...........(3.8)

For comparing two series, the Kelly's coefficient of skewness based on percentile is

$$S_k = \frac{P_{90} + P_{10} - 2P_{50}}{P_{90}-P_{10}} \text{ ....(3.9)}$$

Exercise 3.1

1. Calculate the Karl Pearson's Coefficient of Skewness for following distribution

Score	70-80	60-70	50-60	40-50	30-40	30-20	10-20	0-10
Students	11	22	30	35	21	11	6	5

Use of skewness: Skewness is useful in the following aspects.

- The coefficient of skewness helps in finding the nature and degree of concentration towards the higher or the lower values
- The measure of skewness reveals the extent of emperical relationship between mean, median, and mode.
- It is very useful to find the extent of normality of given distribution (Many statistical errors are based on the assumption of normality so we need to check normality of data, eg. in errors of means).

Test of skewness: A distribution is said to be skewed if:

- The values of mean, median, and mode do not coincide at a single point.
- The frequency curve of a distribution is not symmetric. That is, the folding of curve through mean line do not make two equal halves.
- The first quartile (Q_1) and third quartile (Q_3) are not in equidistance from median.
- Frequencies on either side of the mode are not equal.

3.4.2. Kurtosis

Besides measure of central tendency, dispersion, and skewness, there is another characteristics of frequency distribution to which we termed as kurtosis. The literal meaning of kurtosis refers to the peakedness or flatness of a frequency distribution as compared to the normal distribution. In other words, kurtosis tells us to have an idea about the shape and nature of middle part of a symmetrical distributions as compared to normal distribution. It shows the degree of convexity of a frequency curve in comparison to the normal curve as a standard.

Karl Pearson introduced three types of Kurtic curve

I. **Laptokurtic**: A curve which is more peaked than the normal curve is called laptokurtic. It is said to have negative kurtosis or said to lack kurtosis.

ii. **Mesokurtic:** An intermediate curve which is neither flat nor peaked is called mesokurtic. It is resemble like normal curve.

iii. **Platykurtic:** A curve which is more flat topped than normal is called platykurtic. It is said to have positive kurtosis or said to possess kurtosis in excess.

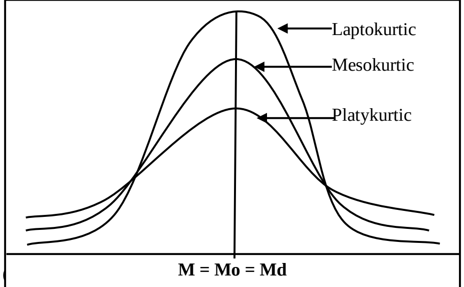


Figure 3.4: Curves showing different Kurtic distribution

Measures of kurtosis: The relative measure and percentiles is the ratio of quartile deviation wi tenth percentiles. Then,

$$\text{Kurtosis } (Ku) = \frac{\text{Quartile Deviation}}{\text{90th Percentile - 10th percentile}}$$

$$= \frac{Q_3 - Q_1}{2\,(P_{90} - P_{10})} \qquad (\because \text{ Q.D.} = \frac{Q_3 - Q_1}{2})\text{.............. (3.10)}$$

In which, Q_3 = Third Quartile

Q_1 = First quartile

P_{90} = Ninetieth percentile

P_{10} = Tenth percentile

From above formula, the value of kurtosis for normal curve is equal to 0.263. If $Ku > 0.263$, the distribution is leptockurtic. If $Ku < 0.263$, the distribution is platykurtic.

Note 3.3: A test scores distribution may be kurtic and skewed due to following reasons.

- When the test items to measure traits are too easier or too difficult, the distribution deviates from normality. For difficult items, the distribution is positively skewed but for easy items the skewness is negative.
- Sampling error may leads abnormality in distribution. When the sampling items are few the distribution is erroneous.
- Administration factors such as cheating, wrong scoring, and inclusion of pre-informed items in test are other factors that deviate normality of any distribution.

Use of kurtosis

- The measure of kurtosis helps (among other things) in the choice of an average. For example, mean is preferable for normal distribution, median is suitable for leptokustic distribution, and the quartile deviation is more appropriate for platykurtic distribution.
- It can be used in describing the characteristics of frequency distribution. Two distribution may have the same average, dispersion, and skewness but there may be difference in concentration of values near mode showing sharper or flatter peak in the frequency curve.
- It is equally helpful in determining the nature of distribution in middle part of the symmetric curve.

3.5 Application of Normal Curve

Normal curve has wide application in the field of education and social sciences. The main applications are discussed below:

1. *To compare scores on different tests*

If we have data selected to the scores of students performance on two or more different tests and we wish to compare these data, we need to convert these scores into comparable standard score to which we call sigma score or z-score. This can be obtained by the following formula.

$$\text{z - score} = \frac{\text{Raw score - Mean}}{\text{Standard deviation}} = \frac{X - \overline{X}}{\sigma}$$

It clearly indicates that how many standard deviation unit a raw is above or below the mean (0). Therefore, it provides a standard scale for the purpose of valuable comparison. This can be illustrated by the following example:

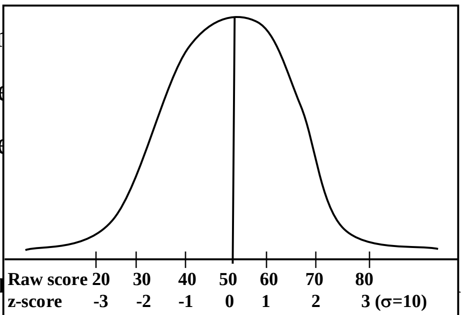


Fig. 3.5: Normal curve with z-score equivalent to raw-score

Example 3.1: A student in M.Ed. obtained 40 m .520) and 30 marks in Measurement & Evaluation (Ed.521). If the mean and standard deviation for the socres in Ed.520 are 35 and 10 and for the scores in Ed.521 are 22 and 8 respectively. Conclude the better achievement of the student.

Solution: Since the measurement scales are not same, we cannot compare the scores in two subjects directly. So convert these scores in standard score form. For this, we have following information on.

	Ed.520	Ed.521
Raw score -	40	30
Mean -	35	22
Standard Deviation-	10	8

Now, *z*-score in Ed.520 $= \frac{40\text{-}35}{10} = \frac{5}{10} = 0.5$

z-score in Ed.521 $= \frac{30\text{-}22}{8} = \frac{8}{8} = 1$

Therefore, the student did better in Measurement & Evaluation then in Research Methodology though he attained higher score in Ed.520.

2. *To ascertain the percentage of individuals whose scores lie between two scores*. Using standard score we can find the percentage of students that falls between two given number in a specified sample.

Example 3.2: Suppose the mean and standard deviation of 500 sample cases are 15 and 3 respectively. If the distribution is assumed to be normal then how many cases fall between 14 and 16 ?

Solution: Convert raw scores 14 and 16 into z-scores form

z-score for raw score 14 = $\frac{14\text{-}15}{3} = -\frac{1}{3} = -0.33\sigma$

z-score for raw score 16 = $\frac{16\text{-}15}{3} = \frac{1}{3} = 0.33\sigma$

Now, see the normal curve table in Appendix A. We see that 1293 cases out of 10000 or 12.93% cases fall between mean and -0.33σ. Similarly, 12.93% cases fall between mean and 0.33σ. Therefore, total number of cases between 14 and 16 = 12.963 + 12.93 = 25.86% or 2586 cases falls out of 500.

Note 3.3: To find the cases in normal curve table, first see the whole number of *z*-value in first column and go horizontally right (on the same row) until required value is obtained for fractional *z*-score. The obtained value is required cases that fall between mean and standard score equivalent to given raw score.

3. ***To find the percentage of individuals who score above a certain point.***

Example 3.3: In a sample of 1000 students in M.Ed. first year, the mean of score in Foundations of Education (Ed.512) is 55 and standard deviation is 5. Assuming normality of distribution find how many students secure first division score (above 60) in Ed.512 ?

Solution: Since we have to find the number of students who secured first division mark in Ed.512 then find standard score for 60 = $\frac{60\text{-}55}{5} = \frac{5}{5} = 1\sigma$ from normal curve table, we see that 3413 out of 1000 or 34.13% cases lie between mean and 1σ. Since 50% cases lie in either side of mean, the cases above 1σ = 50 - 34.13 = 15.87% or 158.7 students out of 1000 were secured first division in Ed.512.

4. ***To find the percentage of cases lying below a given score point (percentile rank).***

Example 3.4: Suppose the mean = 70 and standard deviation = 15 is given for normal distribution of 1000 students. Find the percentile rank of *student scoring 85 and the total number of students whose scores lie* below the score point 35.

Solution: Since the percentile rank is a position of an individual on a scale of 100, then we have to determine the percentage of cases lying below the score point 85. For this, transform the raw score into z-score using formula.

$$\text{z-score} = \frac{X - \overline{X}}{\sigma} = \frac{85 - 70}{15} = \frac{15}{15} = 1\sigma$$

Now from normal curve table, 3513 or 34.13% cases lie between mean and 1σ. Again, to determine the total percentage of cases lying below 1σ = (50+34.13)% = 84.13% of the individuals whose scores lie below the score point 85. Therefore, the percentile rank of the individual whose score 85 is 84.

$$\text{Now, z-score for 35} = \frac{35\text{-}70}{15} = -\frac{35}{15} = -2.33\sigma$$

From normal curve table, 4901 cases or 49.01% cases out of 10,000 lie between mean and -2.33σ. Therefore, 50-49.01 = 0.99% cases lie below 35. That is, $\frac{0.99 \times 1000}{100}$ = 9.9 = 10 individual lie below the score 35.

5. *To find the limits of scores for a given percentage of cases.*

Example 3.5: For a normal distribution with $\overline{X}$ = 100 and σ = 10. Find out the units between which the middle 45 percent of the cases lie.

Solution: It is clear that the middle 45% cases are distributed in such a way that 22.5% cases lie to the left and remaining 22.5% to the right of the mean.

Since 22.5% or 2250 cases lie on the left of mean, from normal curve table we see that 2250 cases lie between mean and -0.59σ. Similarly, remaining 2250 cases lie between mean and right 0.59σ. Therefore, the middle 45% cases fall between the mean and standard score of ±0.59σ.

Now, convert z-score into raw score, we have

$$z = \frac{X_1 - 100}{10}, \qquad z = \frac{X_2 - 100}{10}$$

$$\text{or, } -0.59 = \frac{X_1 - 100}{10} \qquad 0.59 = \frac{X_2 - 100}{10}$$

or, $X_1 = 10\text{-}5.9$ or, $X_2 = 100 + 0.59$

$X_1 = 94\%$ or, $X_2 = 105.9$

Now, the 45% cases fall between point 94.1 and 105.9.

6. *To find out the limits in terms of scores which include the highest given percentage of cases.*

Example 3.6: On a normal distribution with a mean 120 and standard deviation is 20, what limits will includes the highest 10% of the distribution.

Solution: We know that 50% of the cases of norma distribution lie in the right half of the mean. The lower limit of highest 10% is the upper limit of 40% between mean and highest 10%. Therefore, from normal curve table in Appendix A , out of 1000 cases 40% or 4000 cases lie between mean and 1.28σ.

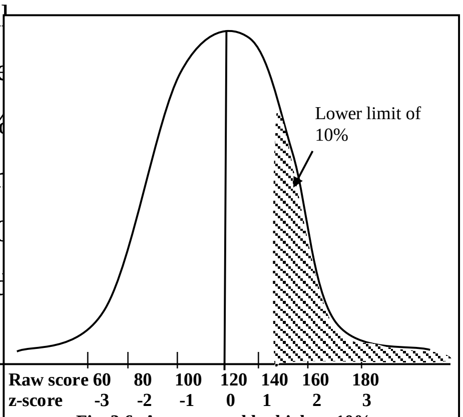


Fig. 3.6: Area covered by highest 10%

Therefore, lower limit of the highest 10% of the c 20 =145.6 and upper limit of highest 10% of the cases will be highest score in the distribution.

7. *To determine the percentile points (or the limits in terms of scores which includes the lowest given percentage of the cases)*

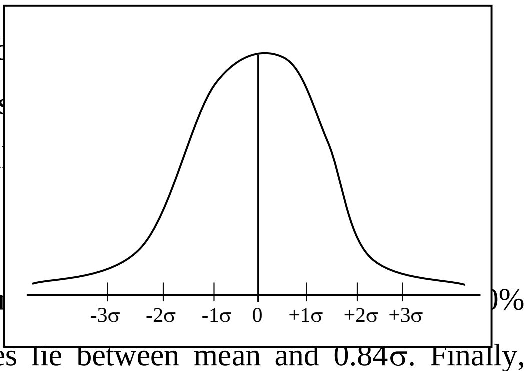
Example 3.7: The mean = 80, σ = 16 of normal distribution of 1000 cases are given Find the limit of 30th percentile (P_{30}).

Solution: To determine P_{30}, we need to look a score point, on the scale of measurement below which 30% of the cases lie. Since 50% cases lie on the left side of mean, the upper limit of the 30% is lower limit of the 20% which lies between 30% and mean. Therefore, out of 1000 cases 20% or 2000 cases lie between mean and -0.525σ (From table). Therefore, the upper limit for P_{30}, $\overline{X}$ - 0.525σ = 80 - 0.525 ×16 =71.6 = 72 and the lower limit will be the lowest score of the distribution.

8. *To find the relative difficulty of test questions, problems, and other test items.*

Example 3.8: Given a test question (A) solved by 10% of a large unrelated group, a second problem (B) solved by 20% of the same group, and third problem (C) solved by 30%. If we assume the capacity measured by the test problems to be distributed normally. What is the relative difficulty of questions A, B, and C ?

Solution: Since 10% of the large group solved question A, 40% of the cases lie between its lower limit and mean. Therefore, from normal curve table the lower limit of 10% is 1.28σ.



Similarly, 20% of the group who passed questio)% above mean. Therefore, 29.95% or 2995 cases lie between mean and 0.84σ. Finally, question C solved by 30% lies above the upper limit of 20% who falls between mean and lower limit of 30%. Therefore, the lower limit of 30% who solved question C is 0.52σ (From table, upper limit of 2000 cases).

Now compare difficulty values (lower limits) for all question we get following result

Questions	Solved by (%)	Difficulty (σ value)	Relative difficulty (σ difference)
A	10	1.28	-
B	20	0.84	0.44

C	30	0.52	0.32

Since the percentage difference between A and B, and B and C are equal but the σ difference in difficulty between A and B is greater than the σ difference between questions B and C. This shows that the difficulty difference between A and B is higher than the difficulty difference between B and C.

9. ***To separate a given group into subgroups according to capacity, when the trait is normally distributed.***

Example 3.9: Suppose that we have administrated, an entrance examination of 200 college students. We wish to classify the group into 5 subgroups: A, B, C, D, and E according to ability, the range of ability to be equal in each subgroup. On the assumption that the trait measured by our examination is normally distributed, how many students should be placed in groups A, B, C, D and E ?

Solution: Since the ability under measurement is normally distributed and most of the a (99.74%) falls

between ± 3σ we may divide this range equa into five subgroups.

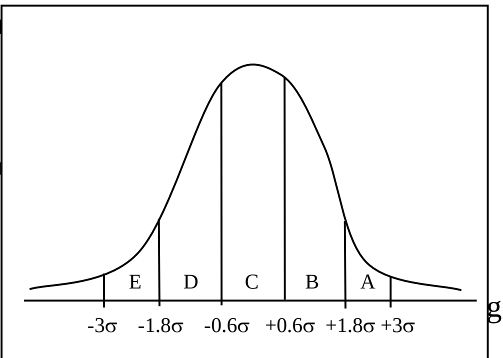


Figure 3.8: Division of subgroups

This division gives 1.2σ as the portion for ge of the each subgroups can be seen clearly from o 3σ, Group B:0.6σ to 1.8σ, Group C: -0.6σ to + 0.6σ, Group D: -1.8σ to -0.6σ, and Group E: -3σ to -1.8σ).

Since Group A extends from +1.8σ to +3σ,

Number of cases between mean and +3σ = 4986

Number of cases between mean and +1.8σ = 4641

Number between +1.8σ and + 3σ = 4986-4641=345

That is, 3.45% out of 10000 falls in Group A.

Now, 3.45% of 200 student = $\frac{3.45\times200}{100}$ = 69 = 7.

In the same way the Group B extends from 0.6σ to 1.8σ.

Number of cases within this groups = 4641 - 2257 = 2384 or 23.84% out of 10000. Now, 23.84% of 200 students = 47.68

The Group C extends from -0.6σ to $\pm\ 0.6\sigma$

Number of cases within this group = 2257 + 2257 = 4514 or 45.14% out of 10000

Now 45.14% of 200 = 90.2.

The group D extends from -1.8σ το $-0.6\sigma\sigma$ and the distribution is normal, Number of cases in this group equals number of cases in Group B.

Similarly, the number of cases in Group E equals, the number of cases in Group A.

Now, we can summarize following results from above computation.

Particular	Groups				
	A	B	C	D	E
Limits	-3σ to -1.8σ	-1.8σ to -0.6σ	-0.6σ to 0.6σ	0.6σ to 1.8σ	1.8σ to 3σ
Case in percentage	3.45	23.84	45.14	23.84	3.45
Number of students (in 200)	6.9	47.68	90.2	47.68	6.9
Students in whole number	7	48	90	48	7

Therefore, the number of students that can be placed in subgroups A, B, C, D, and E are 7, 48, 90, 48 and 7 respectively.

Exercise

Multiple choice questions

1. Area under normal curve covered by mean and +1σ is

   (a) 34.13% (b) 68.26% (c) 45.22% (d) 95.44%

2. The normal curve has its maximum height at

(a) the median (b) the mode

(c) the mean (d) the point of inflection

Short answer questions:

1. Define kurtosis and describe its concept with suitable example.
2. Describe the properties of normal probability curve.

Long answer questions:

1. Define normal probability curve and explain its applications in education with suitable examples.
2. Generally, a normal curve is symmetric about its measure of central tendencies. If the curve is deviate from normality then how can you measure the extent of divergence ? Explain.

Chapter 4

Measure of Relationship

4.0 About this Chapter

This chapter provides the knowledge of different types of relationship useful to describe the data obtained in diverse educational situation. The first section clarifies the concept of correlation between two sets of data linear in nature. Second section describes correlation as a measure of relationship/association. Third section presents the ways of describing relationship between two sets of data using scatter diagram. Fourth section gives the method of computing correlation coefficient between two sets of linear data using different forms of formula given by Karl Pearson. Fifth section provides the method of computing linear relationship between the ranks of scores in two sets using the formula developed by Charles Spearman. The sixth section describes the biserial method of computing association between two variables in which first variable is continuous and quantitative and second is normal and artificially dichotomized. The seventh section provides the way of computing point biserial correlation coefficient between two variables in which first is continuous and quantitative and second is qualitative and naturally dichotomized. The eighth section presents the method of phi-correlation to find the relation between two variables in which both are qualitative and dichotomized. The ninth section states the method of tetrachoric correlation to find the relation between two variables in which both are artificially dichotomized and both cannot be expressed in score. The tenth section explains the method of finding the partial correlation between two variables by nullifying the effect of third variable upon both correlating variables. The final section further provides the method of finding the multiple correlation between a dependent and a number of independent variables.

On completion of this chapter, the following specific objectives are assumed to be achieved.

- To clarify the concept of correlation as a measure of relation.
- To describe the relationship between variables using scatter diagram.
- To be able to use different methods of computing correlation coefficient between two variables with different natures.
- To apply the method of correlation to interpret the given data.

4.1 Concept of Correlation

We encounter with many situations and problems in our daily life. As a professional in different field we may face many problems associated with our working field. For instance, a farmer may need to seek the relation between wheat production and use of certain fertilizer. A merchant wants to know the relation between increase in price change and the amount of selling. Similarly, a teacher may face the problems associated with student and their performance that seek the relation between two or more variables. For, does the increase in study time influence in achievement? Does TV watching influence in performance ? Is there any type of gender influence in attendance in a school?, and so on. These and other similar problems can be tackled with some certain statistical procedure called *method of correlation*.

Historically, the method of finding the relationship between the variables was started with the effort to test the assumptions in evolution theory developed by Charles Darwin in 1859. He tried to test the variation in traits among individuals. To verify the theory developed by him, it become important to study traits in which organisms of the same species differ and to determine how those traits are influenced by heredity. These problems catched the attention of Sir Francis Galton, Darwin's Cousin, and he became interested to study the individual differences. From his laboratory techniques of assessing correlation and making prediction eventually emerged (Menium, King, & Bear, 1970). In the field of education and psychology, therefore, there are many situations that seek the relation between variables.

The relationship between variables can be established for different purposes: To measure the degree of relationship and to find the extent of agreement between the ranks obtained by different individuals. Therefore, these will be discussed below in detail.

4.2 Correlation as a Measure of Relationship

In many educational situations it is important to examine the relationship of one variable to another or to find the relative dependency between variables. For example , is it true that bright children tend to less neurotic than average children ? Measuring the general intelligence of child using standard test, can we say anything about his/her probable scholastic achievement as represented by grades ? Estimating the reliability coefficient between pretest score and score obtained in class test, can we say anything about student future performance. Problems like these, which involve relation among abilities, can be studied by the method of correlation. Therefore, it can be defined as the degree of relationship between two or more set of data representing different traits. The concept of correlation as a measure of relationship can further be clarified using some examples below.

Correlation as a relationship: As we know, the circumstance (c) of the circle can be obtained by the formula, $c = \pi d$ where d is diameter of circle, and $\pi = 3.1416$. If we take different values of d, we will get corresponding values for circumference. That is, the value of circumference is 3.11415 times for each particular value of diameter. This case is true for anywhere in the world. Therefore, the relation between diameter and circumstances of circle is absolute or perfect and is denoted by $r = 1$.

Similarly, in measuring the stability of particular traits among the students of a certain class, we may express the relation. Suppose that twenty students in a class have exactly the same position in first term exam and second term exam, i.e. the student who scores first in the first term test also scores first in second term, the student who scores second in the first term also scores second in the second term and so on. The one-to-one correspondence in two tests signify perfect relationship since the position in first test match exactly the same position in second test. In this case also the relationship is perfectly positive and denoted by $r = 1$.

But if the first student in first test scores last position in second test, second student in second test secures second last in second test and so on. In this case also the relationship is perfect but in negative direction and we denote $r = -1$. Similarly, if the student's position is first in first test and score second in second test, the relationship decreases gradually as the extent of matching decreases. This may indicates positive or negative relationship to some extent. In the same way, if the student's position in first test does not match the position in second test, then there would be no relation at all. In this case we say the relation is zero and indicates $r = 0$. Therefore, the value of correlation ranges from -1 through 0 to 1.

According to Sancheti and Kapoor (1980, p.8.9) the degree of correlation can be described as in the box below:

Degree	Positive	Negative
Perfect correlation	+1	-1
Very high	+0.9≤	- 0.9≤
Fairly high	+0.75 to 0.9	-0.75 to -0.9

In our daily life we can find various such situations in which the scores or traits of first variable can show the relation of the scores or traits of second variable in the extent or degree to which these variables are correlated. Therefore, correlation can be taken as a relationship between variables. Similarly, the extent of degree of correlation enables us to predict the traits in second variables. The greater the association between two variables, the more accurately we can predict the condition or related traits in other.

The degree of correlation and prediction have uses in many areas of psychology, education, sociology, economics, science and management. In education, it can be used to establish the reliability of a test (i.e., the consistency of scores over separated administration, the equivalency of items, the inter-item consistency, and consistency of scoring of different scorer), and to provide validity evidence. In research field, it can be calculated to find the relation of one variable upon other. In medical field a doctor or medical researcher can observe the relation of certain medicine upon particular disease. A farmer can see the impact of certain fertilizer on the amount of wheat production. An investor can observe the effect of advertisement upon the sales amount of produced goods, and so on.

4.3 Scattered Diagram

As stated above, the relationship between variables can be obtained by plotting the value of one variable in *X*-axis against the value of other variable in *Y*-axis. When the number of items (*N*) are large, the method of calculating correlation using direct formula is tedious and too time consuming. In order to avoid this difficulty, we may use diagrammatic method of finding the relationship between variables instead of direct statistical calculation. The diagram or chart thus formed is called *scatter diagram or scattergram*. It represents the joint distribution of two variables so it is also known as *bivariate distribution*. Consider, hypothetical data that the scores obtained by 10 high school student in mathematics and science are given as below.

Table 4.1: Bivariate data obtained by ten high school students

Student	A	B	C	D	E	F	G	H	I	J
Math (*X*)	50	48	46	40	38	37	28	25	20	18
Science (*Y*)	45	44	43	38	37	36	28	25	18	15

The above data can be expressed in the scatter diagram.

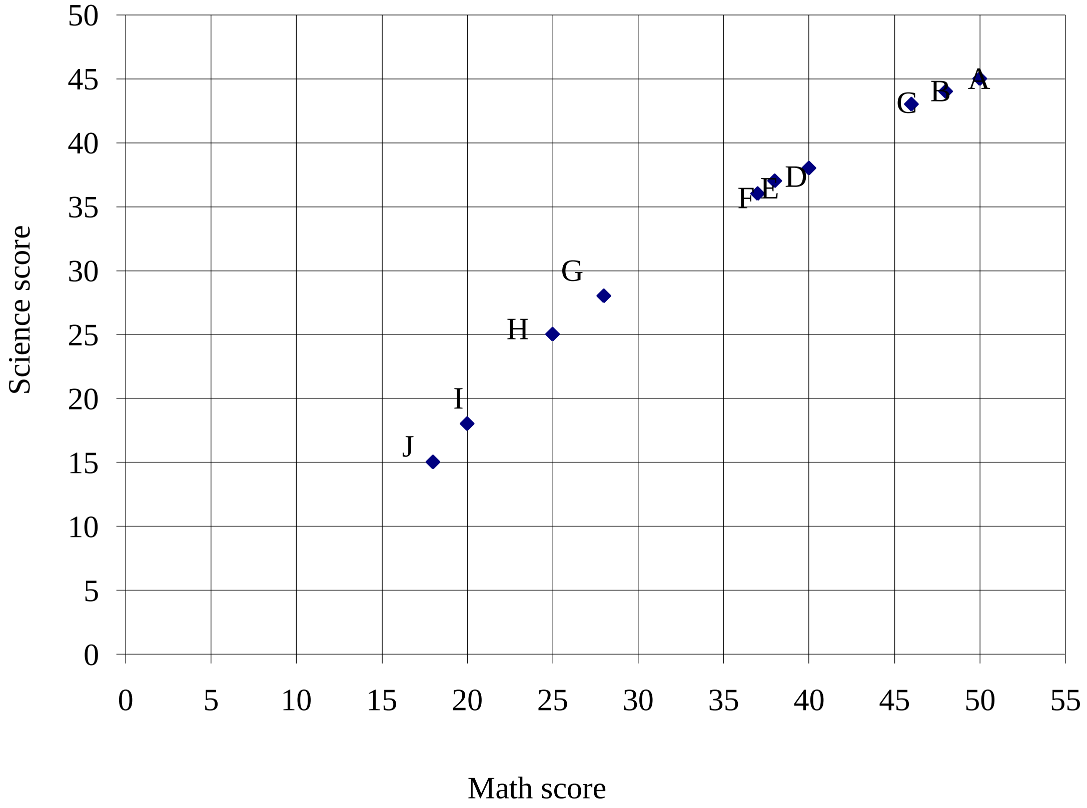


Figure 4.1: Scatter diagram showing positive relation between Mathematics and Science score obtained by ten school students.

The diagram above allows us to easily see the nature of the relationship between the scores in two subjects obtained by the group of same student. That is the student who scores high in mathematics also scores high in science. The one-to-one correspondence between two set of score exhibit the liner relationship between these scores. This means that the scores obtained in math and science are highly correlated. Such relationship is known as positive correlation. If the distribution of patches on the diagram is scattered from left below to right above trend in first quadrant then we can say the relation is positive. As the distribution of patches approaches to linearity the value of correlation coefficient goes near to 1.

In the same way, if we prepared the scatter plot, we will get different diagrams according as the degree or extent of relationship between variables. The scatter plot above shows almost linear relation so we can say these two scores are highly related. If the scores obtained in science are in reverse order, then we will get the scatter diagram as in Figure 4.2 below.

Table 4.2: Scores obtained by ten students in mathematics and science

Student	A	B	C	D	E	F	G	H	I	J
Math (*X*)	50	48	46	40	38	37	28	25	20	18
Science (*Y*)	15	18	25	28	36	37	38	43	44	45

The above data can be expressed in the scatter diagram that shows the different relation than previous one.

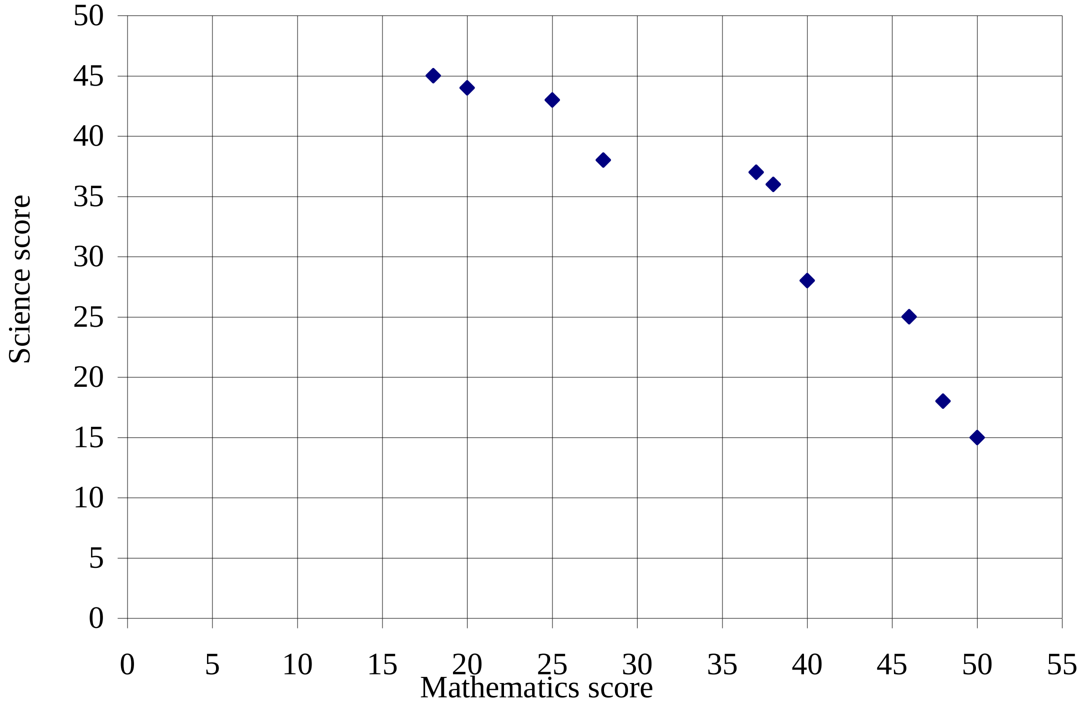


Figure 4.2: Scatter diagram showing negative relation between scores

The distribution points in the above figure is showing certain trend. That is the points are scattered from the right below to left above pattern. Such pattern is known as negative pattern and the relation is called negative correlation ship. The scatter diagram above shows almost linear relation but in negative direction so it is an example of negative correlation.

Again, if the increase of scores in one set does not show any clear direction in another set, then it gives another type of scatter diagram without any pattern to which we call zero relation. For, consider the following data.

***Table 4.3*:** Scores obtained by students in two subjects that show near zero relation

Student	A	B	C	D	E	F	G	H	I	J
Math (*X*)	50	48	46	40	38	37	28	25	20	18
Science (*Y*)	15	45	37	28	37	22	40	16	16	22

The above data can be expressed in the scatter diagram.

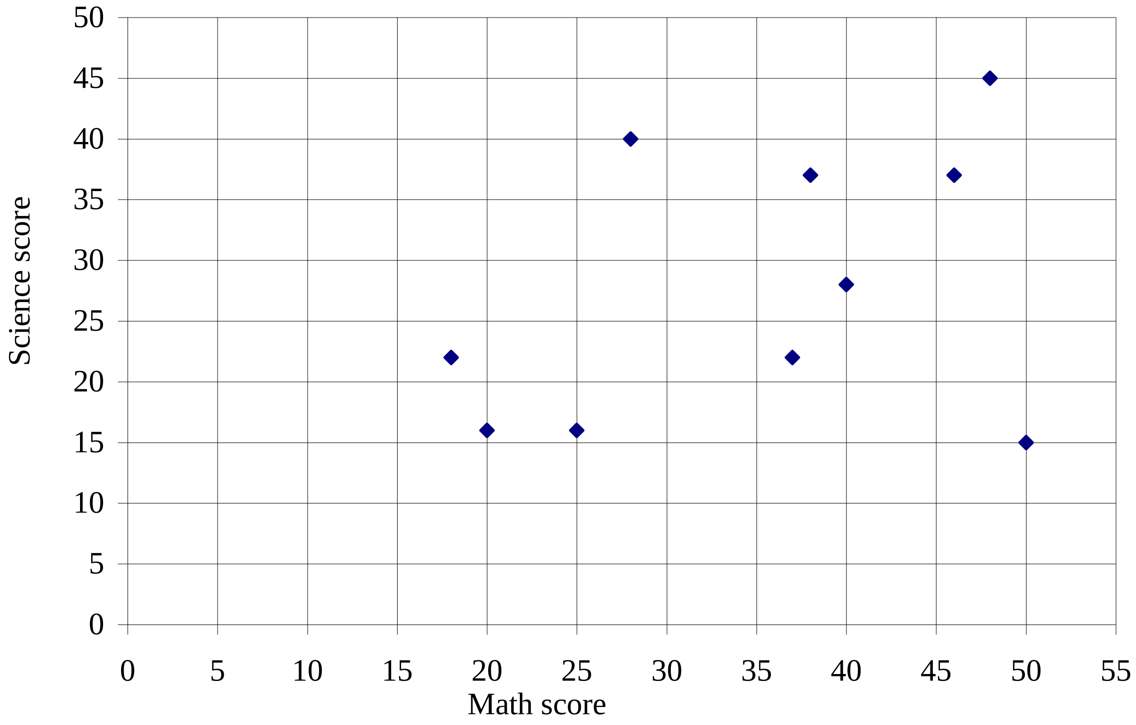


***Figure 4.3*:** Scatter diagram showing poor/zero relationship

Note: If the number of variables / items/scores are large the above procedure of drawing scatter diagram is the tedious and too time consuming. To overcome this difficulty we can draw the scatter diagram keeping one variable in an interval form in *Y*-axis against another variable in frequency form in *X*-axis.

The **steps** in constructing a scatter diagram are as given below.

1. Design one variable as *X* and the other as *Y* . If you have to do prediction then generally choose predictor as *X* variable and another as *Y* variable.
2. Draw *X* axis and *Y* axis having equal length and indicate the range the value of *X* in *X* axis (horizontal line) and value of *Y* in *Y* axis (vertical line). If the numbers are large then make the range /intervals of variables and indicate the interval of one variable in *X*-axis and another variable in *Y*-axis.
3. For each value of *X* in *X*-axis, find the corresponding value of *Y* in *Y*-axis and indicate it with dot or star. For the sake of convenience you may name these points A, B, C, etc as indicated in the table.
4. Name each axis and give the name of entire graph.
5. Analyze the nature of relationship in the graph. If the straight line joining each plotted points makes 45^0 from origin with *X*-axis then the relationship is perfectly positive. If the line makes 45^0 with *X*-axis so that it makes equal intercept with *X*-axis and *Y*-axis then the relation is perfectly negative. If it gives zig zag line towards positive direction, the degree of relation degreases with the increase of zig zag form. If there in no relation, you will get different patches of intersecting points without any pattern (haphazard distribution in scatter diagram).

Different methods to measure association

1. ***Karl Pearson’s correlation (r)*****:** When we have to find the relationship between variables but not specified a particular method of finding the relationship, the Pearson's correlation coefficient is best suited to measure the association between two continuous and quantitative variables. For other circumstances there are alternative measures of association. These are derived directly and indirectly from Pearson's method.

2. ***Rank-order correlation (ρ)*****: T**o measure the relation between two set of scores representing certain variables we may use Pearson's product-moment method. But in some circumstance we need to find the relation between the rank/order roll number of an individual. In such case the Pearson's method is futile and we use new method developed by Spearman to which we call Spearman's rank-order correlation coefficient and is denoted by Greek symbol rho (ρ).

3. ***Biserial correlation* (r_{bis}):** In many instances in education psychology we have to find the association between two variables in which first is *continuous and quantitative* and second is normal and *artificially dichotomized*. In such situation biserial correlation is an estimate of Pearson's *r* and denoted by r_{bis}. Suppose we have to compute biserial *r* between the score obtained on a music appreciation test by a group of student with and without training in music. Here, the artificial dichotomy refers to the arbitrary division without any specific criteria.

4. ***Point-biserial correlation* (r_{pbis}):** If one variable is continuous and quantitative and other is quantitative and naturally dichotomized, then the suitable measure of association is point biserial and is denoted by $\mathrm{r_{phis}}$. Suppose we have to measure the relation between the achievement in mathematics and dichotomy of pass and fail. The division of dichotomy, here, is clear and specific and therefore is called natural or genuine division.

5. ***Phi-correlation* (ϕ):** In some circumstances, we have to find the relation between two variables in which *both are qualitative and dichotomous*. The method designed to measure such relation is known as phi-correlation and is denoted by Greek letter ϕ. Suppose, we have to compute the relation between intellectual status (feeble minded and normal) and marital status (married - unmarried), the phi-coefficient may be appropriate method of finding the association.

6. ***Tetrachoric correlation* (r_t) :** In some situations *both the variables are artificially dichotomized* and *both cannot be expressed in score*, the methods of finding relation using biserial and point biserial are not effective. To find such relation we use a new method of finding the association called tetrachoric correlation and is denoted by r_t. For example, to study the relationship between intelligence and emotional maturity. The first variable 'intelligence' may be dichotomized as above average

and below average and the second variable emotional maturity as emotional mature and emotional immature.

7. ***Partial correlation* ($r_{1.23}$):** In many instances, we find the relation between two variables but the relation between two variables may influenced by the effect of third variable. In such situation we need to nullify the undesired influence of third or any other variable on the relationship of two variables. Therefore, the method of finding the relationship between two variables by nullifying or partial out the effects of third variable upon both variables being correlated is known as partial correlation and is denoted by $r_{1.23}$ where 1, 2, 3 are variables. For, in a study of relation between height and weight of male children in a group may be influenced by the age so with keeping age constant the relation between first two may be more accurate. This is possible by the method of partial correlation.

8. ***Semi-partial correlation*:** The method of removing the influence of one or more additional variables from only one of the two that we are correlating is called semi-partial or part correlation. This method was used by Scarr (1985) to remove the factor or of the mothers' IQ in the relationship between mothers' discipline technique and children's intellectual and social skills.

9. ***Multiple correlation*:** In some cases, we find that a variable is dependent on a number of other variables (called independent variables). For example, if we want to study one's academic achievement, of it may be associated with many variables like intelligence, socio-economic status, education of the parents, the method of teaching, quality of teachers, aptitude, interest, family environment, number of hours spent on studies, and so on. If we want to study inter-correlation between these independent variables and the influence of these variable on single dependent variable (achievement) then we can use an appropriate statistics is called *coefficient of multiple correlation* and is denoted by $R_{1.23}$ where 1, 2, 3 are variables.

10. ***Curvilinear correlation* (η):** In Pearson's r we measure the relation between two variables that are linearly related. But in some situations the relationship may not be linear. Consider the case of finding the relation between age and motor skill. The

first variable - age is not related linearly with second variable - motor skill but the relation is curvilinear because the increase in age will not always increase in motor skill, rather it decreases at certain age. In such case the method of Pearson's *r* underestimate the relation between two quantitative and continuous variables so we use certain method of fining the correlation ratio called curvilinear correlation and is denoted by Greek symbol eta (η).

11. ***The contingency coefficient* (*c*):** The relationship between two variables in which at least one have three or more categories cannot be measured using previous methods of correlation. In such situation we use the contingency method to which we call the contingency coefficient. For example, to study the association between parent's eye colour and offspring's eye colour, this technique would be appropriate.

4.4 Pearson’s Product Moment Correlation

There are various methods of correlation relative to the situation and use of data. In some situations the data for two variables *X* and *Y* are expressed in interval or ratio scale of measurement and the distribution of these variables have a linear relationship. Moreover, the distribution of the variables is unimodal (having single mode) and the variance of the distribution are approximately equal. In such situation we may use the method of finding the relation developed by British Biometrician Karl Pearson (1867-1936).

The method developed by Pearson is also known as product moment method. In science, the word moment refers to the displacement of an object from its centre of gravity. In statistics, therefore, it can be use as the deviation (displacement) of scores form the mean (centre of distribution) of these scores. Therefore, the Pearson's product moment correlation *r* is calculated by taking the products of the paired moments.

We can use various form of Pearson's formula in relation to different situations. When the size of the given data is small or calculating machine is available, we can calculate Pearson's *r* using original data with the help of following raw score formula:

$$r_{xy} = \frac{N\Sigma XY \text{ - } \Sigma X.\Sigma Y.}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]}\sqrt{[N\Sigma Y^2 - (\Sigma Y)^2]}} \text{ .... (4.1)}$$

In which, X = Raw scores in first variable.

Y = Raw scores in second variable.

N = Total number of items or scores in.

But when the number of items and size of data in given variables are large it is tedious and too time consuming to calculate the correlation coefficient using above formula. Therefore, in order to avoid this difficulty, we may use following modified formula which is obtained by reducing the values of raw score from assumed means.

$$r_{xy} = \frac{N\Sigma xy \text{ - } \Sigma x.\Sigma y.}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2]}\sqrt{[N\Sigma y^2 - (\Sigma y)^2]}} \text{ ....(4.2)}$$

In which, x = Deviation of X scores from assumed mean

y = Deviation of Y scores from assumed mean

N = Number of cases or items

But, if standard scores of each distribution are given then we calculate r using z-score formula for correlation coefficient.

$$r_{xy} = \frac{\Sigma(z_x.z_y)}{N} \text{ ....(4.3)}$$

In which, z_x = Standard score corresponding to X-variable

z_y = Standard score corresponding to Y-variable

σ_x = Standard deviation of scores in X-column

σ_y = Standard deviation of scores in Y-column

N = Number of cases or items

By substituting the values of z_x, z_y, σ_x, and σ_y we can reduce the formula (4.3) to new form (4.4)

$$r_{xy} = \frac{\Sigma\left(\frac{X-\overline{X}}{\sigma_x}.\frac{Y-\overline{Y}}{\sigma_y}\right)}{N} = \frac{\Sigma\,(X-\overline{X})(Y-\overline{Y})}{N\,\sigma_x\sigma_y} = \frac{\Sigma\,xy}{N\sqrt{\frac{\Sigma x^2}{N}}.\sqrt{\frac{\Sigma y^2}{N}}}$$

$$\therefore\ r_{xy} = \frac{\Sigma xy}{\sqrt{\Sigma x^2.\Sigma y^2}} \text{...(4.4)}$$

Steps to Calculate Pearson's *r*

1. List the pairs of scores in two columns *X* and *Y*.
2. Find the mean for scores in *X*-column and mean for scores in *Y*-column.
3. Calculate *r* using formula 4.1. For, find sum of scores in each column, sum of product of corresponding scores in two columns, sum of squares of scores in *X*-column, sum of squares of scores in *Y*-column and substitute these values in formula (4.1).

   Again, to calculate *r* using 4.2, reduce each individual score from assumed mean of corresponding column, find required components, and substitute in equation 4.2. To calculate using standard score, find deviation score from actual mean in each column, calculate standard deviation for each column, divide each individual scores of corresponding column by standard deviation, obtain the sum of products of corresponding standard scores from two columns, and finally put the value in formula 4.2.

   Similarly, to find *r* using reduced formula- 4.4, find deviation of individual scores from actual mean, sum of product of deviation scores in two column, sum of squares of deviation of scores in first column, sum of squares of deviation of scores in second column, and put these values in equation 4.4.

4. Interpretate the computed correlation coefficient using different criteria. The criteria presented by Sancheti and Kapoor (1983, p. 89) is as given in the following table:

Table 4.4: Degrees of correlation ship and their interpretation

Degree	Positive	Negative	Interpretation
Absence or no	0	0	No relation
Very little	0 < to + 0.25	- 0.25 + < 0	Negligible
Low degree	+ 0.25 to + 0.50	- 0.50 to - 0.25	Present but slight
Moderate	+ 0.50 to + 0.75	- 0.75 to - 0.50	Substantial but small
Fairly high	+ 0.75 to + 0.90	- 0.90 to - 0.75	Marked relation
Very high	+ 0.90 to < +1	- 1 to < - 0.90	Quite dependable
Perfect	+ 1	- 1	Perfect relation

The interpretation criteria for numerical value of *r* is arbitrary. Mangal (2005, p. 105) and Garrett (2008, p. 175) state the range of criteria that differs with the range above. Therefore, the criteria may be different with different writers, purposes, and situations. Further, the interpretation rely on the aim of the testing. For example, if the reliability coefficient obtained by test retest method is used to describe the intelligence of a group or the coefficient is used to take final decision or the coefficient is used to take irreversible decision then we will require *r* as high as 0.90 and above (Mangal, 2005, p. 106). But in the case of using *r* for predicting validity of a test or for improvement purpose or for reversible decision, we will not require the value of correlation coefficient as much high. More specifically the coefficient of correlation must always be judged with regard to: the nature of variable, significance of coefficient, variability of the group, reliability coefficient of tests used and the purpose for which *r* was computed (Garrett, 2008, p. 170).

Example 4.1: Find correlation coefficient (*r*) between the scores obtained in two different measures by four subjects using raw score method.

Subject	M	N	O	P
Score on *X*	7	6	5	8
Score on *Y*	9	6	7	3

Solution: To find the correlation coefficient using raw score method, we need to compute the components in raw score formula.

Table 4.5: Calculation of *r* using raw-score method

Subject	Scores on		*XY*	X^2	Y^2
	X	*Y*			
M	7	9	63	49	81
N	6	6	36	36	36
O	5	7	35	25	49
P	8	3	24	64	9
$N = 4$	$\Sigma X = 26$	$\Sigma Y = 25$	$\Sigma XY = 158$	$\Sigma X^2 = 174$	$\Sigma Y^2 = 175$

Now, substituting these values in raw-score formula for *r* then

$$r = \frac{N\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]}\sqrt{[N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{4\times 158 - 26\times 25}{\sqrt{[4\times 174 - (26)^2]}\sqrt{[4\times 175 - (25)^2]}}$$

$$= -0.464$$

Conclusion: The value of *r* = - 0.464 indicates negative and low degree of correlation between two measures, i.e. in opposite direction. This means that the first measure and the second measure upon the same subject show slightly opposite result. The scoring tendency of two scorer do not match at all.

Example 4.2: Suppose that a person's scores on personality test designed to measure sociability (*X*) and the rating of that person's sociability made by a close friend (*Y*) are given:

X	30	34	35	36	39	39	40	40	42	46
Y	72	70	76	80	73	79	76	83	85	81

What the result says about validity of the personality test?

Solution: To test the validity of personality test we need to find the dependency of scores obtained in personality test with the rating on sociability made by close friend. For this we calculate the correlation coefficient using reduced formula (to avoid the difficulty of calculation)

Table 4.6: Calculation of *r* using deviation method

X	*Y*	*x* = *X*-A	*y* = *Y*-A	x^2	y^2	*xy*
30	72	-8	-5	64	25	40
34	70	-4	-7	16	49	28
35	76	-3	-1	9	1	3
36	80	-2	3	4	9	-6
39	73	1	-4	1	16	-4
39	79	1	2	1	4	2
40	76	2	-1	4	1	-2
40	83	2	6	4	36	12
42	85	4	8	16	64	32
46	81	8	4	64	16	32
Σ*X*=381	Σ*Y*=775	Σ*X* = 1	Σ*y* = -5	Σx^2=183	Σy^2 =221	Σ*xy* = 137

Now, on substituting the values from table in deviation formula, we get

$$r = \frac{N\Sigma XY - \Sigma X.\Sigma Y}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]}\sqrt{[N\Sigma Y^2 - (\Sigma Y)^2]}}$$

$$= \frac{10\times 137 - 1\times(-5)}{\sqrt{[10\times 183 - (1)^2]}\sqrt{[10\times 221 - (-5)^2]}}$$

$$= -0.68$$

Alternatively, using reduced formula,

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2.\Sigma y^2}} = \frac{137}{\sqrt{183\times 221}} = 0.68$$

Exercise 4.1

1. Seven students secured the following scores on two quizzes, *X* and *Y*:

Students	A	B	C	D	E	F	G
Score on *X*	3	9	7	8	4	6	7
Score on *Y*	2	7	8	6	6	5	7

a) Construct scatter diagram of the relationship between these two variables. What is the direction of relationship ?

b) Compute *r* to two decimals using raw score method and the deviation score method. Which method is easier ? Why ?

c) If the two quizzes cover the same subject matter, what is it about students' performance that *r* measures ?

2. Find the correlation between the two sets of memory-span scores when deviations are taken from assumed mean

Test	Scores														
Digit span	15	14	13	12	11	11	11	10	10	10	9	9	8	7	7
Letter span	12	14	10	8	12	9	12	8	10	9	8	9	7	8	6

3. Ten students have obtained the following scores on tests in History and Nepali. Plot these scores on scatter diagram and interpret the relation.

Individuals	A	B	C	D	E	F	G	H	I	J
History (*X*)	13	12	10	8	7	6	6	4	3	1
Nepali (*Y*)	7	11	3	7	2	12	6	2	9	6

4.5 Spearman's Rank Order Correlation

When we have to find the relation between two set of scores, we generally use Pearson's *r*. In many situations in education and psychology, it is not possible to measure the variable under consideration quantitatively or to ascertain the magnitudes of each items in a given statistical series but it is possible to rank the different aspects of variables. For example, it may be possible for the judges to rank their students for scholarship provision, whereas it may be difficult to assign them a numerical grade. Similarly, when the ranking of students with respect to particular trait is known and their scores in class test are given then the scores in class test can be possible to change in rank order form but it is not possible to change the ranks in score form for correlation purpose. Therefore, in these situations, it is not possible to measure the relationship between the traits by product-moment correlation coefficient. In such case we need to new technique of computing relationship between two set of ranks and that technique was developed by *Charles Edward Spearman* in 1904.

Therefore the method of ascertaining the relationship between two set of ranks is known as rank order correlation or generally Spearman's correlation coefficient and usually denoted by Greek symbol rho (ρ).

More clearly, this method is useful when the data are available only in ordinal form of measurement (possible to present only in ranking) rather than interval or ratio form or if the number of paired variables is fewer than 30 (Koul, 2009, p. 345). This means that the method is suitable for the qualitative data such as honestly, beauty, efficiency, intelligence etc. For example, the students in two classes can be ranked in order of their intelligence and the degree of correlation can be computed between their ranks using rank method. Similarly, when ranks or rank differences between two set of scores representing different traits are available, this technique is only a method to assess the relationship between these traits. But this does not mean that Spearman's ρ cannot be used while raw scores are available when the ranking of first set of scores exactly match with the ranking ao scores in second set then Pearson's r and Spearman's ρ agree exactly.

The following formula can be used to compute rank order coefficient of correlation (ρ).

Case I: When rank is not repeated.

$$\rho = 1 - \frac{6\Sigma D^2}{N\ (N^2 - 1)} \quad \text{...(4.5)}$$

In which, D = The difference between the paired ranks.

N = Number of paired ranks.

Case II: When ranks are repeated

If any two or more individuals are found to be equal in ranking with respect to some trait, then the formula in case I will not work properly. In such case, common ranks should be given to the repeated items. The common rank is the average of the repeated ranks. The rank to the item next to repeated items is serial rank as if it would not be repeated. As a result of the common ranking the formula for ρ has to be corrected adding $\frac{m\ (m^2 - 1)}{12}$

for each repeated value to ΣD^2, where m is the number of times an item has repeated. Hence the above formula for non-repeated case reduce to the following form.

$$\rho = 1 - \frac{6\,[\Sigma D^2 + \frac{m\,(m^2-1)}{12} + \frac{n\,(n^2-1)}{12} + \frac{p\,(p^2-1)}{12} + ...]}{N\,(N \;\; -1)} \quad .... \ (4.6)$$

In which, m = Number of times for first repeated items.

n = Number of times for second repeated items.

N = Number of pairs in ranks.

D = Difference between the paired ranks.

Steps for rank-difference method

1. Examine the given data to know whether ranks or actual values are given. If actual values are given, rank them according to their magnitude.
2. Take the difference of two corresponding ranks and put these difference in D column.
3. Square each difference and sum these to find ΣD^2
4. For non-repeated case, put the value of ΣD^2 and N in formula 4.5, and find ρ. For repeated or tied case, substitute the required values including the number of times for each repeated items and calculate ρ.
5. Interpret the obtained rank order correlation coefficient. As the value of r, the value of rank correlation coefficient also varies between −1 and +1. When the value of ρ is +1, it implies complete agreement in the order of ranks and the ranks will be in the same direction. When ρ = -1, the ranks will be in opposite directions showing complete disagreement in the order of ranks.

Example 4.3 : Ten top student's answer sheet were distributed to two checker to maintain reliability in examination. Decide the extent of their agreement using the ranks assigned by these two examiner.

Students	A	B	C	D	E	F	G	H	I	J
Ranks assigned by *X*	1	5	4	6	3	2	7	9	10	8

Ranks assigned by Y	2	3	5	4	6	1	7	8	9	10

Solution: Since the given ranks are not repeated the first case to compute rho is applicable. Therefore, to find required components for ρ the given ranks are tabulated as in the following table

Table 4.7: Calculation of rank difference correlation (case I)

Students	Ranks assigned by X	Rank assigned by Y	D	D^2
A	1	2	-1	1
B	5	3	2	4
C	4	5	-1	1
D	6	4	2	4
E	3	6	-3	9
F	2	1	1	1
G	7	7	0	0
H	9	8	1	1
I	10	9	1	1
J	8	10	-2	4
$N = 10$			$\Sigma D^2 = 26$	

Now, we have for ρ when scores are non-repeated

$$\rho = 1 - \frac{6\Sigma D^2}{N\,(N^2 - 1)} = 1 - \frac{6 \times 26}{10\,(100 - 1)} = 0.84$$

The scoring style of two examiner agree 84%. That is, the relation between the ranks of two set of scores is found to be high. Because of their high reliability we can use them for consistent scoring purpose.

Example 4.4: Seven students were ranked on the basis of their accuracy in English and Speed test taken on same subject. Calculate the relationship in two tests using obtained data.

Students	A	B	C	D	E	F	G

Order in accuracy test	1	2	3	4	5	6	7
Score in speed test	21	23	30	35	30	25	20

Solution: Since the scores obtained by C and E in speed test are same (30) in second and third position, common ranks are provided to them taking average of two ranks, i.e. (2 +3)/2 = 2.5. Using given data following table can be constructed.

Table 4.8: Calculation of rank difference correlation (case II)

Students	Rank in accuracy (*X*)	Rank in speed (*Y*)	*D*	D^2
A	1	6	-5	25
B	5	5	-3	9
C	4	2.5	0.5	0.25
D	6	1	3	9
E	3	2.5	2.5	6.25
F	2	4	2	4
G	7	7	0	0
$N = 7$			$\Sigma D^2 = 53.5$	

Since the rank is repeated, we must use the formula for repeated case

$$\rho \qquad = 1 - \frac{6\left[\Sigma D^2 + \dfrac{m\,(m^2-1)}{12}\right]}{N\,(N^2-1)}$$

$$= 1 - \frac{6\,[53.5 + \left[\dfrac{2(2^2-1))}{12}\right]}{7\,(7^2-1)} = 0.035$$

There is poor relation between speed test and accuracy test. This means that the students who did better in accuracy test are poor in speed test and vice versa.

Exercise 4.2

1. Two judges at an art show are asked to place the six pictures that reached the 'finals' in order of merit. Their remaining are:

Picture	A	B	C	D	E	F
Judge A	3	2	6	4	1	5
Judge B	2	1	5	3	4	6

   a. Find the agreement between the ranking of two judges using Spearman method.

   b. If you calculate Pearson's *r* as the index of agreement between two judges, would you expect *r* to be identical with ρ? Explain.

2. The rank correlation coefficient of marks obtained by 10 students in Curriculum and Psychology was found to be 0.2. It was later discovered that the difference in ranks in the two subjects obtained by one of the students was wrongly taken as 9 instead of 7. Find the correct value of the coefficient of ρ correlation. (Corrected ρ = 0.394)

3. The ranks obtained by 10 individuals before and after training of some courses are given below. Decide whether the training is equally effective for all individuals.

Individuals	A	B	C	D	E	F	G	H	I	J
Ranks before training	1	6	3	9	5	2	7	10	8	4
Ranks after training	6	8	3	7	2	1	5	9	4	10

4. A teacher ranked seven of his pupil according to their academic achievement. The order of achievement together with their family income for each pupil is given below. What would you say about the relation between family income and academic achievement ?

Pupil	A	B	C	D	E	F	G
Ranking	1	2	3	4	5	6	7
Family income (Rs)	9200	4500	6500	8000	27000	17500	16700

4.6 Biserial Correlation

As already stated we have many problems related to correlation in which first variable is continuous and quantitative and second variable is normal and dichotomous. Here the dichotomous refers to the variable that can be cut into two categories. For example, social adjustment can be divided into socially adjusted and socially maladjusted, training situation can be categorized into trained and untrained, economic status can be reduced into poor and not poor, morality can be divided into moral and immoral, effort can be put in successful and unsuccessful. Similarly, other categories are athletic-non athletic, radical-conservative, introvert- extrovert, social minded-mechanical minded, drop-outs – stay ins, fifth grade - above / below fifth, etc. All these categories are artificial because we divided them according to our convenience. There is no clear and specific demarcation for division or objectivity for categorization.

To find the relation between any trait and such artificial dichotomous variable, previous methods of correlations are not applicable and then we should use new technique. For example, a teacher may be interested to find the relation between academic achievement and economic status of students in certain grade. He can take achievement score as first variable and categorization of economic status in poor and not poor as second variable.

Similarly, in many test questions and various sort of items are scored in dichotomous form. For, problems are marked as passed or failed, statements are true or false, personality inventory items are yes or no, interest inventory items are like or dislike, and so on. When two categories split cannot be regarded as representing an underlying normal distribution upon which an arbitrary division has been imposed but it is in fact two discrete groupings, the point biserial *r* is the appropriate measure of correlation (Garrett, 2008, p. 376). Therefore, biserial is a technique to find relationship between the set of scores representing a trait and any other trait with two-fold categorization. If there is continuity in dichotomized trait, normality of distribution underlying the dichotomy, a split that is not to extreme (the closer to 0.50 the better) and a large *N* then biseral *r* is an estimate of product-moment *r* (p. 380).

Besides, it has some limitations such as it cannot be used in regression, the biserial coefficient has no standard error of estimate and the score predicted for all the member of group is simply the mean of that category. The value of r_{bis} is not limited to ±1. The value may be larger and more reliable then the value of r.

The biserial correlation coefficient can be computed using following formula.

$$r_{bis} = \frac{M_p - M_q}{\sigma_t} \times \frac{p.q}{y} \text{ ....(4.7)}$$

In which, p = Proportion of cases in large category of dichotomous variable

q = Proportion of cases in lower group = 1-p

M_p = Mean of proportion of first category

M_p = Mean of proportion of second category

σ_t = Standard deviation of total group

y = Height of ordinate separating p & q in normal curve.

Alternative method

$$r_{bis} = \frac{M_p - M_t}{\sigma_t} \times \frac{p}{y} \text{...(4.8)}$$

In which, M_t = Mean of the entire group.

Steps for biserial correlation

1. Prepare table showing first variable and dichotomous variable.
2. Find proportion(p) for the case in larger group and q for the case in smaller group. Compute M_p mean of p, and M_q mean of q. To calculate r_{bis} using alternative method compute mean of entire group M_t.
3. Find the height of the normal curve in ordinate (in Y-axis) representing p and q. (For this see table in the Appendix B in which standard scores and ordinates corresponding to divisions of the area under normal curve is given)
4. Substitute these values in formula and get r_{bis}.
5. On the basis of obtained result analyze the extent of relation between variables.

Example 4.5. The scores obtained by a group of trained and untrained students in music appreciation test are given below. Interpret the association of training in test scores on music appreciation test.

Scores	60-64	55-59	50-54	45-49	40-44	35-39	30-34
Trained	4	0	7	6	0	4	2
Untrained	12	15	30	35	20	27	10

Solution: Formation of table to compute required components for r_{bis}.

Table 4.9: Calculation of r_{bis} for the data obtained in music appreciation test

Scores	Trained (f_q)	Untrained (f_p)	Total (f)	Value (m)	d	f_pd	f_qd	fd	d^2
60-64	4	12	16	62	3	36	12	48	144
55-59	0	15	15	57	2	30	0	30	60
50-54	7	30	37	52	1	30	7	37	37
45-49	6	35	41	47	0	0	0	0	0
40-44		20	20	42	-1	-20	0	-20	20
35-39	4	27	31	37	-2	-54	-8	-62	124
30-34	2	10	12	32	-3	-30	-6	-36	108
	$\Sigma N_1 =$ 23	ΣN_2 = 149	ΣN =172			Σf_pd = -8	Σf_qd = 5	Σfd = -3	Σd^2 =493

Now, Proportion of larger group $(p) = \frac{N_2}{N} = \frac{149}{172} = 0.866$

Proportion of smaller group $(q) = \frac{N_1}{N} = \frac{23}{172} = 0.134$

Height of ordinate separating p and q near median (y) = 0.2171

Mean of larger group $(M_P) = A + \frac{\Sigma f_P.d}{N_2} \times i$

In which, A = Assumed mean = 47

I = Class interval = 5

N_2 = Number of items in larger group = 149

$d = \left(\frac{m - A}{i}\right)$

$$\therefore M_p = 47 + \frac{(-8)}{149}\times 5 = 46.73$$

Similarly,

Mean of smaller group (M_q) = $A + \frac{\Sigma f_q.d}{N_1}\times i = 47 + \frac{5}{23}\times 5 = 48.08$

Mean of entire group (M_t) = $A + \frac{\Sigma fd}{N}\times i = 47 + \frac{(-3)}{172}\times 5 = 46.91$

Standard deviation of entire group (σ_t)

$$= \sqrt{\frac{fd^2}{N} - \left(\frac{fd}{N}\right)^2}\times i = \sqrt{\frac{493}{172} - \left(\frac{-3}{172}\right)^2}\times 5 = 8.764$$

$$\therefore r_{bis} = \frac{M_p - M_q}{\sigma_t}\times\frac{pq}{y} = \frac{46.73-48.08}{8.464}\times\frac{0.866\times 0.134}{0.2171} = -0.085$$

Using alternative method

$$r_{bis} = \frac{M_p - M_t}{\sigma_t}\times\frac{p}{y}$$

$$= \frac{46.73-46.91}{8.464}\times\frac{0.866}{0.2171}$$

$$= -0.0848$$

Therefore, the biserial correlation coefficient is slightly negative from both methods. This means that there is no relationship (negligible) between achievement and music appreciation test.

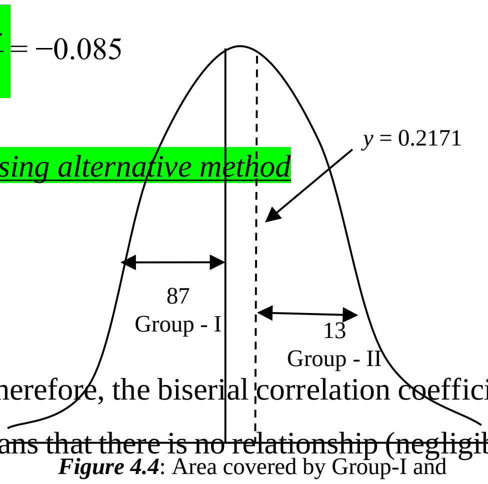


Figure 4.4: Area covered by Group-I and Group II

Exercise 4.3

1. The performance scores obtained by a group of 200 students with intelligence below normal, and above normal are given.

Performance	130 -139	120 -129	110 -119	100 -109	90 -99	80 -89	70 -79	60 -69	50 -59	40 -59
Below normal	0	0	3	7	16	21	11	4	6	2
Above normal	5	7	7	26	30	27	10	3	1	0

Decide the influence of intelligence on performance

2. The distribution of scores on an achievement test earned by those students who answered 50% or more and those who answered less than 50% of the items in an arithmetic test correctly are given below. Compute r_{bis}.

Achievement Scores	185- 194	175- 184	165- 174	155- 164	145- 154	135- 144	125- 134	115- 124	105- 114
50% ≤	7	16	10	35	24	15	10	3	0
> 50%	0	0	6	15	40	26	3	5	5

4.7 Point-biserial Correlation

In same situations we have to measure the relationship between two variables in which one variable is continuous and the other is reducible to pure or genuine dichotomy, e.g. when items are indicated as true or false, or scored as simply 1 if correct and 0 if incorrect, individual separated as male or female, living or dead, psychotic or normal, colour blind or normal, loyal or disloyal and so on. In above example, the division of dichotomous variable in each case is clear and natural but not artificial. Hence, if we are sure that the division of dichotomous variable does not belong to the category of artificial or forceful dichotomy, then we should try to compute point biserial correlation coefficient.

Therefore, the relation between continuous variable and natural dichotomous variable is point biserial correlation and generally denoted by r_{pbis}. The value of point biserial can be obtained using following formula,

$$r_{pbis} = \frac{M_p - M_q}{\sigma_t} \sqrt{p.q} \text{ .... } (4.9)$$

Alternatively,

$$r_{pbis} = \frac{M_p - M_t}{\sigma_t} \sqrt{p/q} \text{ ....}(4.10)$$

In which, p = Proportion of cases in larger group.

q = Proportion of cases in smaller group.

M_p = Mean of larger group in dichotomous variable.

M_q = Mean of smaller group in dichotomous variable.

M_t = Mean of the total group in dichotomous variable.

σ_t = Standard deviation of entire group.

Properties of point biserial and comparison to biserial

1. It has no strict assumption for the distribution of dichotomous variable. For e.g., continuity, normality and large N that split near median.
2. It can be computed easily relative to biserial.
3. It can be use instead of product-moment r.
4. The range of r_{pbis} coefficient lies between ±1, but this is not true for biserial. Due to this property it can be compare with the values from other correlation methods.
5. Standard error of r_{pbis} can be estimated and significance of the hypothesis can be tested.
6. Although both are applicable in item analysis, r_{pbis} is more valid measure.
7. It can be used in regression equation.
8. When we are not sure whether the split of dichotomous variable is natural or artificial, we can safely use r_{pbis} but cannot r_{bis}.

Steps for point biserial correlation

1. As in biserial method tabulate the given value in the form of first variable and second (dichotomous) variable.
2. Find proportion p dividing sum of larger group score by sum of entire scores and find q = 1- p.

3. Compute mean of the first (larger) group (M_p), second (smaller) group (M_q) and total group (M_t) separately.
4. Calculate standard deviation of entire group (second variable).
5. Substitute these values in eq. 4.9 and 4.10 to find the value of r_{pbis}.

Example 4.6: Scores secured by a group of 10 students in intelligence test and result of mechanical aptitude test (1 for pass and 0 for fail) are given. Compute point biserial correlation coefficient to find the relation between intelligence and mechanical aptitude.

Students	A	B	C	D	E	F	G	H	I	J
Intelligence test (*X*)	105	100	117	120	119	95	110	115	108	98
Mechanical aptitude (*Y*)	1	0	1	1	0	0	1	1	0	1

Solution: To compute point biserial correlation we need the values of p, q, M_p, M_q, M_t, and σ_t for which the given data can be organized in the following way.

***Table 4.10*:** Computation of r_{pbis} between intelligence score and mechanical aptitude test score.

Students	X	Y	X^2	Y^2	XY
A	105	1	11025	1	105
B	100	0	10000	0	0
C	117	1	13689	1	117
D	120	1	14400	1	120
E	119	0	14161	0	0
F	95	0	9025	0	0
G	110	1	12100	1	110
H	115	1	13225	1	115
I	108	0	11664	0	0
J	98	1	9604	1	98
	$\Sigma X = 1087$	$\Sigma Y = 6$	$\Sigma X^2 = 118893$	$\Sigma Y^2 = 6$	$\Sigma X\,Y = 665$

Now,

Proportion of larger group (p) $= \frac{\Sigma 1}{N} = \frac{6}{10} = 0.6$

Proportion of smaller group (q) $= 1 - p = 1 - 0.6 = 0.4$

Mean of larger group (M_p) $= \frac{\text{Sum of scores in } X \text{ who passed in } Y}{\text{Number students passed in } Y}$

$$= \frac{105+117+120+110+115+98}{6}$$

$$= \frac{665}{6} = 110.83$$

Mean of smaller group (M_q) $= \frac{\text{Sum of scores in } X \text{ who failed in } Y}{\text{Total students who failed in } Y}$

$$= \frac{100+119+95+108}{4} = 105.5$$

Standard deviation of total (σ_t) $= \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2}$

$$= \sqrt{\frac{118893}{10} - \left(\frac{1087}{10}\right)^2}$$

$$= 11889.3 - 11815.69 = 8.579$$

Now, $r_{pbis} = \frac{M_p - M_q}{\sigma_t}\sqrt{p.q}$

$$= \frac{110.83 - 105.5}{8.573} \times \sqrt{0.6 \times 0.4} = 0.30$$

The result shows that the relation between intelligence and mechanical aptitude is poor.

To calculate r_{pbis} using alternate formula, we must compute mean of entire group, than

$$M_t = \frac{\Sigma X}{N} = \frac{1087}{10} = 10.87$$

$$\therefore r_{pbis} = \frac{M_p - M_t}{\sigma_t} \times \sqrt{\frac{p}{q}}$$

$$= \frac{110.83 - 108.7}{8.579} \times \sqrt{\frac{0.6}{0.4}} = 0.30$$

Example 4.7: The distributions of scores obtained by 65 students on a certain test (*X)* and another dichotomous test (*Y*) is given below. The test *Y* was scored as 1 for right and 0 for wrong responses. Measure the relationship between the scores obtained in two tests.

Scores on *X*		49-45	44-40	35-39	30-34	25-29	20-24
Scores on *Y*	Correct	7	6	3	3	1	0
	Incorrect	6	8	6	9	4	12

Solution: Here, the distribution of scores in first set (*X*) is continuous and second set (*Y)* is splitted naturally on correct and incorrect responses. So, it is better to use the point-biserial correlation formula to measure the association between two variables. Therefore, we can use the following process to compute required relation. Using given data we can tabulated the given value as below.

Table 4.11: Computation of r_{pbis} of given continuous data

Scores (*X*)	Students (*Y*)		Total (*f*)	Value (*m*)	*d*	f_pd	f_qd	*fd*	fd^2
	Correct(*p*)	Incorrect(*q*)							
49-45	7	6	13	47	2	14	12	26	52
44-40	6	8	14	42	1	6	8	14	14
35-39	3	6	9	37	0	0	0	0	0
30-34	3	9	12	32	-1	-3	-9	-12	12
25-29	1	4	5	27	-2	-2	-8	-10	20
20-24	0	12	12	22	-3	0	-36	-36	108
	N_p = 20	N_q = 45	*N* =65			$\Sigma f_p d$ = 15	$\Sigma f_q d$ = 33	Σfd =18	Σfd^2 =206

Now, Proportion of students in first group $(p) = \frac{20}{65} = 0.308$

Proportion of students in second group $(q) = \frac{45}{165} = 0.692$

Mean of scores who responded correctly $(M_p) = A + \frac{\Sigma f_p d}{N_2} \times i$

In which, A = Assumed mean = 37

N_2 = Number of students with correct responses = 20

i = Class interval = 5, $d = \left(\frac{m - A}{i}\right)$

$\therefore M_p = 37 + \frac{15}{20} \times 5 = 40.75$

Similarly , (M_q) $= A + \frac{\Sigma f_q . d}{N_q} \times i = 37 + \frac{(-33)}{45} \times 5 = 33.33$

Standard deviation of entire group (σ_t)

$$= \sqrt{\frac{fd^2}{N} - \left(\frac{fd}{N}\right)^2} \times i = \sqrt{\frac{206}{65} - \left(\frac{-18}{65}\right)^2} \times 5 = 8.78$$

$$\therefore r_{pbis} = \frac{M_p - M_q}{\sigma_t} \times pq = \frac{40.75 - 33.33}{8.78} \times \sqrt{0.308 \times 0.692} = 0.39$$

To calculate r_{pbis} using alternate formula, we need mean of the entire group (M_t)

$$\therefore M_t = A + \frac{\Sigma fd}{N} \times i = 37 + \frac{(-18)}{65} \times 5 = 35.61$$

$$\therefore r_{pbis} = \frac{M_p - M_t}{\sigma_t} \times \sqrt{\frac{p}{q}}$$

$$= \frac{40.75 - 35.61}{8.78} \times \sqrt{\frac{0.308}{0.692}} = 0.39$$

Exercise 4.4

1. The data in the table below presents the scores on Miller Analogies Test (*X)* and the corresponding number of individuals participated in VA training (*Y*) who failed in program and obtained Ph. D. Compute r_{pbis} and r_{bis}. Which is the more appropriate coefficient for these data ?

	X	95-99	90-89	85-89	80-84	75-79	70-74	65-69	60-64	55-59	50-54	45-49	40-44	35-39	30-34
Y	Failed in program	0	1	0	2	4	6	8	3	2	6	2	3	1	1
	Obtained Ph.D.	1	1	6	11	6	9	3	2	1	-	-	-	-	-

2. A group of students with and without training, obtained the following scores on a performance test. Find out the biserial correlation between training and performance.

Performance test scores	90-99	80-89	70-79	60-69	50-59	40-49	30-39	20-29
Trained	6	19	31	58	40	18	9	5
Untrained	0	3	5	17	30	14	7	4

3. A group of individuals were asked to answer 'Yes' or 'No' for a particular item. Compute the point biserial correlation coefficient between the item and total score from the following data:

Scores on opinion scale	95-99	90-94	85-89	80-84	75-79	70-74	65-69	60-64
Yes	0	1	0	2	4	6	8	3
No	1	1	6	11	6	9	3	2

4.8 Phi-correlation

In some situations we have to "compute correlation between two variables in which both the variables are genuinely dichotomous. Then the appropriate method to measure the association is phi-correlation and usually denoted by Greek letter ϕ.

The ϕ coefficient is an appropriate measure of correlation when the two fold classification is truly discrete with no possibility of intermediate value. For example, living-dead, pass-fail, employed-not employed, like dislike, agree-disagree, yes-no, etc. Phi may also be used with normally distributed continuous variables when the categories are genuinely dichotomous. For instance two tests are split as the median (right or wrong). The relation between ϕ and tetrachoric is same as the relation between biserial and point-biserial. The ϕ, since r_{pbis}, is a product - moment r and can be checked directly against r obtained from same table.

It is most useful in item analysis when we want to know the item to item correlation. The values of ϕ coefficient ranges between −1 to +1. For better measure of relationship, it does not require a split near median. When there is any doubt regarding the nature of dichotomy of variable, it is always safe to compute ϕ instead of tetrachoric coefficient (r_t). This mean that ϕ is more dependable than r_t. The relation between ϕ and chi-square (χ^2) can be expressed as $\chi^2 = N\phi^2$.

The ϕ-coefficient can be obtained using the following formula.

$$\phi = \frac{AD - BC}{\sqrt{(A+B)\,(C+D)\,(B+D)\,(A+C)}} \text{ ....(4.11)}$$

Where, A, B, C, D represent the frequencies in each cell of fourfold table.

When the four fold table is expressed in proportional form, the ϕ can be computed using

$$\phi = \frac{ad - bc}{\sqrt{pqp'q'}} \text{ ....(4.13)}$$

Example 4.8: Two items X and Y are part of a test of 100 items Item X is passed by 100 and failed by 125 students in a group of 225. Item Y is passed by 135 and failed by 90 in the same sample. Find the correlation between X and Y.

Solution: The given information gives following 2 × 2 table

		Item *X*		
		Failed	Passed	Total
Item *Y*	Passed	B (55)	A (80)	A+B (135)
	Failed	D (70)	C (20)	C+D (90)
	Total	B+D (125)	A + C (100)	A+B+C+D (225)

Substituting the values of A, B, C, D in above formula, we get

$$\phi = \frac{80\times 70 - 55\times 20}{\sqrt{135\times 90\times 125\times 100}} = 0.36$$

Therefore, the above value of phi indicates the poor relationship between two items.

Example 4.9: One hundred individuals in a survey sample responded to items No 1 and 2 of interest inventory (in yes or no) as in the following table. Find the relation between two items using ϕ.

		Item 1		
		No	Yes	Total
Item 2	Yes	27	20	47
	No	24	29	53
	Total	51	49	100

Steps for ϕ - correlation

1. Prepare 2 × 2 table using given data.
2. Find values of A, B, C, D and other components from the table.
3. Substitute these values and compute ϕ.
4. Analyze and interpret the obtained value.

4.9 Tetra-choric Correlation

Sometime we have to find the correlation ship between two variables in which both the variables are dichotomous in artificial form. That is the categories cannot be expressed in score form. In such situation we use a technique of measuring relationship called tetra-choric correlation and generally denoted by (r_t).

Tetra-choric r_t is, especially, useful when we wish to find the relation between two attributes neither of which is measurable in scores but both of which are capable of being separated into two categories. For instance, if we want to study the relationship between intelligence and emotional maturity. The first variable intelligence may be dichotomized as average and below average and the second variable emotional maturity as emotionally mature and emotionally immature. Similarly, to study relationship between adjustment and success in job, we may divide the variables as adjusted - maladjusted, and success – failure. In the same way the variables poverty and delinquency may be divided into poor- not poor and delinquent-non delinquent respectively. To be r_t applicable, both the variables under study should essentially be continuous and normally distributed if the variables can be expressed in frequency distributions and should be in score form. This method is more applicable only when N is large and the splits are almost equal.

The appropriate formula for tetrachoric r_t are

Case I. When AD > BC

$$r_t = \text{Cos} \left(\frac{180 \times \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}} \right) \text{.......................................(4.14)}$$

Case II. When AD < BC

$$r_t = \text{Cos} \left(\frac{180 \times \sqrt{AD}}{\sqrt{AD} + \sqrt{BC}} \right) \text{...(4.15)}$$

Case III. When AD = BC

$$r_t = \text{Cos} \left(\frac{180 \times \sqrt{AD}}{2\sqrt{AD}} \right) \text{..(4.16)}$$

Where, A, B, C, D are frequencies of cells in 2 × 2 table.

Steps for computation

1. Compete AD and BC
2. Substitute these value in appropriate form of above formula.
3. Find the value of r_t and describe the value.

Example 4.10: Find the relationship between social adjustment and job success with the help of following data.

	Unsuccessful	Successful
Well adjusted	25	35
Poorly adjusted	30	10
Total	55	45

Solution: On the basis of above data following fourfold table can be constructed.

	X-variable (Job success)			
Y-variable (Adjustment)		Unsuccessful	Successful	Total
	Well adjusted	B (25)	A (35)	A + B = (60)
	Poorly adjusted	D (30)	C (10)	C + D (40)
	Total	B + D (55)	A + C (45)	100

Now substituting these values in formula

$$r_t = \text{Cos}\left(\frac{180 \times \sqrt{BC}}{\sqrt{AD} + \sqrt{BC}}\right) = Cos\left(\frac{180 \times \sqrt{250}}{\sqrt{1050} + \sqrt{250}}\right) = 0.52$$

The result shows that there is positive relation between success and social adjustment capability.

Exercise 4.5

1. Compute tetra-choric r_t for the following tables that exhibits relation of alcoholism and health in 811 fathers and sons. Entries are expressed as proportions.

		Sons		
		Unhealthy	Healthy	Total

Fathers	Nonalcoholic	0.343	0.405	0.748
	Alcoholic	0.102	0.150	0.252
	Total	0.445	0.555	1.000

Note: Conditions for applying r_t

As ϕ and r_t are both used to determine the relationship in a fourfoldtable, doubt may create which technique is applicable for the given data. To avoid this confusion the following points will be helpful.

- When the distribution is normal, variable is continuous, and *N* is large, we may safely use r_t.
- If r_t holds above three assumptions, it is good estimate of *r*.
- Tetra-choric coefficient ranges from +1.00 to -1 regardless of the relative size of the marginal total (divisions).
- The standard error of r_t is difficult to compute and is always greater than the *SE* of comparable *r*.

4.10 Partial and Multiple Correlation

Partial Correlation: We have discussed different types of measures of relationship between variables. In linear relationship the measure ascertain the association between two variables irrespective of third variable. However, in education and psychology, the relationship between two variables is greatly influenced by a third or additional valuables. In such situation the estimation of correlation between these two variables may not be accurate and the prediction on the basis of such result is liable to lead error. In order to overcome from such error we must nullify the effect of third or additional variable. There are two methods of controlling the influence of undesirable variables. First, we can do this experimentally by selecting the subjects from same intervening variable situation (age, socio-economic status, gender, etc). But this matching process is impracticable in some situation because of its tediousness. Alternatively, to find the relationship between required variables controlling the undesired effect of third or additional variable, we have another more convenient statistical technique called partial correlation.

Suppose a researcher wants to find out the relationship between general intelligence and school achievement, then the relationship may be influenced by age factor (additional variable). To find more accurate coefficient he/she should partial out the effect of age. In the same way, to find the relation between height and weight, age may be additional variable. Similarly, in a study that shows the relationship between participation in co-curricular activities and academic achievement a number of variables like intelligence, socio-economic status, environmental difference, age, health, and other factors may be responsible for increasing or decreasing the relationship between major variables. In partial correlation we attempt to reduce the effect of such variables and make the result more reliable.

Therefore, partial correlation is based on the assumption that "controlling the main variables (two or three) automatically control other related intervening variables". Based on this assumption we can compute partial correlation using following formulae

When only one undesirable variable is held constant, we compute first order partial correlation using the formula,

$$r_{12,3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{(1-r_{13}{}^{2})\,(1-r_{23}{}^{2})}} \text{ .... } (4.17)$$

Where the numbers 1 and 2 signify the major variables and 3 the undesirable variables. Therefore, $r_{12.3}$ means the partial correlation between major variables 1 and 2 controlling the effect of intervening variable 3.

When two variables are kept constant we compute second order partial correlation using the formula.

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}.r_{24.3}}{\sqrt{(1-r_{14.3})\,(1-r_{24.3})}} \text{ .... } (4.18)$$

Where $r_{12.34}$ denotes the relationship between two variables 1 and 2 controlling two intervening variables 3 and 4.

Similarly, when three variables are partial out, we compute third order partial correlation using following formula,

$$r_{12.345} = \frac{r_{12.34} - r_{15.34}.r_{15.34}}{\sqrt{(1 - r_{15.34}{}^{2})\,(1 - r_{25.34}{}^{2})}} \text{ .... } (4.19)$$

Where $r_{12.345}$ is the partial correlation between two variables 1 and 2 reducing the effect of three variables 3, 4 and 5.

Steps for partial correlation

1. Obtain data related to main variables and undesired variables.
2. If you have to compute partial correlation using raw data then find correlation between variables 1 and 2, 1 and 3, 2 and 3, and so on.
3. If you have given the correlation coefficient between different two variables, then substitute these values directly in partial correlation formula.
4. Describe/interpret obtained value of partial correlation to get conclusion.

Example 4.11: A researcher wants to find an independent correlation between intelligence and academic achievement of certain group of students. The correlation between intelligence and achievement was found to be 0.80; intelligence and age was found to be 0.70; and achievement and age was found to be 0.60. Find out partial correlation between intelligence and achievement.

Solution: Suppose the variables intelligence, achievement, and age are denoted by index 1, 2, and 3 respectively. Then by given condition, correlation between intelligence and achievement becomes $r_{12} = 0.80$, correlation between intelligence and age becomes $r_{13} = 0.70$, and correlation between achievement and age equals $r_{23} = 0.60$. Now partial correlation between intelligence and achievement reducing age effect equals.

$$r_{12,3} = \frac{r_{12} - r_{13}.r_{23}}{\sqrt{(1 - r_{13}{}^2)\,(1 - r_{23}{}^2)}} = \frac{0.80 - 0.70 \times 0.60}{\sqrt{[1 - (0.70)^2\,][1 - (0.60)^2\,]}} = 0.67$$

Exercise 4.5

1. An investigator, during his study collected data and found simple correlation as given below:

a) Between height and weight = 0.80

b) Between height and age = 0.60

c) Between weight and age = 0.50

Among three variables which variables are more related ? Decide.

2. A sample of 500 students of grade VIII were evaluated in terms of their academic achievement and their participation in co-curricular activities. Their IQ's were also tested and inter correlation among these three variables were obtained as r_{12} = 0.82, r_{13} = 0.80. Find out the net correlation between first two variables.

Multiple Correlation: In partial correlation, we find the relationship between two major variables excluding the effect of one or two or three undesirable variables. But, in many situations in education, we need to find the effect of one or more variables upon single variable. For example, the achievement of students may be influenced by many variables like intelligence, family education, economic status, neighborhood environment, teacher quality, students interest, number of study hours, and so on. The first variable 'achievement' is dependent variable and remaining other variables are independent variables. If we want to study an aggregate effect of two or more independent variables upon single dependent variable, we need a special technique or statistics called coefficient of multiple correlation Therefore, multiple correlation shows the degree of relationship between single dependent variable and combined effect of two or more independent variables.

If we denote dependent variable by X_1 and independent variables by X_2 and X_3 then the multiple correlation denoted by R can be computed by

$$R_{1.23} = \frac{r_{12}{}^{2} - r_{13}{}^{2} - 2r_{12}r_{13}r_{23}}{\sqrt{1 - r_{23}{}^{2}}} \text{ .... (4.20)}$$

Where, $R_{1.23}$ denotes the coefficient of multiple correlation between dependent variable X_1 and combined effect of two independent variable X_2 and X_3. Similarly, r_{12} is correlation between X_1 and X_2, r_{13} is correlation between X_1 and $X_{3,}$ and r_{23} is correlation between X_2 and X_3.

Characteristics of multiple correlation

1. It measures the strength of association of single dependent variable with two or more independent variables.
2. It shows the relationship with correlation between independent variables as well as the correlations of these variables with dependent variables.
3. It shows accurate value only for large number of sample or cases.
4. The multiple correlation coefficient is always positive and less than 1 (i.e. $0 < R < 1$).
5. It estimates combined effect of independent variables upon dependent variable.

Example 4.12: Suppose an investigator collects scores secured by a group of students in a particular subject in three different tests and compute correlation coefficient as follows:

a) Correlation between scores of final exam (X_1) and first term exam (X_2) = 0.72

b) Correlation between final exam (X_1) and second term exam (X_3) and first term exam (X_2) = 0.72

c) Correlation between first term (X_2) and second term exam (X_3) = 0.54

Compute $R_{1.23}$ and interpret the result.

Solution: From given condition,

Correlation between X_1 and X_2 gives $r_{12} = 0.72$

Correlation between X_1 and X_3 gives $r_{13} = 0.36$

Correlation between X_3 and X_2 gives $r_{23} = 0.54$

But we have, multiple correlation coefficient formula involving three variables as

$$R_{1.23} \quad = \sqrt{\frac{r_{12} + r_{13}^{\ 2} - 2r_{12}.r_{13}.r_{23}}{1-(r_{13})^2}}$$

Substituting above values in multiple correlation formula, then

$$R_{1.23} \quad = \sqrt{\frac{(0.72)^2 + (0.36)^2 - 2\times 0.72\times 0.36\times 0.54}{1-(0.54)^2}} = 0.52$$

The multiple correlation coefficient 0.52 between final test and unit test indicates that the scores secured by students in final test is influenced to some extent by the combined effect of scores in first term and second term test.

Example 4.13: Consider a researcher wishes to study the multiple correlation of self-esteem (X_1) with achievement-motivation (X_2) and locus of control (X_3), and found r_{12} = 0.54, r_{13} = 0.37 and r_{23} = 0.26.

Solution: The multiple correlation coefficient for three variables,

$$R_{1.23} \quad = \sqrt{\frac{r_{12}^{\ 2} + r_{13}^{\ 2} - 2r_{12}.r_{13}.r_{23}}{1-r_{13}^{\ 2}}}$$

Substituting given values in above formula gives

$$R_{1.23} \quad = \sqrt{\frac{(0.54)^2 + (0.37)^2 - 2\times 0.54\times 0.36\times 0.26}{1-(0.26)^2}} = 0.59$$

The coefficient 0.59 indicates that self-esteem is related to some extent to achievement - motivation and locus of control taken together.

Alternatively, the multiple correlation coefficient can be computed using another less convenient formula then previous one when standard deviation of scores are also available. This can be clearer from following example.

Example 4.14: Find coefficient of multiple correlation using following data collected by researcher related to three variables.

Self-esteem (X_1)	Achievement Motivation (X_2)	Locus of control (X_3)
σ_1 = 19.64	σ_2 = 2.89	σ_1 = 7.05
σ_{12} = 0.54	σ_{23} = 0.26	σ_{12} = 0.37

Solution: We have new formula to compute multiple correlation of a variable with two other variables:

$$R_{1.23} = \sqrt{1 - \frac{\sigma_{1.23}{}^{2}}{\sigma_1{}^{2}}} \text{ ....(4.21)}$$

In which,

σ_1 = Standard deviation of first variable (X_1)

$\sigma_{1.23}$ = $\sigma_1 \sqrt{(1\text{-}r_{12}{}^{2})\ (1\text{-}r_{13.2}{}^{2})}$

r_{12} = Correlation between X_1 and X_2

$r_{13.2}$ = Partial correlation between X_1 and X_3 keeping X_2 constant

Now using given values

σ_1 = 19.64, r_{12} = 0.54

$$r_{13.2} = \frac{r_{13} - r_{12} \times r_{23}}{\sqrt{(1 - r_{12}{}^{2})\,(1 - r_{23}{}^{2})}} = \frac{0.37 - 0.54 \times 0.26}{\sqrt{[1 - (0.54)^2]\ [1 - 0.26]}} = 0.28$$

$$r_{1.23} = \sigma_1 \sqrt{[1 - (r_{12})^2\ [1 - (r_{13.2})^2} = 19.64 \sqrt{[1 - (0.54)^2\,][1 - (0.28)^2\,]} = 15.86$$

Substituting these values in formula (4.20) we get,

$$r_{1.23} = \sqrt{1 - \frac{\sigma_{1.23}{}^{2}}{\sigma_1}} = \sqrt{1 - \frac{(15.86)^2}{(19.64)2}} = 0.59$$

This shows that the new formula gives the same coefficient of multiple correlation as in the previous formula (4.19).

Exercise 4.6

1. Test A correlates 0.60 with criterion and 0.50 with test B, which correlates only 0.10 with the criterion what is the multiple R of A and B with the criterion ? Why it is higher than the correlation of A with criterion ?

2. Let X_1 be a criterion and X_2 and X_3 be two other tests. Compute $R_{1.23}$ using the correlations, and standard deviations given below.

   $\sigma_1 = 5$ $\sigma_2 = 8$ $\sigma_3 = 7$

   $\sigma_{12} = 0.55$ $\sigma_{23} = 0.20$ $\sigma_{13} = 0.45$

3. For a large group of students, scores in theory (X_1), scores in method (X_2), and scores in field work (X_3) are collected and the inter correlation between them are computed as $r_{12} = 0.69$, $r_{13} = 0.45$ and $r_{23} = 0.58$ compute multiple, correlation $R_{2.12}$.

Exercise (Model)

Multiple choice questions

1. To find the association between the rank order of two set of variables the .... correlation method is useful.

   (a) Product moment (b) Phi (c) Rank order (d) Biserial

2. The tetra-choric correlation method is useful to find the relation between two variables in which

   (a) first variable is continuous and quantitative and second is qualitative and naturally dichotomized.

   (b) first is continuous and quantitative and second is normal and artificially dichotomized

   (c) both the variables are qualitative and dichotomized

   (d) both the variables are artificially dichotomized and cannot be expressed in scores.

Short answer questions:

1. The total score obtained by a group of 16 student in a multiple choice test and their respective scores in particular item (say in item no. 4) are given.

Student	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Score in item	1	1	0	0	1	1	0	0	1	1	1	0	1	1	0	1
Total score	34	19	22	26	17	24	11	27	23	14	22	25	14	26	25	16

2. Clarify the concept of tetrachoric correlation with suitable example.

Long answer questions:

1. Describe the concept of correlation as a measure of relationship. "Biserial correlation is an estimation of Pearson's correlation when first variable is continuous and quantitative and second is normal and artificially dichotomized". Justify this statement with suitable example.

2. Describe different cases of rank difference method of computing correlation. Compute rank correlation of the following data obtained before and after particular training.

Individual	A	B	C	D	E	F	G	H	I	J
Ranks before training	2	6	3	9	5	2	7	10	8	4
Score after training	26	28	22	27	22	11	15	19	14	30

Chapter 5

Regression and Prediction

5.0 About the Chapter

In previous chapter the degree of relationship between two variables is discussed but not about the nature of relation. This chapter provides the knowledge regarding the nature of relation and helps predict the value of unknown variable using the value of known variable. To provide the knowledge related to regression and prediction the chapter is divided into five sections. The first section describes the concept, use, and types of regression analysis. This section also shows clear distinction between correlation and regression. Second section presents the standard score forms of regression equation. Third section deals with the raw score form in different equations. Fourth section is discussed about the error that occurs while predicting the nature of dependent variable using the value of independent variable. The fifth section is about some conditions that influence the estimated value to some extent.

On completion of this chapter following specific objectives are assumed to be fulfilled.

- To clarify the concept of regression analysis.
- To explain the use of regression analysis.
- To state types of regression and its relation with correlation.
- To derive regression equation in standard and raw score form.
- To estimate the standard errors of prediction.
- To identify the criteria of best fit.

5.1 Concept of Regression Analysis

In previous chapter we have discussed about correlation between two or more variables. The correlation coefficient shows the degree of relationship between two or more variables but does not show the nature of relationship between these variables. In another word, using correlation coefficient we can explain the degree of relationship between two or more variables but cannot say anything about unknown variables using the value of another known variable unless the variables are perfectly related. For instance, suppose if the correlation coefficient between the scores obtained in Mathematics and Science is 1 then the prediction that a student who secured highest score in mathematics may also secured highest in Science is perfect. But in practice the perfect relation is rare and in such case our prediction is less than perfect. Therefore for reliable prediction we need another technique that use the relationship between variables to predict the value of unknown variable using the value of known variable. This technique is generally known as regression method.

Literally, the word "regression" refers to "moving backward", "going back", "stepping back", or "return to the mean value". This shows that regression is the relationship with average (mean) between variable in terms of original units of the data. British doctor and biometrician Sir Francis Galton (cousin of Darwin) used regression method initially to study the hereditary. According to him,

> “with a correlation coefficient of 0.8 between heights of fathers and children if the average height of certain set of fathers is *x* cm above the general average, the average height of children shall be 0.8 *x* cm above the general average. There is thus a move towards mediocrity" (Sancheti & Kapoor, 1983, p. 9.1).

Galton found that children of tall parents tend to be less tall, and children of short parents less short, than their parents. In other words, the heights of the offspring tend to "move back" toward the mean height of the general population. This tendency toward maintaining "the mean height" Galton called the principle of regression, and the line describing the relationship of height in parent and offspring was called a "***regression line***" (Garrett, 2008, p.153). Thus regression is a line describing the average relationship between two variables.

Now, turn to an example to be clear how regression is different from correlation and the regression line shows the average relationship between two variables. For this, assume the relationship between height and weight of 100 students.

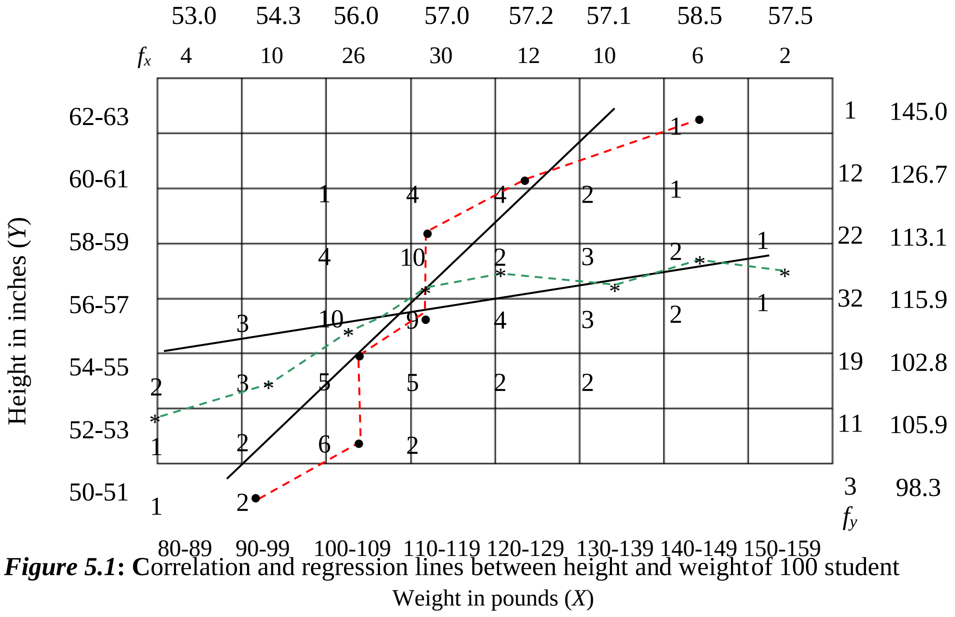


***Figure 5.1*: C**orrelation and regression lines between height and weight of 100 student

In above figure, along the left hand margin from bottom to top are class intervals of the height distribution measured in inches. Along the bottom of the diagram from left to right are class-intervals of weight distribution measured in pounds. Each of the students are plotted in the graph with respect to height and weight. Suppose that a student weights 85 lbs and is 52 inches tall. This weight lies in first column from the left, and his height in the second row from the bottom. In similar manner, location of other students can be identified in graph. Along the top of figure, the f_x row indicates number of students who fall in each weight interval. Similarly, along the right hand margin, f_y column indicates the number of students who fall in each height interval.

The distribution of students in respective height and weight interval forms a scatter gram to which we call correlation table. Analyzing this diagram we can observe some facts. For, the trend of number distribution to the diagram indicates correlation between height and weight. It is clear that the drift of paired height and weight from upper right hand section of the diagram to lower left hand section reveals the positive relationship between height and weight. The more scattered the drift, the less these two variables are related.

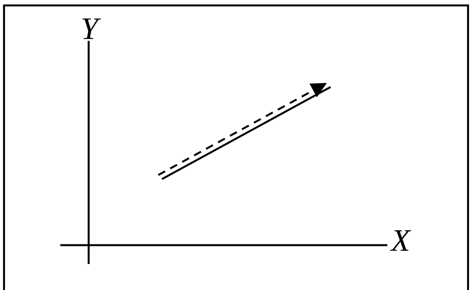
In the same way the relation can be related with the help of average of each height and weight interval in relation to weight and height respectively. The average of each height interval is given on the right margin of diagram as row means and the average of each weight interval is given at the top of diagram as column means. This means can be calculated at hand using assume mean method.

From bottom and top of the diagram we can see that an increase of approximately 70 lbs (154.5 - 84.5) of an actual weight corresponds to an increase of 4.5 inches (57.0-53.0) in mean height. That is, the increase in the lightest to the heaviest man is parallel by an increase of 4.5 inches in height. Therefore, the correlation between weight and height is positive. Similarly, with an actual increase of height by 12 inches parallel to an increase of mean weight by 46.7 lbs reveals positive relation between height and weight. That is, the taller the students, the heavier their weight. The correlation of height in relation to weight and the correlation of weight in relation to height both are positive.

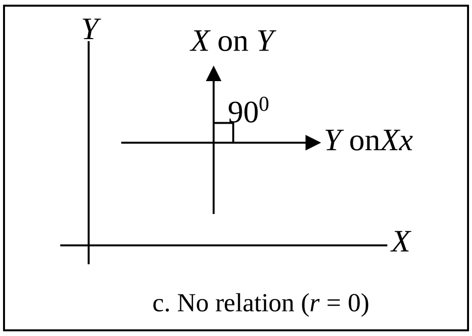
Now, if we find the average of number in *X*-column (weight average in each interval) against their respective height interval and indicate by small crosses, then the joining of these crosses gives an irregular line. This time represents the change in mean value of *Y* over the given range of *X*. Similarly, if we indicate the means of scores in each *Y*-raw by small circle and joined these circles by dotted line, we get another irregular line. This represents the change in the mean value of *X* over a given range of *Y*. These two lines shows linear relationship between variables *X* and *Y*.

Now draw two straight lines, one as close as possible through stars / crosses and another through circles, to describe the general trend of two irregular lines. These lines are two regression lines (Regression of *X* on *Y* and regression of *Y* on *X*). Therefore, regression lines are running means representing series of means. These two lines can be drawn by the "method of least squares". The least square method is based on the principle that the sum of the squares of the deviations of the observed *Y* values from the fitted line shall be minimum." Similar condition applies for line *X* on *Y*. This criterion is known as least-square or minimum squared criterion. Graphical or algebraic method can be used to draw the lines under the assumption of least square criterion. These lines, therefore, are also called lines of "best fit". Using these lines we can describe the nature of relationship.

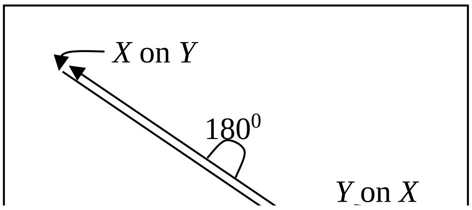
These two lines forms an angle like a scissor blades. If the relationship between two variables is perfect, then these two lines overlap at a single line. As the relationship decreases, the angle between these two lines increases. If one line is perpendicular upon another, then it shows no relationship between these valriables. This concept will be more clearer from following figures.



a. Perfectly positive (r=1)



c. No relation (*r* = 0)



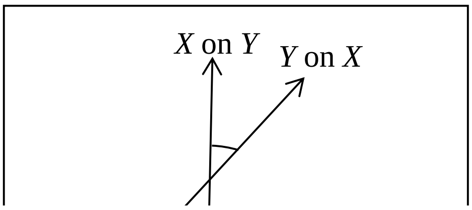


Figure 5.2: Different relationship of regression lines

When the means of *Y*-scores over the whole range of *X*, and the means of *X*-scores over the whole range of *Y* are unchanged, $r = 0$ and two regression lines will be perpendicular to each other (Fig. 5.2c). If the angle increase to 180^0 then the lines in this situation, again overlap in reverse way ($r = -1$) (Fig. 5.2 d).

Now, with the help of these lines we can predict or estimate the value of unknown (dependent) variable using the value of known (independent) variable. Suppose, we want to know the height of a student whose weight is known to be 125 lbs. From figure 5.1 we know that the mean height of all the 12 students who fall in the interval 120-129 is 57.2 inches so the most probable height of a student who weighs, 125 lbs is 57.2 inches. Similarly, the most likely weight of a student whose height is 58.5 inches is 113.1 lbs. This means that the most probable weight of any student in the group is the mean weight of all the students who have approximately same height (within same interval)

In sum, regression is the measure of the average relationship between two or more variables in terms of the original units of the data. Regression analysis, therefore, is the mathematical measure of average relationship between two or more variables in terms of the original units of data. The variable whose value is to be predicted is called dependent and the variable which influence the predicted value is called independent variable. There are always two lines of regression, one of *X* on *Y* and other of *Y* on *X*. The line of regression of *X* on *Y* is used to estimate the value of dependent variable *X* for any known value of independent variable *Y*. Similarly, the line of *Y* on *X* can be used to predict the value of dependent variable *Y* with the help of known value of independent variable *X*. Therefore, these two lines cannot be use interchangeably. The two line intersect through the average value of two variables.

Use of Regression Analysis: Regression analysis has great practical value then correlation analysis in natural science and social science as well. Correlation helps determine the extents of association between variables but not to determine the nature of relationship. Regression analysis, however, is a very useful technique to find the functional relation between two or more variables and thereby help predict the effect of known (independent) variable upon unknown (dependent) variable. More specifically, the use of regression analysis can be given in the following way.

1. Since many problems in education and psychology are linked between two or more variables. The decision about dependent variable without knowing the effect of independent variable is erroneous. Therefore, it is helpful to identify the causing variable/s and to determine its effect.
2. In education it can be used for prediction purpose. For, does a student who secured highest score in particular test would do better in future in a certain area ? Certainly, without some basis we cannot answer about it. Regression analysis will help as a dependable measure.
3. In determining validity of a test it is more powerful tools especially to determine concurrent and predictive validity.
4. For guidance and counseling purpose regression effect can be used more effectively.
5. For research purpose it is useful to establish cause and effect relationship and to determine the nature of relation.

***Types of Regression*:** Regression analysis can be classified into three categories:

a) ***Simple and multiple*:-** The regression analysis related to the study of only two variables is termed as simple regression. For example the effect of study hours on mathematics achievement. In correlation we measure the relationship but in regression we study the effect of independent variable (study hours) upon dependent variable (achievement).

But the regression analysis related to the study of more than two variables at a time is known as multiple regression. In such case one variable (achievement) is a dependent variables and remaining two or more (study hours teaching method, interest, etc.) are independent variables.

b) ***Total and partial:*** When the effect of all the independent variables that are influential upon dependent variables are studied the analysis is known as total. Normally, they take the form of multiple relationship.

On the other hand, in partial regression analysis we study the effect of all the relevant independent variables upon single dependent variable excluding the effect of irrelevant variable. For example, study of effect use of teaching materials reducing the effect of environment, interest, teaching method, etc.

c) ***Linear and non-linear***: If we plot the value of given two variables on the graph then we obtain certain curve joining these intersecting points gives certain path. If the path is in the form of straight line then the regression analysis is linear. Fortunately most of the relations in education and psychology are linear.

However, in some situations, joining these plotted points gives certain type of curve called non-linear path. The analysis to study such relation is known as non-linear regression. For example, the relation between age and intelligence gives curvilinear relation. In other word, the regression is known as non linear if the curve of regression is not a straight line.

Difference between correlation and regression**:** The following are the some major differences between correlation and regression:

1. Correlation shows the general relationship between two variables whereas regression express average relationship between these variables.
2. Generally, correlation analysis is related to linear relationship whereas regression analysis studies linear as well as non-linear relationship.
3. Correlation analysis need not show cause and effect relationship between the variables under study whereas regression analysis clearly indicates the cause (dependent) and the effect (independent) relationship.
4. Correlation coefficient is symmetric, i.e. $r_{xy} = r_{yx}$ but regression coefficients arenot symmetric, i.e. $b_{xy} \neq b_{yx}$.

5. Correlation coefficient is relative measure of the linear relationship between two variables whereas regression coefficient are absolute measures. Correlation is independent of units and its value varies between -1 to 1. But in regression, a unit change in the value of variable implies certain change in the value of *Y* variable.
6. Correlation can be used to find the degree of relationship whereas regression is useful for prediction purpose.

5.2 Regression Equation: Standard Score Form

we have already know that standard scores are the deviations of raw scores from its mean expressed on standard deviation units. Since regression line is the average relationship between two variables or the line of running means representing series of means, we can express these scores in standard score form keeping its means as zero. In other word, to every straight line, there corresponds an equation. Regression line, is a straight line obtained according to least squares criterion, also corresponds and equation in standard score form as given below:

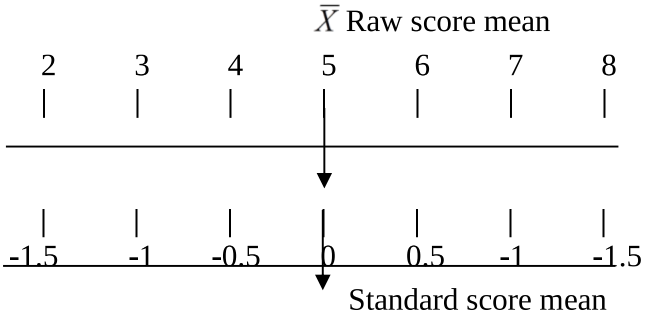


Figure 5.3: Comparison of raw scores and standard scores

The regression equation of *Y* on *X* in standard score form is

$z_y = rz_x$ .... (5.1)

In which, z_y = Predicted standard score value of *Y*

z_x = Standard-score value of *X* form which z_y is predicted.

r = Coefficient of correlation between *X* and *Y*.

Similarly, regression equation of *X* on *Y* in standard score form can be given as:

$z_x = r\ z_y$ ....(5.2)

In which, z_x = the predicted standard score value of X (the regression) z_y = the standard score value of Y from which z_x is predicted.

Note:

- Practically, we do not use the standard-score equation for prediction purpose because the scores from which prediction is made are usually in the form of raw scores. However, it gives some message about prediction.
- If we use $X = \overline{X}$ as a predictor, then

  $Zx = \frac{X - \overline{X}}{\sigma_x} = \frac{\overline{X} - \overline{X}}{\sigma_x} = 0$ gives $z_y = r\,(0) = 0\,r$.

  That is for all values of r, the regression equation shows that the mean of X as predictor always gives the means of Y as predicted value.
- If $r = 0$, then $z_y = 0(z_x) = 0$

  This means that the predicted standard score value is zero. In raw score terms, the predicted value of Y is the mean of Y for any value of X. In this case, knowing the value of X has no advantage in predicting the value of Y.

5.3 Regression Equation: Row Score Form

We know that standard score is the deviation of any given raw score from its mean in standard deviation unit. If X is raw score, $\overline{X}$ its mean, and σ_x the standard deviation of scores then standard score $Z_x = \frac{X - \overline{X}}{\sigma_x}$.

Similarly, the standard score for Y values, $Z_y = \frac{Y - Y}{\sigma_y}$

Substitute these values in equation (5.1) and (5.2) we get the regression equation of Y on X in raw-score form as $\frac{Y - \overline{Y}}{\sigma_y} = r\frac{X - \overline{X}}{\sigma_x}$

or, $Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$ ....(5.3)

In which, Y = predicted raw score,

$\overline{Y}$ = mean of Y scores,

r = correlation between X and Y

The regression equation of *X* on *Y* in raw score form becomes

$$\frac{X-\overline{X}}{\sigma_x}=r\frac{Y-\overline{Y}}{\sigma_y}$$

$$\text{or, } X-\overline{X}=r\frac{\sigma_x}{\sigma_y}(Y-\overline{Y})\text{ ....(5.4)}$$

In which, X = predicted raw score,

$\overline{X}$ = means of *X* scores,

r = correlation between *X* and *Y*,

σ_x and σ_y = standard deviations of *X* and *Y*.

The equation (5.3) helps predict the raw-score value of *X* corresponding any raw score value of *Y*. Similarly equation (5.4) helps estimate the raw score value of *Y* for corresponding value of *X*.

Note: In case of perfect correlation, positive or negative, i.e. $r \neq 1$, the two regression lines concides and become identical. In this case, the equation of lines of regression of *Y* on *X* and *X* on *Y* becomes.

$$\frac{Y-\overline{Y}}{\sigma_y}=\pm\left(\frac{X-\overline{X}}{\sigma_x}\right)$$

$$\text{or } Y-\overline{Y}=\pm\left(\frac{\sigma_y}{\sigma_x}\right)(X-\overline{X})\text{ ....(5.5)}$$

$$\text{and } \frac{X-\overline{X}}{\sigma_x}=r\frac{Y-\overline{Y}}{\sigma_y}$$

$$\text{or, } X-\overline{X}=\pm\left(\frac{\sigma_x}{\sigma_y}\right)(Y-\overline{Y})\text{ ....(5.6)}$$

When the correlation between two variables is zero, i.e., $r = 0$, two regression lines becomes perpendicular to each other and $Y=\overline{Y},\ \ X=\overline{X}$

Example 5.1: A group of five students obtained the following scores on two achievement texts *X* and *Y*.

Students	A	B	C	D	E
Scores in *X*	10	11	12	9	8

Scores in *Y*	12	18	20	10	10

a) Determine regression equation of *Y* on *X* and of *X* on *Y*.

b) If a student scores 15 in test *X*, predict his probable score in test *Y*.

c) If a student scores 10 in *Y*, estimate his probable score in test *X*.

Solution: To compute regression, first we need to calculate mean and correlation coefficient between the scores of two test.

***Table 5.1*: Computation of mean and correlation for regression.**

Students	*X*	*Y*	$x = X - \overline{X}$	x^2	$y = Y - \overline{Y}$	y^2	*xy*
A	10	12	0	0	-2	4	0
B	11	18	1	1	4	16	4
C	12	20	2	4	6	36	12
D	9	10	-1	1	-4	16	4
E	8	10	-2	4	-4	16	8
ΣX= 50		ΣY = 70		Σx^2 = 10		Σy^2 = 88	Σxy = 28

$$\overline{X} = \frac{\Sigma X}{N} = \frac{50}{5} = 10 \qquad \sigma_x = \sqrt{\frac{\Sigma X^2}{N}} = \sqrt{\frac{10}{5}} = 1.4$$

$$\overline{Y} = \frac{\Sigma Y}{N} = \frac{70}{5} = 14 \qquad \sigma_y = \sqrt{\frac{\Sigma Y^2}{N}} = \sqrt{\frac{88}{5}} = 4.2$$

To compute correlation, we have formula in standard score form.

$$r = \frac{\Sigma xy}{N\sigma_x \sigma_y} = \frac{28}{5 \times 1.4 \times 4.2} = 0.98$$

a) Now, line of regression of *Y* on *X* on raw score form

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

or, $Y = 14 = 0.98 \times \frac{4.2}{1.4}$ $(X\text{-}10)$

or, $Y = 2.94X - 15.4$ .... (i)

Similarly, line of regression of X on Y in raw score form is

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

or, $X - 10 = 0.98 \times \frac{1.4}{4.2}$ $(Y - 14)$

or, $X = 0.33Y + 5.42$..(ii)

b) If a student scores 15 in the test X we estimate the value of Y using equation (i)

or, $Y = 2.94 \times 15 - 154 = 28.7$

Therefore, the estimated score in test Y is 28.7.

c) If a student scores 10 in test Y then we use equation (ii) to predict the value of X, therefore, $X = 0.33 \times 10 + 5.42 = 8.72$

That is the student who obtains 10 in Y may obtain 8.72 in test X.

Example 5.2: Given the following data for two tests.

Social study (X)	Nepali (Y)	
$\overline{X} = 75.0$	$\overline{Y} = 70.0$	$r_{xy} = 0.72$
$\sigma_x = 6.0$	$\sigma_y = 8.0$	

a) Find the regression equation in raw score form

b) Predict the probable grade of a student in Nepali whose mark in social study is 65.

Solution:

a) Using given data we can work out for two regression equation as given below:

The regression equation of Y on X is $Y - \overline{Y} = r_{xy}\frac{\sigma_y}{\sigma_x}(X - \overline{X})$

or, $\text{Y} - 70 = 0.72 \times \frac{8}{6}$ $(\text{Y} - 75)$

or, X = 0.96X - 2...(i)

Similarly, the regression of X on Y is $X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$

or, $X - 75 = 0.72 \times \frac{6}{8}\ (Y - 70)$

or, $X = 0.75Y + 22.5$...(ii)

b) Now, we have to find the estimated score of a student in Nepali who obtained 65 in Social study. We use equation (i),

i.e. $Y = 0.96 \times 65 - 2 = 60.4$

Therefore, the student who secured 65 in Social study may obtain 60.4 in Nepali.

Exercise 5.1

1. Following are the marks in Mathematics and English in an annual examination.

	Math (*X*)	English (*Y*)	
Mean	40	50	*r* = 0.5
S.D.	10	16	

Give the equations of lines of regression and estimate the score in English when obtained score in Math is 50 and estimate the score in Math when score in English is 30.

2. Assuming that there is a linear relationship between height (*X*) and weight (*Y*) of pupil in a certain area, estimate the height of an individual whose weight is 150 lbs using the following data.

	Height (*X*)	Weight (*Y*)	
Mean	65 inches	160 lbs	*r* = 0.60
S.D.	3 inches	20 lbs	

Also find the most likely weight of a person whose height is 68 inches.

3. Following are the data associated with a group of students related to an entrance test and a class test.

	Entrance (*X*)	Class test (*Y*)	
Mean	38.00	70.00	*r* = 0.78

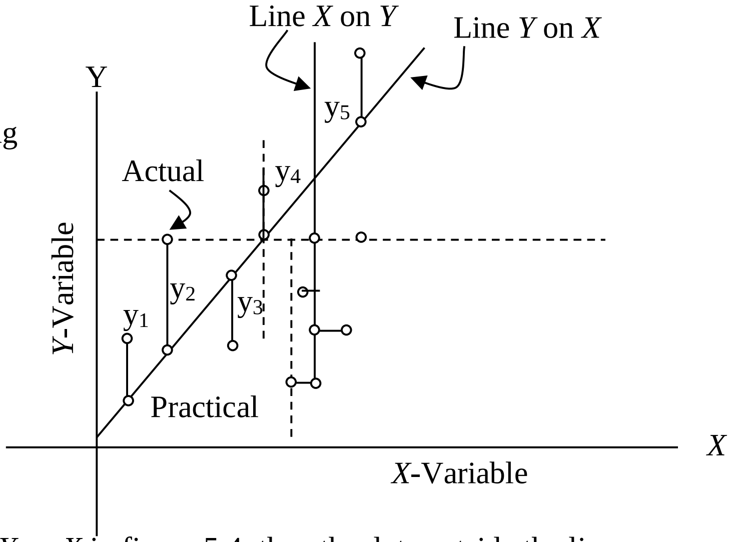
S.D.	10.35	8.28	

Estimate the score of a student in entrance test who obtained 71 in class test. If the student had been obtained 60 in actual entrance test, what is your reflection about the predictive power of entrance test with respect to standardized criterion of class.

5.4 Errors of Prediction : The Standard Error of Estimate

From previous discussion we know that the lines of regression represent the average of values related to two variables. The lines are drawn with principles of least-square criterion so that sum of squares of all the deviations of actual points from predicted points are minimum. Therefore, the regression lines shows average relationship between two variables but not actual relation. This can be clear from Figure 5.1, that the height of a student can be predicted to be 54.3 inches whose weight is 85 lbs. But if we take another student with weight 86 lbs then his height also should be estimated to be 54.3 inches. Because both the students fall in the same interval (10-11). Thus, we predict the same height for all students whose weight fall in the same interval. But the predicted height is average so there remains more or less difference of this predicted value with actual value. This shows that we cannot calculate accurate value but we can estimate that is approximately accurate. *The larger the deviation between actual value and predicted value the lesser the accuracy of predicted value.* This discrepancy is generally known as ***error***.

Figure 5.4: Diagram showing errors in prediction



Consider the regression line *Y* on *X* in figure 5.4, then the dots outside the line represent actual values of the *Y*- variable. The corresponding dots on the line show the average or best fitting points (predicted values) drawn with the principle of least-square method such that the average of the sum of square of their distance with actual values is always minimum than from any other line. These differences y_1, y_2, y_3, ... are then errors and the standard deviation of the errors is known on standard error of estimate of *Y*on *X*. If we take *Y* for actual value, *Y'* for predicted and *N* number of values then standard error of estimate of line *Y* on *X* becomes.

$$\sigma_{yx} = \sqrt{\frac{\Sigma\,(Y - Y')^2}{N}} \text{ ....(5.7)}$$

Similarly for line *X* on *Y*

$$\sigma_{xy} = \sqrt{\frac{\Sigma\,(X - X')^2}{N}} \text{ ....(5.8)}$$

Alternatively, we can write equivalent form for formula (5.7) and (5.8) in correlation form as given below

Y on $X \Rightarrow \sigma_{yx} = \sigma_y \sqrt{1\text{-}r_{xy}^2}$ ....(5.9)

In which, σ_y = Standard deviation of scores for *Y* distribution

$$= \frac{\sqrt{\Sigma\,(Y - \overline{Y})^2}}{N}$$

r_{xy} = Coefficient of correlation between variables *X* and *Y*

Similarly, standard error of estimate of *X* on *Y*

$$\sigma_{xy} = \sigma_x \sqrt{1 - r_{xy}^2} \text{ ....(5.10)}$$

Where, σ_x = Standard deviation of *X*-scores.

Note: When $r = 1$, from (5.7) we can see $(Y - Y') = 0 \Rightarrow \sigma_{yx} = 0 \Rightarrow$There is no error of prediction. When $r = 0$, $Y' = \overline{Y}$ for all values of *X*. In this case formula (5.7) becomes, $\sigma_{yx} = \sqrt{\dfrac{\Sigma\,(Y - Y)^2}{N}} = \sigma_y$

Therefore, the values of σ_{yx} ranges from 0 to σ_y as the correlation decreases from 1 to 0.

Example 5.3: The scores obtained by 10 students in two unit tests *X* and *Y* are given. Find regression equations and estimate standard error of *Y* on *X*.

Student	A	B	C	D	E	F	G	H	I	J
Test *X*	17	8	8	20	14	7	21	22	19	30
Test *Y*	9	13	7	18	11	2	5	15	26	28

Solution: On the basis of given scores, different components can be computed on the following way.

Students	X	Y	$X - \overline{X}$	$Y - \overline{Y}$	$(X - \overline{X})^2$	$(Y - \overline{Y})^2$	$(X - \overline{X})\times(Y - \overline{Y})$	Y'	$(Y\text{-}Y')^2$
A	17	9	0.4	-4.4	0.16	19.36	-1.76	13.714	22.1841
B	8	13	-8.6	-0.4	73.96	0.16	3.44	6.67	40.0689
C	8	7	-8.6	-6.4	73.96	40.96	55.04	6.67	0.1089
D	20	18	3.4	+4.6	11.56	21.16	15.64	16.06	3.7636
E	14	11	-8.6	-2.4	6.76	5.76	109.44	11.36	0.1296
F	7	2	-9.6	-11.4	92.16	129.96	36.96	5.88	15.0544
G	21	5	4.4	-8.4	19.36	70.56	8.64	16.85	140.4225
H	22	15	5.4	1.6	29.16	2.56	30.24	17.63	6.9169
I	19	26	2.4	12.6	5.76	158.76	195.64	15.28	114.9184
J	30	28	13.4	14.6	179.56	213.16	385.60	23.89	16.8921
$\Sigma =$	166	134	0	0	492.40	662.40		134.0	360.4594
$\Sigma/N =$	16.6	134			49.24	66.24		134	36.04

Calculation:

Standard deviation of *X* scores *(*σ_x) = $\sqrt{\dfrac{\Sigma\,(X-\overline{X})^2}{N}} = \sqrt{\dfrac{492.40}{10}} = 7.017$

Standard deviation of *Y* scores *(*σ_y) = $\sqrt{\dfrac{\Sigma\,(Y-\overline{Y})^2}{N}} = \sqrt{\dfrac{662.40}{10}} = 8.139$

Correlation between *X* and *Y* (r$_{xy}$) =

$$\frac{\Sigma\,(X-\overline{X})\,(Y-\overline{Y})}{N\boldsymbol{\sigma}.\boldsymbol{\sigma}} = \frac{385.60}{10\times 7.017\times 8.139} = 0.675$$

Predictor score of *Y* on *X* (y') =

$$r.\frac{\boldsymbol{\sigma}}{\boldsymbol{\sigma}_x}(X-\overline{X})+\overline{Y} = 0.675\times\frac{8.139}{7.017}(17-16.6)+13.4$$

Standard error of *Y* on *X* (σ_{yx}) = $\sqrt{\dfrac{\Sigma\,(y-y')^2}{N}} = \sqrt{\dfrac{360.45}{10}} = \sqrt{36.04} = 6$

Alternatively, $\boldsymbol{\sigma}_{xy} = \boldsymbol{\sigma}_x\sqrt{1-r_{xy}{}^2} = 7.017\sqrt{1-(0.675)^2} = 5.177$

and $\boldsymbol{\sigma}_{yx} = \boldsymbol{\sigma}_y\sqrt{1-r_{yx}{}^2} = 8.139\sqrt{1-(0.675)^2} = 6$

This shows that the result is same on using both formula but the alternative method is more convenient.

5.5 Cautions Concerning Estimation of Predictive Error

While we estimate standard error of prediction using regression line or predictive value, we must keep some points to work used method properly.

1. ***Linearity :*** The relationship between two variables *X* and *Y* must be liner. Otherwise, for most values of *X*, the estimated value of *Y* increase or decrease artificially.

2. ***Variability:*** The variability or scatterness of actual *Y* values about its predicted value (*Y*') must be same for all values of *X*. This means that the distribution must be homosedastic. Otherwise, the standard error of prediction approximate the

standard deviation of *Y* values only for intermediate values of *X*. In such case it will overestimate predictive error in *Y* for low values of *X* and underestimate for high values. It is clear from formula $\sigma_{xy} = \frac{\sqrt{(Y - Y')^2}}{N}$

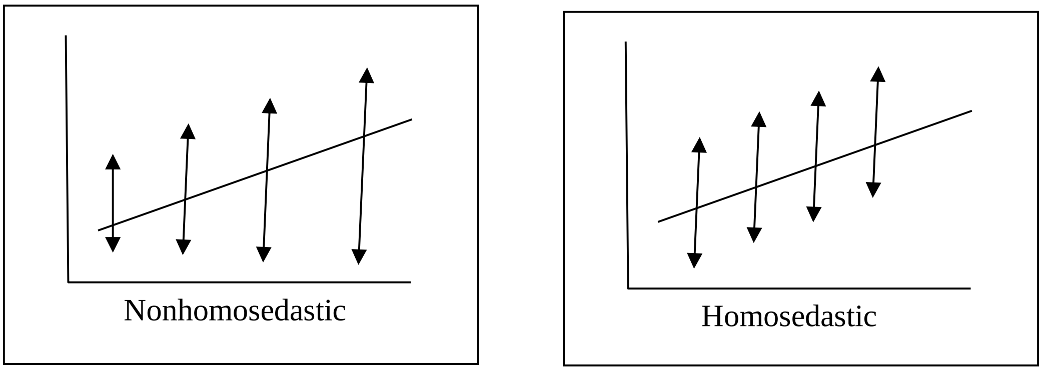


Figure 5.5: Diagram showing homosedastic and non-homosedasticdistribution of scores.

3. ***Normality:*** The actual *Y* score must be normally distributed for all values of *X*. Otherwise, we cannot determine certain proportion of cases above or below the given point. If the distribution is normal we can conclude that:

68% of the *Y*-scores fall within $\overline{Y} \pm 1.00\sigma_y$

95% of the *Y*-scores fall within $\overline{Y} \pm 1.96\sigma_y$

99% of the *Y*-scores fall within $\overline{Y} \pm 2.58\sigma_y$

Since standard error of estimate is a standard deviation of *Y*-scores about its predicted score, the above distribution of area under normal curve holds equally and we can write

68% of the *Y*-scores fall within $Y' \pm 1.00\sigma_{yx}$

95% of the *Y*-scores fall within $Y' \pm 1.96\sigma_{yx}$

99% of the *Y*-scores fall within $Y' \pm 2.58\sigma_{yx}$

Thus, only in the condition of normality the prediction of one variable on the basis of another is more accurate.

4. ***Partiality:*** The estimation of standard error of regression do not cover the errors due to sampling. In many situations on research we do not work with the total population but with the sample. If we work with total population for correlation, regression, standard error of estimate, etc. then the result will be more accurate than the result from sample. To reduce this effect we can use method of sampling error. Therefore, the above procedure of estimating error is partial but not total.

5. ***Sampling Variability:*** The variability in sampling influence the size of error. In other words the additional variability due to sampling variation extends the limits within which actual *Y* values fall. The limits extends less for central values of *X* and more for those values far from *X*. If larger sample and more accurate procedure are used the interval that contain central 95% of actual *Y*-value would wider then in small sample. Therefore, predicting and estimating the error of prediction are best done when the sample size are large enough to reduce the margin of error to a tolerable amount A larger sample size also makes it easier to determine whether the conditions of linearity, homosedasticity and normality are reasonably met. A sample with $N = 100$ is really rather small for these purposes (Minium, et al., 2010, p. 187). The larger the sample, the less the sampling variation.

Exercise 5.2

1. In the following correlation table, compute coefficient of correlation and regression equation. Also show the regression effect by calculating the regression equation in standard - score form. For *X*'s of $\pm 1.0\sigma$ and $\pm 2.0\sigma$ from the mean in arithmetic, find the corresponding score in σ-score form.

Group I.Q. Arithmetic	≤ 84	85-89	90-94	95-99	100-104	105-109	110-114	115-119	110-124	125≤	Totals
90 & L				3	3	15	12	9	9	5	56
85-89				8	17	15	24	13	6	6	89
80-84			4	6	22	21	20	10	5	1	89

75-79			7	25	33	23	10	7	4		109
70-74		4	10	18	14	22	12	1	1		82
65-69	1	3	3	18	7	8	8	1			43
60-64			2	5	3	1	1	-			12
Total	1	7	26	77	99	105	87	41			480

Exercise (Model)

Multiple Choice questions

1. If the value of correlation is zero, then the value of predicted standard score is

   (a) 1 (b) 0 (c) -1 (d) None of the above

2. If we take *Y* for actual value, *Y*' for predicted, and *N* for number of values then standard error of estimate line *Y* on *X* becomes:

   (a) $\sigma_{xy} = \sqrt{\frac{\Sigma\,(X - X')^2}{N}}$ (b) $\sigma_{yx} = \sqrt{\frac{\Sigma\,(Y - Y')^2}{N}}$

   (c) $\sigma_{xy} = \sigma_x \sqrt{1 - r_{xy}{}^2}$ (d) $\sigma_{yx} = \sigma_y \sqrt{1 - r_{xy}{}^2}$

Short answer questions:

1. Define regression analysis and state its relation with correlation coefficient.
2. Derive regression equation in standard score form.

Long answer questions:

1. Discuss the conditions that must be fit fulfilled for best fit of predicted value.
2. Find the regression equation and estimate of standard error of *Y* on *X* using following data.

Student	A	B	C	D	E	F	G	H	I	J
Test *X*	17	8	8	20	14	7	21	22	19	30
Test *Y*	9	13	7	18	11	2	5	15	26	28

Chapter 6

Inferential Statistics

a. About the Chapter

In previous chapters we deal with the descriptive statistics in which the data obtained from particular group of individuals observed or studied were used to derive the conclusion and the result was limited for the generalization to that group only. The purpose of descriptive statistics is to organize and summarize: observations so that they will be easier to comprehend and describe for particular situations. But in many situations in education, we cannot work with total population and use representative sample to infer the result for that population. The statistics that is used to infer the result for population is divided into two types: parametric and non-parametric. The whole chapter covers these two types of tests and is separated into three sections: The first section describes the concept of inferential statistics in parametric and non-parametric form. The second section is provided with the concept, characteristics and derivation of variants of parametric tests: *t*, *z* and *F* test (One way and two ways). The third section explains the concept and computation of χ^2 as a form of non-parametric test.

Particularly, on complete study of this chapter the following specific objectives are assumed to be achieved.

- To clarify the concept of inferential statistics.
- To point out the differences between parametric and non-parametric tests.
- To use *t*, *z*, and *F* test for making inferences about parametric data.
- To apply chi-square test for making inferences about non-parametric sets of data.

6.1 Concept

In previous chapters we discussed about descriptive statistics in which the computed value were need to describe the properties of particular samples. That is, the results obtained from measure of central tendency, measure of spreadness or variability, measure of relative position and measure of correlation ship (Product-Moment) were used for the purpose of description of certain traits of subject from which the data were taken. It means that the descriptive statistics limit on generalizing the result to particular group of individuals observed or studied. However, in many situations in education and social science we do not work with total population but with representative sample from it and derive the result for inferring population parameter. The whole section is devoted to clarify the concept of statistics that uses data from sampling. Therefore, the branch of statistics that allows to generalize or infer about unknown data (population) from known data (sample) is inferential statistics. Among various methods of such an inferential statistics included in present course are *t*-test, *z*-test, *F*-test, and χ^2-test.

The inferential statistics are divided into two parts. Parametric and non-parametric. The statistics applied to situations in which assumption of normality is tenable, i.e. variables are distributed normally are parametric. In such statistics the representative sample is also assumed to be normal, as the population from which it is drawn is normal. Most of the variables like intelligence, personality, achievement, etc. with which we work in education are assume to follow the pattern of normal distribution. The statistics under parametric category is significance of means (*t*-test, *z*-test, and *F*-test). Such statistics work under following assumptions.

1. Population from which sample have been drawn should normally distributed. This can be known by the assumption of normality.
2. Variables measured must be in interval or ratios scale.
3. The observation must be independent. That is the scores in one test do not influence scores on another test or the inclusion and exclusion of any case in sample should not affect the result.
4. Should met the criteria of homosedasticity. That is the population must have same variance or known ratio of variance.

On the other hand, non-parametric tests are those statistical tests in which the assumption of normality in a distribution is not tenable. When the distributions are plotted graphically they will show gross deviation from a normal curve and are influenced by different factors including chance variation. So parameters like mean, standard deviations, etc. can not be used to describe the distributions. The distribution pattern of variables is not uniform so such tests are also called distribution free test. Parametric tests are not suitable for such distribution in that these distributions are generally based on frequency data (nominal and ordinal. Therefore, the branch of statistics dealing with distributions in which assumption of normality is not tenable is called non-parametric.

Different tests like χ^2-test, contingency coefficient, rank-order correlation (ρ), sign test, median test, the Mann-Whitney *U* test, Run test, and Kolmogorov Sminrov (*KS*) two sample test are the statistics that fall under the category of non-parametric tests. The non-parametric tests can be used under following assumptions.

1. When the assumption like the normality of the distribution of scores in the population are doubtful. That is, the distribution of scores are free or values are not distributed in a certain pattern.
2. When the variables are measurable in the form of nominal or ordinal scale. That is, variables are expressed in frequency data.
3. When *N* is quite small. That is, if the size of the sample is as small as *N* = 5 or *N* = 6 then there is no alternative of use of non-parametric test.

Note: Though non-parametric tests are simpler and easier to compute, their use should be restricted even in case of failure of the conditions for the parametric tests. This is because of the non-parametric tests being less powerful to detect a true difference than parametric test in some situations. However, the use of non-parametric test in statistical inference cannot be underestimated.

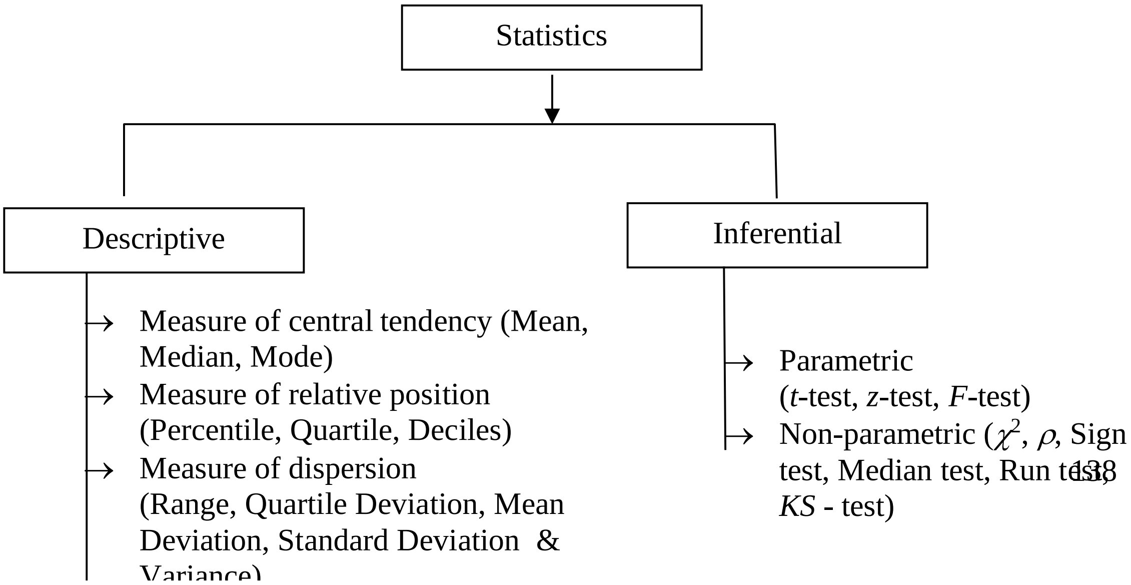

Statistics
Descriptive
Inferential
→ Measure of central tendency (Mean, Median, Mode)
→ Measure of relative position (Percentile, Quartile, Deciles)
→ Measure of dispersion (Range, Quartile Deviation, Mean Deviation, Standard Deviation & Variance)
→ Parametric (t-test, z-test, F-test)
→ Non-parametric (χ², ρ, Sign test, Median test, Run test, KS - test)

Figure 6.1: Diagram showing relation between types statistics

b. Parametric Tests

As discussed above parametric tests for inference purpose are available in different forms. Under this section the standard error of means, the significance of the difference between two means, and the significance of the difference between more than two means at a time will be discussed.

6.2.1 The Standard Error of Means

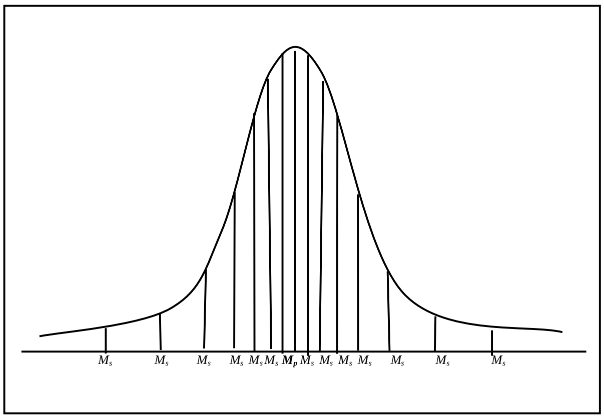
The nature of errors rely on the number of samples that taken from population so we discuss the standard error of means in two different cases below.

Case I: The standard error of means for large samples ($N \geq 30$)

If we compute mean from population then the result is accurate and there is no error. However, in real life situation it is not possible to work with all the population so that we need to select certain amount of sample representative to population. If the sample is more representative there is less error in sampling mean and if the sample is less representative then there is more error in sampling mean. Therefore, sampling error is the difference between population mean (M_p) and sample mean (M_s). The difference or sampling error may occurs from variation due to chance selection of individuals and the error need not to include the result of mistakes in sampling procedure.

If we take large number of equal samples from the population and measure the mean value of each sample, the mean value will not be identical. Some relatively high and some relatively low from population mean but many of them cluster around the M_p (Figure 6.2)

Figure 6.2: Normal curve showing the sample mean (M_s) around population mean (M_p)



The distribution of sample mean they generated is known as a random sampling distribution of mean. As the size of the sample increases, the distribution becomes almost normal and the average means of the sample means will be approximately the same as population mean ($M_p = M_s$). In such situation 95% of all the sample mean fall between ±1.96 standard error of population (*S.E.*) (and also of standard error of sample means (SE_M). That is, 95% chance of falling single sample mean is within limit of population mean. Therefore, the standard deviation of the random sampling distribution; called *standard error of mean*, depends on the standard deviation of population (σ) and sample size. (N)

$$\therefore SE \text{ or } \boldsymbol{\sigma_p} = \frac{\sigma}{\sqrt{N}} \text{ .... (6.1)}$$

However, as we are usually unable to ascertain the σ of the total population, the standard deviation of samples (S) is used instead. In such case equation (6.1) becomes,

$$SE_M = \frac{S}{\sqrt{N}} \text{ ....(6.2)}$$

This clearly shows that the smaller the standard deviation of sample (S), the smaller the sampling error. The larger the N, the smaller the sampling error. As it is difficult to influence the size of *S*, though it is possible to increase the size of *N* to reduce standard error. When standard deviation is very large, then *N* too needs to be very large to counteract the error. It is suggested that, unless there are very unusual distribution, the sample of 25 or greater usually yield a normal sampling distribution of mean.

Example 6.1: Assume the mean of the intelligence test scores of a sample of 100 students in a school is 30 and standard deviation is 5. how dependable is this mean with population mean ?

Solution: We have, the standard error of mean (SE_M) = $\frac{S}{\sqrt{N}}$ = $\frac{5}{100}$ = 0.5

This SE_M can be thought of as the standard deviation of distribution of sample means around population mean. In case of large samples, the sampling distribution of sample means is assumed to be normal. From Figure 6.3, it can be seen that the sampling distribution is centered at the unknown population mean with $\sigma = 0.50$.

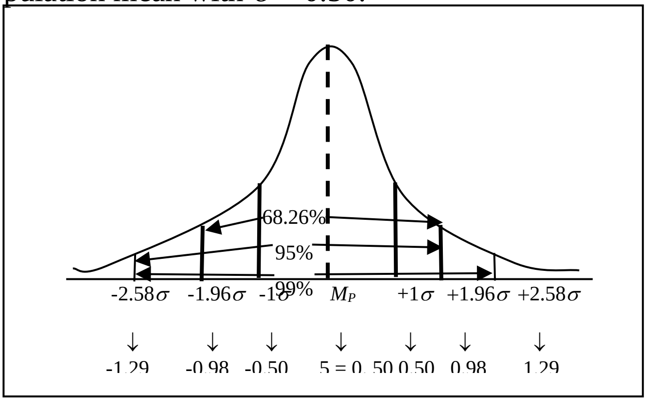


Figure 6.3: Variability of sampling distribution of means around population means.

The sample means fall equally on positive and negative side of population mean. Exactly 68.26 percent of sample means will fall within $\pm 1\sigma = \pm 1 \times 0.5 = \pm 0.5$. Furthermore, 95 percent of the 100 samples will fall between $\pm 1.96\sigma = \pm 1.96 \times 0.5 = 1.29$. This shows that the probability of 0.95 that our mean of 30 does not miss the population mean by more than ±0.98. Similarly, the probability of 0.99 that our sample mean of 30 does not miss the M_p by more than ±1.29. Since the magnitude of the probable deviation of sample mean from its population mean is the measure of probability with which we are able to estimate the M_p from the sample. In both cases the magnitude is under the range of M_p, the M_s is dependable to M_p.

Case II: The standard error of means for small samples (N < 30)

For small sample we may estimate the value of population mean using the following formula $SE_M = \frac{S}{\sqrt{N}}$

In which, S = Standard deviation of small sample.

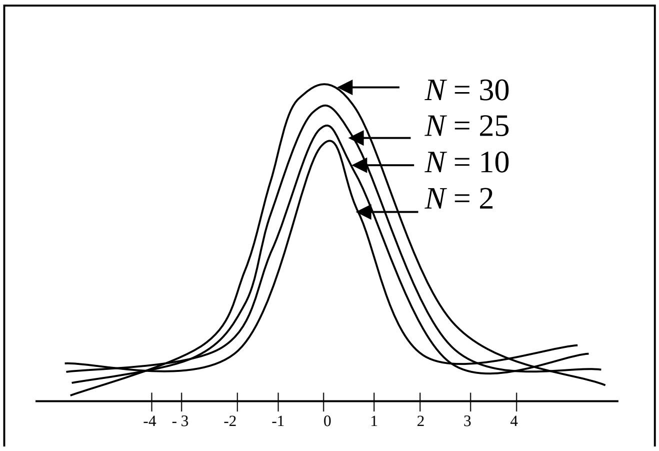
N = The number of cases in sample.

Since the standard deviation of random sample (S) is smaller than the corresponding population standard deviation (σ) for small N, then to correct the underestimated value and to get a better approximation to σ, we should compute the SD of a sample by the formula $S = \sqrt{\frac{\Sigma x^2}{N-1}}$ instead of usual formula $SD = \sqrt{\frac{\Sigma x^2}{N}}$. If the SD has been computed with N in the denominator then the same correction formula should be use in SE_M,

i.e., $SE_M = \frac{\sigma}{\sqrt{N-1}}$ ……………………………………(6.3)

For small samples, it is not necessary that the sampling distribution of means to be normal. In 1815, William Sedy Gosset of Ireland developed the concept of small sample size. He found that the distribution curves of small samples means were somewhat different from the normal curve. Such distribution was named as *t*-distribution. When sample size is small, the *t*-distribution lies under normal curve but the tails curve are higher than the corresponding parts of the normal curve. The relation between normal curve and *t*-distribution with the change of N can be seen clearly from Figure 6.4.

Figure 6.4: Diagram showing the effect of increase *N* in *t*-distribution.



For small samples it is necessary to consult the *t*-table from Appendix C for the tabulated value of selected points. To use *t*-value we need degree of freedom (*df)*. The concept of *df* will make clearer in the following section. As *N* increases the *t*-value approaches *z*-values of the normal probability curve. This can be use from another table in Appendix D.

Degrees of Freedom (df): It is a number of values that can be chosen freely. For example, if we deal with two samples X_1, X_2 whose mean is considered to be 20. If X_1 is 8 then $(X_1 + X_2)/2 = 20$ implies $X_2 = 32$. In this case we are free to choose only one value of $X_{2,}$ i.e., 32. In two samples, therefore, the degree of freedom (*df*) is 2-1 = 1. For *N* samples, the *d.f.* = *N*-1.

Similarly, we can consider another example to be more clearer. Suppose we have 5 scores : 1, 2, 3, 4, and 5 then their mean is 3. The deviations of these scores from mean are -2,-1, 0, 1, 2. And the sum of all the deviations equals zero. Keeping zero constant we can choose independently (freely) only four numbers and cannot choose last number to make sum zero in our desire. This means that first four numbers are changeable according as our interest and fifth number is fix to make sum zero. The degrees of freedom, therefore, is 5-1=4.

When a statistics is used to estimate parameters the number of *d.f.* depends upon the restriction placed upon the observation. One *d.f.* is lost for each restriction imposed. In estimating M_p from M_s the *d.f.* is *N*-1. Therefore, the *d.f.* will vary from one statistics to another or from one case to another.

Confidence Intervals and Levels of Confidence : Suppose large number of equal samples from a population are drawn and mean for each sample is computed then the means are distributed normally around population means. Using the concept of normal curve and statement of central limit theorem (it states that *if random large samples of equal size are repeatedly drawn from any population, then the mean of those samples will be approximately normally distributed*) the *Ms* has 95% chance of being within ± 1.96σ units from *Mp*. Similarly, *Ms* has 99% chance of being ±2.58σ units from *Mp*. As sample mean assume population mean in normal distribution *Ms = Mp* and the above condition can be stated as 0.95 probability a sample mean falls within *Ms* ± 1.96 and for 0.99 probability that a sample mean falls within M_S ±2.58σ units from *Mp*. Then the range between (M_s-1σ)

and (M_s + 1σ) are called *confidence interval* or *fiduciary limits* and the confidence value for defined interval is called *fiduciary probability*. The value outside these limits is called *level of confidence*. For M_S ± 1.96σ confidence interval the level of confidence is 5% or 0.05 level. Similarly, for M_S ± 2.59σ confidence interval the level of confidence is 1% or 0.01 level. The 0.05 level of confidence indicates that the probability 0.95 in which M_p lies within interval M_s ±1.96σ and 0.05 that it fall outside these limits. Same reasoning applied for 0.01 level of confidence.

Example 6.2: The scores obtained by 5 children in a test of perception are 5,5, 7, 4 and 4. Determine the 0.95 and 0.99 confidence intervals for the population mean.

Solution: The mean of the scores obtained in perception test is (5+7+7+4+4) / 5 = 5 = M_s
Standard deviation of the given score can be computed as:

X	$x = X\text{-}M$	x^2
5	0	0
5	0	0
7	2	4
4	-1	1
4	-1	1
	$\Sigma x = 0$	$\Sigma x^2 = 6$

We know that the formula of standard deviation for small sample

$$S = \sqrt{\frac{\Sigma x^2}{N-1}} = \sqrt{\frac{6}{5-1}} = 1.5$$

Standard error of mean (SE_M) = $\dfrac{S}{\sqrt{N}} = \dfrac{1.5}{\sqrt{5}}$ = 0.67

Degree of freedom for 5 numbers (df) = N - 1 = 5-1 = 4

From t-table in Appendix C, the tabulated value of t = 2.78 for 4 *d.f.* at 0.05 level of confidence. Similarly, t = 4.60 for 4 df at 0.01, level of confidence. From first value of t, we know that 95 out of 100 sample means will fall within ±2.78 SE_M = ±2.78×0.67 = ±1.86

of the population mean and 5 out of 100 fall outside these limits. Therefore, the probability is 0.95 that our sample mean 5 does not miss M_p by more than ±2.78×0.67 = ±1.86.

Similarly, for t = 4.60, 99% of our sample means will lie between ± 4.60 SE_M = ± 4.6×0.67 = ±3.82 of M_p and 5% fall outside these limits. Therefore, the probability is 0.99 that the sample mean 5 does not fall outside ±4.60×0.67 = ± 3.82.

Now, for 0.95 confidence interval the limits $M_s \pm 2.78\ SE_M = 5 \pm 2.78 \times 0.67 = 5 \pm 1.86$. Therefore, the probability is 0.95 that M_p is not less than 3.14 and not greater than 6.86. Similarly the probability is 0.99 that M_p is not less than 1.18 and not greater than 8.82 ($1.18 \leq M_p \leq 8.82$).

6.2.2 Significance of the Difference Between Means

Concept of Standard Error of Difference : In previous section we discussed about the significance of the mean. That is, if we select large number of samples from given population, compute their means, and arrange these means into frequency distribution, the result is normal curve (for large sample means) and *t*-distribution curve (for small sample means).We also know that the standard deviation of the distribution of these sample means as a standard error of means (SE_M) and this SE_M was taken as yardstick for testing the significance of a given sample mean.

But another case arises, if we have data related to difference between means of a number of samples. If we arrange the distribution of these difference between sampling means in frequency distribution the result is also a normal for large sample and *t*-distribution for small sample. The standard deviation of the distribution of these differences is the standard error of the difference between means. In this case too the error can be taken as a tool to test the significance of the difference between means. Suppose we have to find the difference in achievement between the boys and girls. For this, we may select large number of samples of boys, administer achievement test, and compute mean of their achievement scores. Apply same procedure for girls group and find difference between these groups. Suppose the mean of the boys is 22 and that of the girls is 20. Without some basis we cannot say anything about the observed difference of 2. For this we need to decide if the difference between means of boys and girls score is actual or due to some sampling error. To estimate error we must have standard error of the difference of two sampling means. Then with the help of the observed difference of two means and its standard error we can decide whether the difference actually exists between the population means.

Hypothesis: As discussed in previous section parametric tests are powerful tools to infer or to generalize the results obtained from sample to parent population from which sample are drawn. But the decision making about the characteristics of the population on the basis of study sample may lead wrong decision if the sample is not representative of the population. Suppose, we may want to decide if the average achievement of a group pre and post the treatment (say use of slide to teach particular content) is really different. Without same basis we cannot decide though these two achievements are different roughly. The method of statistics which helps in arriving at the logical conclusion for such decision is called *statistical decision making* or *hypothesis testing* or *test of significance* or *test of hypothesis.*

Generally, hypothesis is an assumption that we make about population parameter. It is drawn logically on the basis of available evidence regarding any parameter of the population. Here, the measures descriptive of population are called *parameters* and the measures computed form sample are called *statistics*. The greater the representativeness of population the lesser the difference between parameter and statistics. Therefore, the degree to which a sample mean represents its parameter is an index of the significance or trustworthiness of the computed sample mean. In hypothesis we assume regarding the discrepancy between the result of these two conditions.

Types of hypothesis: Theoretically, there should be only one type of hypothesis - research hypothesis, the basis of our investigation. However, because of convention on research and because of convenience in testing it can be classified basically in two types: research and test hypothesis.

The research hypothesis is a speculative statement subjected to verification through a research study. It is formulated before data collection and guide the whole research activities as a major statement. But for the purpose of statistical testing an investigator write test hypothesis in null (H_0) and alternate form (H_1) and tries to reject null form to keep the decision strong. If the researcher fails he/she accept null hypothesis. The null form states that there is no relation between variables or there is no difference between the result of sample mean and population mean or there is no actual difference between two means even though they are different in number. And if the null form fails, the alternate form (H_1) is accepted. In this case the research hypothesis is accepted without any change. Consider an example: Effects of different types of incentives on pupil achievement.

The null form is

H_0 : There is no effect of incentive on achievement

And the alternate forms are

H_1 : Incentive influence pupil achievement (Non-directional).

: Incentive increase pupil achievement (Directional).

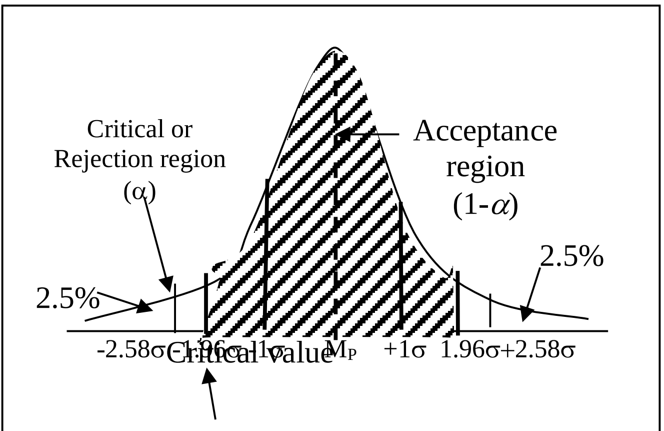
In above example, the H_0 claims no relation between two variables in which first is incentive and second is achievement. Researcher tries to reject the null form and if fails, he must accept either of the alternative forms. The first statement of alternative form is assuming that there is relation between variables but without specific direction. Therefore, it is non-directional in nature. But the second statement of the alternative forms is expecting positive effective of incentive upon achievement so it is directional. Researcher decides the nature of alternative hypothesis according as the need of problem and applies statistics (one tail or two tails) accordingly.

To conclude this problem an investigator may use any of the several techniques. One of these he may use pre-test post-test control group design. For this, he can take certain number of pupil as a representative sample of whole pupil adopting certain method. Then he may record the achievement score before and after giving incentive. Suppose the results (means) from two test are slightly different then the question arise whether the difference between two means indicates really valid difference which will help drawing conclusion or the difference is the result of sampling fluctuation which have occurred due to error ? To answer this question researcher first need to find the standard error of means (SE_M).

Level of Significance: For decision making, an investigator has to set certain arbitrary standard or accepting or rejecting a null hypothesis. Commonly used standards are 5% (0.05) and 1% (0.01). If we use 5% as a basis for decision, it means that out of 100 cases 5 cases are likely to reject null hypothesis or we are 95% confident that the decision of accepting H_0 is correct

Similarly, if we set standard of 1%, we are confident that our decision of accepting H_0 is 99% correct. These standards chosen arbitrarily by researcher are called *level of significance*. The rejection or acceptance of null hypothesis depends upon the level of significance adopted.

Figure 6.5: Diagram showing different areas and values under normal curve.



If we set 5% level of significance for decision making about null hypothesis, then it takes the value of ±1.96σ in base line of normal curve. That is if we take ±1.96σ from mean it covers 95% area under normal curve and left remaining 5%. If the significant value obtained by our calculation is greater than the value at 5% then we reject null hypothesis assuming the difference significant. It is small then we accept it with 95% confidence.

The region of standard normal curve corresponding to a predetermined level of significance is called *critical region or* ***rejection region***. The region under normal curve not occupied by rejection region is called ***acceptance region*** (shadowed area in Figure 6.5). When the computed test statistics lies in the acceptance region, we believe the hypothesis is true and accept null hypothesis. Otherwise, we reject null hypothesis. The value of test statistics which separate acceptance region from rejection is called ***critical value***.

Note: For large samples ($N \geq 30$) we use critical value from z-table and for small sample we use critical value from t-table.

One-tailed or two-tailed test: From hypothesis, we know that the null hypothesis is a statement that states no difference in relation between variables. In option of null hypothesis, we set another hypothesis namely alternate hypothesis (basically it is equivalent with research hypothesis). The alternate form takes two pattern of hypothesis: directional and non-directional. When we are hypothesizing a direction of difference, rather than the more existence of difference, we make use of one tailed test. For such test 1% or 5% area of rejection is either at the upper tail or at the lower tail of the curve. Remember previous examples:

"Incentive improve achievement of pupil."

“Over TV watching decrease achievement of pupil".

In both cases, we use one-tail or one sided test. To conclude these statements we may set two groups (control and experimental), provide treatment (incentive or opportunity to TV watching) to experimental and compute the means of experimental and control group separately, say, M_1 and M_2. Then we test hypothesis for $M_1 > M_2$ or $M_1 < M_2$ instead $M_1 \neq M_2$. When $M_1 > M_2$ we use right tailed test and when $M_1 < M_2$ we use left tailed test.

We use two tailed test if we are only concerned with the absolute magnitude of the difference between means regardless of sign. The above examples may restated as:

"Incentive influence achievement of pupil".

“Over T.V. watching influence achievement of pupil".

In both examples we need not obtain the value of difference with proper sign but merely the significant difference. In such case we use two-sided test and write the alternate hypothesis as $M_1 \neq M_2$. The decision whether to use one-sided or two-sided depends upon the nature of given statement in our problem. If we take one-sided or two-sided test for our decision then we use critical values on respective column of z-table in Appendix D.

Size of Sample and Decision Making: As we know that the sampling distribution of the differences between means may look like a normal curve (for large sample) and look like *t*-distribution curve (for small sample). If $N \geq 30$, then the distribution of differences between means will be a normal one. In such case the value of the standard error used to determine the significance of the difference between means will be in the forms of standard scores z-scores). Therefore, we use *z*-test for large samples and use the critical value or tabulated value of standard error for 0.05 or 0.01 level of significance respectively. Since the curve obtained from sampling distribution of differences for large sample is normal then there is no effect of degree of freedom so we need not *df* for large sample.

On the other hand, when $N < 30$ the curve of distribution becomes like *t*-distribution and it will vary with the number of *d.f.* In this case small sample statistics to which we called *t*-statistics or *t*-test are applicable for decision making. We obtain the critical value using certain degree of freedom from *t*-table at 0.05 or 0.01 level of significance. This value helps us to decide for rejection or acceptance of null hypothesis.

For decision making, we reject, null hypothesis if our calculated *z* or *t*-statistics is greater than the critical value of *z* or *t* obtained from their respective columns in the tables given in Appendices. In this case we accept alternative hypothesis. Again, if the calculated value lies within the range of standard or critical value then we accept null hypothesis.

***Errors in Decision Making*:** In testing hypothesis or making decision, researcher generally commit two types of errors: *Type I* and *Type II* errors.

Type I errors are made when we reject null hypothesis by marking a difference significant, although there is no actual difference between two means, such type of errors is also called **α-error**.

The difference is marked significant when the gap between two sample means signifies a real difference between the parameters of the population from which sample were drawn. As already stated, before making judgment significant or non-significant we must choose some relative value or probability scale which helps to separate these two categories. Researcher may select 0.01, 0.02, 0.03, 0.04, 0.05, etc. as the arbitrary standards for decision making to which we called level of significance. But the most often used levels of significance on education are 0.01 and 0.05. This means that, the rejection or acceptance of null hypothesis or decision making regarding hypothesis depends upon the level of significance chosen as an standard. If a researcher takes 0.05 or 5 percent level of significance or decision making and rejects null hypothesis at that standard, then he is accepting risk of 5% that his decision may be wrong. In this case researcher is 95% confident that the null hypothesis is wrong and only 5 chances out of 100 it may be true. Similar argument apply for 1% level of significance. The hypothesis is rejected at 5% level of significance may be accepted at 1% level of significance. Therefore, if we set up a wider confidence interval, then we reduce the possibility of α-error (i.e. possibility of α-error is less in 0.01 level than in 0.05).

Type II errors are made when we accept null hypothesis by marking a difference not significant, when there exist a true difference between two means. Such errors are also called ***β-errors***. When we set up more wider confidence interval we may commit this type of error (i.e. possibility of β-error is more in 0.01 level than in 0.05 level).

The summary of these two types of errors can be presented in the following decision table.

Table 6.1: Possibility of different errors in decision making

Actual	Decision	
	Accept H_0	Reject H_o
H_0 is true	Correct (No error) Probability = 1 - α	Wrong (Type I or α-error) Probability = α
H_0 is false	Wrong (Type II or β–error) Probability = β	Correct (No error) Probability = 1-β

While making decision about to accept or reject null hypothesis, we should try to minimize α and β errors. But the probability of β-error is much more risky than α-error because we accept false hypothesis with β-error. Therefore, our concern should be to minimize β-error when the decision is more critical. In this case we may take relatively narrow confidence interval or significance level (i.e. 0.05).

The effort to reduce one type of error increase the possibility of another type of error. Therefore, researcher must decide which kind of error or wrong inference he should avoid. *Type I errors have more serious effect in many research program than Type II errors*. If a researcher claims a significant difference erroneously and terminate the program with low level of significance (i.e. 0.05) then the program will be futile when high level of significance (i.e. 0.01) is demanded. Similarly, *β-errors must be watched carefully whom experimental factors are potentially dangerous*. For example, if one is studying the psychological effects of a drug and decides that there is no effect even when there exists positive effect (β-error) then the result would have disastrous (because it assumes psychological effect as an actual effect). Similarly, β-error is more important when one have to challenge the result previously established. Therefore, *in most experimental research it is wise to set up a significance level of at least 0.01 whereas in most preliminary works the satisfactory significance level is 0.05*. However, a researcher or an experimenter always should try to avoid errors in drawing inferences. This can be done by demanding more evidence for certain result when the significance of the result is doubtful or uncertain. These evidences are possible with the effort of obtaining additional data (by more trails), repetition of experiment, and controls of variables.

From above discussion, it is clear that the probability of accepting false null hypothesis is β then obviously the probability of rejecting false null hypothesis is 1-β. That is; 1-β is the probability of claiming significant difference when a true difference really exists. The value of 1-β is called *power of the test*. Hence, β and power of test are complementary, i.e. any condition that decreases β increases the power of the test, and vice versa. Therefore, among several ways of conducting a test, the one good way is increasing the power of test by decreasing the value of β. Several factors are accountable to influence the power of the test.

- *The discrepancy between true mean and hypothetical mean*: The larger the discrepancy, the greater the power.

- *The size of the sample*: The larger the sample size, the greater the power. Researcher can increase the size of the sample by keeping other conditions constant.
- *The standard deviation of the variable:* The smaller the standard deviation, the greater the power. Standard deviation can be reduced by improving the reliability of the measuring instrument.
- *Relation between samples:* Dependent or correlation samples can increase power. Therefore, power can be increased by choosing correlated groups.
- *Level of significance:* The larger the value of α, the lower the value β and the greater the power. Although it is possible to change the value of α for desired value of β, it is not better way. Rather one ought to set up the value of α maintaining the previously stated condition and should improve accuracy through sample size and standard deviation.

Independent and Dependent Means: Two means are said to be independent or uncorrelated when these means are computed from samples drawn randomly from totally different or unrelated groups.

However, the means are said to be dependent or correlated when the values are obtained from related groups. For example, if we compute means from the scores obtained by two different groups, say experimental and control group, at the same time the means are independent. But if the results are computed by the use of scores obtained by same group in two different times (say in pre-test post test design) then two means are dependent. Similarly, if two groups are different but the individual are matched one-to-one by one or more characteristics to make equivalent samples and same test is administrated, then the means so obtained are also dependent.

6.2.3 Test of Significance of Difference between Two Means

The appropriate selection of statistics to test the significance of difference between two means depend upon the number of samples from which means are derived and the dependency of variables. Therefore, we categorized these tests on these two basis and test the significance of difference in the following ways:

Case I: Significance of the difference between two uncorrelated (independent) means ($N \geq 30$)

Example: Test the significance of the difference in means of scores obtained by two groups taught by different methods.

Group	Mean	*SD*	Student
A	43	8	65
B	30	7	65

Solution:

Step I: Hypothesis formulation

Since we have to test the difference without any specific direction then we can use two-tail test and formulate hypothesis accordingly. If we denote null hypothesis by H_0 and alternative by H_1 then

H_0 : There is no significant difference between the means taught by two methods, i.e. $M_1 = M_2$

H_1 : The difference between two means is significant, i.e. $M_1 \neq M_2$

Step II: Computation of Statistics

Since $N > 30$, we apply the formula for large samples to estimate standard error of difference (SE_D),

$$SE_D = \sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}} \text{ ....(6.4)}$$

$$= \sqrt{\frac{(8)^2}{(65)^2} + \frac{(7)^2}{(65)^2}} = 1.32$$

Now, the value that represents a point of sampling distribution on base line under normal curve (z) = $\frac{\text{Difference between means}}{\text{Standard error of difference}} = \frac{M_1 - M_2}{SE_D}$

$$\therefore \text{z} = \frac{43\text{-}30}{1.32} = 9.85$$

Step III: Decision making

Since the samples are large, the values of z from z-table are 1.96 and 2.58 for 0.05 and 0.01 level of significance respectively. The computed value |z| = 9.85 is greater then the tabulated value for standard normal curve, i.e. $|z|_{cal} > z_{tab}$, we reject null hypothesis for both the level of significance. That is, the mean from two methods are significantly different, or the difference is not merely due to chance factors or due to sampling fluctuations.

Note: If we state alternative hypothesis as ‘the mean of group A is significantly higher than the mean of group B’ then we must use the value of z for two tailed test that is available in another column in z-table and make decision accordingly.

Example 6.4: A science teacher divides his class into two random groups. He uses special method to teach experimental group hoping that such a method will promote special skill in science and teach another group using traditional method. He found following data after taking achievement test.

Group	Mean	*SD*	Number
Experimental	35	4	48
Control	30	3	45

Is the difference between means significant to prove the effectiveness of new method ?

Solution: In this problem teacher is interested to bring positive change through new method so the problem clearly indicates the direction. Therefore, we formulate directional hypothesis to apply one-sided test.

Step I: Formulation of hypothesis

H_0: There exists no real difference between the means of two samples.

*H*1: The achievement of experimental group is significantly higher than control group.

Step II: Computation

Since *N* > 30, and two groups are different we can use formula or z test to compute the standard error of difference between two independent means (same as in previous example), Then,

$$z = \frac{M_1 - M_2}{\sqrt{\frac{S_1^{\ 2}}{N_1} + \frac{S_2^{\ 2}}{N_2}}} = \frac{35-30}{\sqrt{\frac{4^2}{48} + \frac{3^2}{45}}} = \frac{5}{0.73} = 6.85$$

Step III: Decision Making

Since the tabulated value of *z* for one-tailed test are 1.65 and 2.33 at 5% and 1% level of significance respectively, then the computed value is greater than these values in both cases. We reject the null hypothesis at both level of significance and conclude that the new method used by science teacher is more effective than old ones.

Case II: Significance of the difference between two independent means (N < 30)

Since *N* < 30, the sampling distribution makes *t*-distribution. Therefore we use the new formula to estimate standard error of difference, use the degree of freedom to find an appropriate tabulated value of *t* (since for small sample the critical value varies as the *df* changes), and compute *t*-value using difference in mean and error of difference. Other process is same as in large samples.

Example 6.5: Test the significance of difference of following data obtained by two groups A and B in attitude test in which each group consists of 10 students.

Group A	10	9	8	7	7	8	6	5	6	4
Group B	9	8	6	7	8	8	11	12	6	5

Solution:

Step I: Hypothesis formulation

*H*0: The difference between the means of two groups is not significance.

H_1: The attitude of two groups are significantly different.

Step II: Computation

To obtain the requirement for mean and standard deviation we proceed in the following way. Let the scores of first group be X_1 and second group be X_2 then

X_1	X_2	X_1 - M_1 = x_1	X_2- M_2 = x_2	x_1^2	x_2^2
10	7	10-7 = 3	7-8=-1	9	1
9	8	9-7 = 2	8-8=-0	4	0
8	6	8-7 = 1	6-8=-2	1	4
7	7	7-7 = 0	7-8=-1	0	1
7	8	7-7 = 0	8-8 = 0	0	0
8	8	8-7 = 1	8-8= 0	1	0
6	11	6-7 = -1	11-8= 3	1	9
5	12	5-7 = -2	12-8= 4	4	16
6	6	6-7 = -1	6-8= -2	1	4
4	5	4-7 = -1	5-8 = -3	9	9
ΣX_1=70	ΣX_2=80			Σx_1^2 = 30	Σx_2^2 = 44

$$\text{Mean of first group } (M_1) = \frac{\Sigma X_1}{N_1} = \frac{70}{10} = 7$$

$$\text{Mean of second group } (M_2) = \frac{\Sigma X_2}{N_2} = \frac{80}{10} = 8$$

Now, standard error of difference (SE_D) for independent small samples,

$$SE_D = S\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$$

$$= \sqrt{\frac{(N_1-1)S_1^{\,2}+(N_2-1)\,S_2^{\,2}}{N_1+N_2-2}} \times \sqrt{\frac{1}{N_1}+\frac{1}{N_2}} \text{........................(6.5)}$$

$$= \sqrt{\frac{(N_1-1).\frac{\Sigma x_1^{\,2}}{N_1-1}+(N_2-1).\frac{\Sigma x_2^{\,2}}{N_2-1}}{N_1+N_2-2}} \times \sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$$

$$= \sqrt{\frac{\Sigma x_1^{\,2}+\Sigma x_2^{\,2}}{N_1+N_2-2}} \times \sqrt{\frac{1}{N_1}+\frac{1}{N_2}} \qquad \left(\because S_1 = \sqrt{\frac{\Sigma x_1^{\,2}}{N_1-1}}\right)$$

$$= \sqrt{\frac{30+44}{10+10-2}} \times \sqrt{\frac{1}{10}+\frac{1}{10}}$$

$$= \sqrt{\frac{74}{18}} \times \sqrt{\frac{2}{10}}$$

$$= 0.908$$

$$\therefore t = \frac{M_1-M_2}{SE_D} = \frac{7-8}{0.908} = -1.1$$

Since the two groups are unrelated, required degree of freedom (*d.f.*) = *d.f.* for first group + *d.f.* for second group.

i.e. *d.f.* = (N_1-1) + (N_2-1)= N_1 + N_2 - 2= 10 + 10 - 2= 18

Step III: Decision Making,

The critical value of *t* with 18 *d.f.* at 5% and 1% level of significance are 2.10 and 2.88 respectively. Since $|t|_{cal} < |t|_{tab}$ in both cases, we accept null hypothesis. Therefore, the scores obtained by two groups are not significantly different. The obtained difference between means is merely due to some chance factors or sampling fluctuations.

Case III: Significance of the difference between two dependent means (For all N).

As already said, two means are dependent when the scores are obtained from the administration of same test upon same group on different occasions or from the administration of same test upon two equivalent groups on same time. The procedure for testing the significance of the difference between two dependent means is same as previously used except the use of particular formula to estimate error. The formula are different for different methods.

- *To determine the significance of the difference between the means obtained from the initial and final testing,* we use the formula,

$$SE_D = \sqrt{SE_1^{\ 2} + SE_2^{\ 2} - 2r_{12}\, SE_1 SE_2} \text{.......................................(6.6)}$$

$$= \sqrt{\frac{S_1^{\ 2}}{N_1} + \frac{S_2^{\ 2}}{N_2} - 2r_{12}\frac{S_1 S_2}{\sqrt{N_1}\sqrt{N_2}}}$$

In which, SE_1 = Standard error of means of first test.

SE_2 = Standard error of means of second test.

r_{12} = Correlation between initial and final testing.

- *To test the significance of the difference between means of two matched group*. Two groups are said to be matched if a group as a whole is matched with the other group in terms of mean and standard deviation of some other variables instead of one-to-one matching of individuals in two groups. In such case the above formula reduce to

$$SE_D = \sqrt{[(SE_1)^2 + (SE_2)^2\ (1 - r^2)} = \sqrt{\left[\left(\frac{S_1}{\sqrt{N_1}}\right)^2 + \left(\frac{S_2}{\sqrt{N_2}}\right)^2\right](1 - r^2)} \text{.......(6.7)}$$

Example 6.6: Suppose, twelve students of eight grade of a certain school were administrated an achievement test in mathematics and then all the students were provided remedial instruction in mathematics. Three weeks later, the test was administered for the second time. Test the hypothesis that "remedial instruction in mathematics increases test scores of students" using following data.

Test	Scores											
Pre-test (X_1)	33	57	50	65	36	55	37	30	21	46	37	45
Post-test (X_2)	45	67	58	79	46	63	47	25	30	56	35	57

Solution: Since we have to test the hypothesis 'remedial instruction in mathematics increases test scores of students' we formulate pair of hypothesis as below.

Step I: Hypothesis formulation

H_0: There is no effect of remedial instruction in mathematics.

H_1: The remedial instruction in mathematics increase test scores of students.

Step II: Computation of *t* using given data:

Since *N*=12, the formula for *t*-distribution for dependent case is applicable. For this the given data is organized as given below in the table.

X_1	X_2	$X_1 - A = d_1$	$X_2 - A_2 = d_2$	d_1^2	d_2^2	d_1d_2
33	45	-9	-5	81	25	45
57	67	15	17	225	289	255
50	58	8	8	64	64	64
65	79	23	29	529	841	667
36	46	-6	-4	36	16	24
55	63	13	13	169	169	169
37	47	-5	-3	25	9	15
30	25	-12	-25	144	625	30
21	30	-21	-20	441	400	420
46	56	4	6	16	36	24
37	35	-5	-15	25	225	75
45	57	3	7	9	49	21
ΣX_1 =512	ΣX_2 =608	Σd_1=8	Σd_2=8	Σd_1^2 =1764	Σd_2^2 =2748	Σd_1d_2 = 2079

Mean of first group (M_1) = $\frac{\Sigma X_1}{N_1} = \frac{512}{12} = 42.66$

Mean of second group (M_1) = $\frac{\Sigma X_2}{N_2} = \frac{608}{12} = 50.66$

Suppose, assume means A_1 for $X_1 = 42$ and A_2 for $X_2 = 50$ then,

$$S_1 = \sqrt{\frac{\Sigma d_1^{\,2}}{N_1} - \left(\frac{\Sigma d_1}{N_1}\right)^2} = \sqrt{\frac{1764}{12} - \left(\frac{8}{12}\right)^2} = \sqrt{147 - 0.44} = 12.10$$

$$S_2 = \sqrt{\frac{\Sigma d_2^{\,2}}{N_2} - \left(\frac{\Sigma d_2}{N_2}\right)^2} = \sqrt{\frac{2748}{12} - \left(\frac{8}{12}\right)^2} = \sqrt{229 - 0.44} = 15.11$$

$$r_{12} = \frac{\Sigma(d_1 d_2)}{N.S_1.S_2} = \frac{2079}{12 \times 12.10 \times 15.11} = 0.94$$

Substituting these values in formula, we have

$$t = \frac{M_1 - M_2}{\sqrt{\frac{S_1^{\,2}}{N_1} + \frac{S_2^{\,2}}{N_2} - 2r\frac{S_1 S_2}{\sqrt{N_1}\sqrt{N_2}}}}$$

$$= \frac{42.66 - 50.66}{\sqrt{\frac{146.5}{12} + \frac{228.5}{12} - 2 \times 0.94 \frac{12.10}{\sqrt{12}} \times \frac{15.11}{\sqrt{12}}}}$$

$$= \frac{-8}{\sqrt{31.25 - 28.64}} = -\frac{8}{1.615} = -4.95$$

Since the case is dependent and number of students are 12 in each group, degrees of freedom (*d. f.*)= *N* -1 = 12 - 1 = 11

Step III: Decision Making

Since the alternative hypothesis states that the remedial instruction in mathematics increase achievement scores, i.e. indicates clear direction so we can use one-tailed test. The tabulated value for 11 *d.f.* at 0.05 and 0.01 level of significance in one-tailed test are 1.80 and 2.72 respectively. Since $|t|_{cal} > |t|_{tab}$ in both cases then we reject null hypothesis. The alternative hypothesis is accepted, i.e. the remedial instruction in mathematics increases achievement score is true.

Note: An alternative method, namely ***difference method*** can also be used to test the significance of the difference between the means of test scores obtained by administration of the same test to the sample upon two occasions. This method is easier and convenient when sample is small. The formula in this situation can be written as: $t = \frac{M_D}{SE_{MD}}$ ....(6.8)

In which, M_D = Mean of difference of two scores.

SE_{MD} = Standard error of mean of difference $= \frac{\mathrm{SD}}{\sqrt{\mathrm{N}}}$

Example 6.7: Two groups of students are matched for mean and standard deviation on a group of intelligence test. While using learning test battery upon these groups, following data were found.

Group	Mean	*SD*	Number	
A	48.52	10.6	58	$r = 0.50$
B	53.61	15.35	72	

Test the significance of the difference between groups A and B at 0.05 and 0.01 levels of significance.

Solution: From given condition $N > 30$ and groups are matched so case is dependent. Therefore, we proceed accordingly.

Step I: Hypothesis formulation

H_0: The difference of means of two matched group is not significant.

H_1: The difference of means is significant.

Step II: Computation

Since the case is dependent and N is large, the formula is applicable,

$$z = \frac{M_1 - M_2}{\sqrt{[(SE_1)^2 + (SE_2)^2](1-r^2)}}$$

$$= \frac{48.52 - 53.61}{\sqrt{\left[\left(\frac{10.6}{\sqrt{58}}\right)^2 + \left(\frac{15.35}{\sqrt{72}}\right)^2\right]\left[1-(0.5)^2\right]}}$$

$$= \frac{-5.09}{1.97} = -2.58$$

Since N is large, we need of compute *d.f.*

Step III: Decision Making

The critical value of *z* at 0.05 and 0.01 levels of significance are 1.96 and 2.58 respectively. Since the computed value |z| = 2.58 is greater than 1.96 and equal to 2.58. Therefore, the difference is significant at 0.05 and quite significant at 0.01 level of significance.

Exercise 6.1

1. A research methodology teacher is interested to analyze the score obtained by mathematics students and he found the following data from department record.

Group	Mean	*SD*	Number
Mathematics	38.5	7.3	108
Science	34.8	6.4	45

Decide what his conclusion was regarding the performance of students at 0.05 and 0.01 levels of significance.

2.Two groups each made up of 20 from fifth grade students were matched on the basis of I.Q. Filmstrips were used to teach the experimental group, the control group was exposed to a conventional "read and discuss" method. The following data were found.

Group	Mean	Variance	
Control	53.20	54.76	*r* = 0.6
Experimental	49.80	42.85	

Decide statistical significance of difference at 0.05 and .01 level of significance (2.093 for 19 *df.*)

6.2.4 Analysis of Variance (ANOVA)

In previous section, we discussed *t* and *z* test in which we compute the significance of the difference between means of two random samples. However, in real life situations. We encounter in such cases in which we have to find the significance of the difference between more than two sample means. In such case we have two ways.

First, we may infer the significance of the difference between two samples means at a time. That is, compute *t*-test for first two sample means, then for first and third sample means, and so on. For example, if we have to find the significance of the difference between the achievement of four caste groups children, assume A, B, C, and D, then we need to apply *t*-test between the means of A and B, A and C, A and D, B and C, B and D, C and D. This can be more clearer from following induction method.

when there are two samples we apply 1 test.
when there are three samples we apply 3 tests.
when there are four samples we apply 6 tests.
…………………………………....................
when there are *n* samples we apply $\frac{n\,(n\text{-}1)}{2}$ tests.

However, the above process is tedious and too time consuming. Therefore, we use second way or alternative procedure which enables us to determine whether the sample means differ from one another (between group variance) to a greater extent than the test scores differ from their own sample means (within group variance) using the ratio.

$$F = \frac{\text{Variance between groups}}{\text{Variance within groups}} \quad \text{……………………..(6.9)}$$

In which, between group variance is the variance of the group means around the total mean of all groups and within group variance is average variance of the member of each group around their respective group means. The ratio between these two variance is known as *critical ratio* and this ratio can be used to decide the significance of the difference between several means. Therefore, *the composite procedure for testing simultaneously the* *difference between several sample means is known as the analysis of variance (ANOVA)*. This procedure was developed by a great statistician *Sir Ronald A. Fisher* and therefore the variance ratio is generally known as *F-ratio* or *F-test*.

Basic assumptions of ANOVA: The analysis of variance works properly under following assumptions.

- *The population distribution should be normal*. The population for each sample must be normally distributed with the identical mean and variance except for large sample size. But this condition of normality is not strict. *Eden and Yates* showed that even with a population departing considerably from normality, the effectiveness of normal distribution still held. Similarly, *Norton* states that *F* is less sensitive to variations in the shape of population distribution.

- *All the groups of a certain criterion should be randomly chosen from the sub-population having the same criterion*. For example, if we wish to select three groups in school population, first group from grade I, second from grade II, and third from grade III, then we must choose randomly these groups from respective subpopulation (say from same caste group). Failure to fulfill this condition gives incomparable variance and therefore gives biased results.

- *The subgroups under investigation/study should have nearly equal variability*. Otherwise, the interpretation gives false result. The equality of variability is known through homogeneity test (*Bartlett's* or *Hartley's test*). In other words, the within group variance must be approximately equal.

- *The variance ratio must be greater than unity*. That is, the variance between sample is greater than the variance within samples. If the variance within samples is greater than the variance between samples, the larger variance should be kept in numerator. So that *F* is always greater than 1. The value of *F* cannot be negative since both the terms of *F*-ratio are squared values. The value of *F* lies between zero to infinity, i.e., $0 \leq F \leq \infty$ but at $F = 0$ the variance analysis is no meaningful.

***Types of ANOVA*:** The analysis of variance is mainly carried out under one-way classification (One way ANOVA) and two-way classification (Two-way ANOVA). In one-way, the effect of one factor is taken into consideration while in two-way we study the effect of two factors at a single attempt. For example, if we want to study the effect of three different incentives in learning, we take three groups of students randomly from a class. These groups are taught by the same teacher with three different incentives. Achievement tests are taken and mean scores of these three groups are computed. Now, to find the significance of the difference between the means of these three groups we use one-way analysis of variance. Instead, if we have to study the effect of one more variable say school system (public and private) or gender (male and female), we need 2 × 3 = 6 groups, i.e. three groups in each type of schools. In such case, we have to study the impact of two experimental variables (incentives and school types), each having three and two levels respectively. *Therefore, this procedure of studying the effect of two or more variables at the same time is two way analysis of variance.*

Example 6.8 (One way): Suppose that an investigator conducted an experimental study to determine the effect of three different incentives in learning of particular skill. Test the significance of the difference between the means of the following performance scores obtained by different groups.

Groups	Performance scores						
I	3	5	3	1	7	3	6
II	4	5	3	4	9	5	5
III	5	5	5	1	7	3	7

Solution:

Step I: Hypothesis formulation

H_0 : The means of scores of three groups are not significantly different.

H_1: Three means are significantly different

Step II: Computation of F-ratio

Now, for the sake of convenience, assume given groups are X_1, X_2 and X_3 and organize given data for different values in the following table.

X_1	X_2	X_3	Total (X)	X_1^2	X_2^2	X_3^2	Total (X^2)
3	4	5	12	9	16	25	50
5	5	5	15	25	25	25	75
3	3	5	11	9	9	25	43
1	4	1	6	1	16	1	18
7	9	7	23	49	81	49	679
3	5	3	11	9	25	9	43
6	5	7	18	36	25	49	110
ΣX_1 $= 28$	ΣX_2 $= 35$	ΣX_3 $= 33$	ΣX $= 96$	ΣX_1^2 $= 138$	ΣX_2^2 $= 197$	ΣX_3^2 $= 183$	ΣX^2 $= 518$

Since, $N_1 = N_2 = N_3 = 7$, then $N = N_1 + N_2 + N_3 = 21$

Group means, $\frac{\Sigma X_1}{N_1} = \frac{28}{7} = 4, \frac{\Sigma X_2}{N_2} = \frac{35}{7} = 5, \frac{\Sigma X_3}{N_3} = \frac{33}{7} = 4.71$

Correction factor (C) = $\frac{(\Sigma X)^2}{N} = \frac{(96)^2}{21} = 438.85$

Total Variance (S_t^2) = Total sum o squares

$$= \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 518 - 438.85 = 79.15$$

Between group variance (S_b^2) $= \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N_2} + \frac{(\Sigma X_3)^2}{N_3} - C$

$$= \frac{(28)^2}{7} + \frac{(38)^2}{7} + \frac{(33)^2}{7} + \frac{(33)^2}{7} - 438.85$$

$$= \frac{784 + 1225 + 1089}{7} - 438.85 = 3.72$$

Within group variance (S_w^2) $= S_t^2 - S_b^2 = 79.15 - 3.72 = 75.43$

Now, degree of freedom (*d.f.)*

df for $S_t^2 = N - 1 = 21 - 2 = 20$

df for $S_b^2 = K - 1 = 3 - 1 = 2$

df for $S_w^2 = (N - 1) - (K-1) = (N - K) = 21-3 = 18$

Mean square variance

$$\text{Between group} = \frac{S_b^{\ 2}}{df} = \frac{S_b^{\ 2}}{K-1} = \frac{3.72}{2} = 1.86$$

$$\text{Within group} = \frac{S_w^{\ 2}}{df} = \frac{S_w^{\ 2}}{N-K} = \frac{75.43}{18} = 4.19$$

$$F \quad = \frac{\text{Mean square variance between groups}}{\text{Mean square variance within groups}} = \frac{1.86}{4.19} = 0.444$$

Step-III: Decision Making

From *F*-table (see *df* on row for larger mean square variance and *df* on column for smaller mean square variance for specific level of significance) the critical value of *F* = 19.43 at 0.05 level of significance and *F* = 99.44 at 0.01 level of significance. Since both the critical values of *F* from table are strictly greater than the calculated value of *F* = 0.44, we accept null hypothesis that the differences between the means are not significant.

Exercise 6.2

1. Suppose we want to study the effects of eight different experiments conditions (A, B, C, D, E, F, G, H) upon performance on a sensory motor task. From total of 48 subjects, 6 are assigned at random to each of 8 groups and the same test is administrated to all. Do the mean scores achieved under 8 experimental conditions differ significantly ? The obtained scores are given below.

Subjects	Scores on different experimental conditions							
	A	B	C	D	E	F	G	H
1	64	73	77	78	63	75	78	55
2	72	61	83	91	65	93	46	66
3	68	90	97	97	44	78	41	49
4	77	80	69	82	77	71	50	64

5	56	94	79	85	65	63	69	70
6	95	67	87	77	76	76	82	68

Example 6.9 (Two way): The data below represent the marks of eight students by three teachers in terms of their performances on a particular skill.

Teachers	Marks obtained by different students							
	1	2	3	4	5	6	7	8
A	8	6	8	10	12	14	16	20
B	4	3	4	8	6	10	8	11
C	2	6	8	8	8	9	10	8

Analyze the variance of scores among teachers themselves and among the students in their markings.

Solution:

Step I: Hypothesis formulation

H_0: The scoring of students by different teachers and the skill among students are not significantly different.

H_1: The results in both the variables are significant.

Step II: Computation of F-ratios.

a. Now, for the sake of convenience, arrange the given data for required values of F-ratios in two-way analysis of variance.

Students	Rating of teachers			Total	X_1^2	X_2^2	X_3^2	X^2
	A(X_1)	B(X_2)	C(X_3)	(X)				$=(X_1^2+X_2^2+X_3^2)$
1	8	4	2	14	64	16	4	84
2	6	3	6	15	36	9	36	81
3	8	4	8	20	64	16	64	144
4	10	8	8	26	100	64	64	228
5	12	6	8	26	144	36	64	244
6	14	10	9	33	196	100	81	377
7	16	8	10	34	256	64	100	420

8	20	11	8	39	400	121	64	585
	ΣX_1 =94	ΣX_2 =54	ΣX_3 =33	ΣX =207	ΣX_1^2 =1260	ΣX_2^2 =426	ΣX_3^2 =417	$\Sigma X^2 = 2163$

Here, $N = N_1 + N_2 + N_3 = 8 + 8 + 8 = 24$

Correction factor (C) = $\dfrac{(\Sigma X)^2}{N} = \dfrac{(207)^2}{24} = 1785.37$

b. Total Variance (S_t^2) = Total sum of squares

$$= \Sigma X^2 - \frac{(\Sigma X)^2}{N} = 2063 - 1785.37 = 377.63$$

c. Variance for rows (S_r^2) = Sum of squares for students

$$= \frac{(207)2}{3} - 1785.37 = 194.29$$

d. Variance of columns (S_c^2) = Sum of square for teachers

$$= \frac{(94)^2 + (54)^2 + (59)^2}{8} - 1785.37 = 118.75$$

e. Residual sum of squares = $S_t^2 = (S_r^2 + S_c^2)$

$$= 377.63 - (194.29 + 118.75) = 64.59$$

f. Variance in table form and F-ratios

Variation	*df*	Variance	Mean Variance
Rows (students)	r - 1 = 7	194.29	194.29/7 = 27.75
Columns (Teachers)	c - 1 = 2	118.75	118.75/2 = 59.37
Interaction	(r-1) (c-1) = 14	64.59	65.59/14 = 4.61

Now, F-ratio for rows $= \dfrac{\text{Mean squre variance for students}}{\text{Mean square residual variance}}$

$$= \frac{27.75}{4.61} = 6.01$$

F-ratio for columns = $\dfrac{\text{Mean squre variance for teacher}}{\text{Mean square residual variance}}$

$$= \frac{59.37}{4.61} = 12.87$$

g. F-ratios and critical values

			Critical value at	

F-ratio for	*df* for greater value	*df* for smaller value	0.05	0.01	Calculated *F*
Rows	7	14	2.77	4.30	6.0
Columns	2	14	3.74	6.51	12.87

Step III: Decision Making

Since the calculated *F*-ratios for both cases are greater than the respective critical values at .05 and 0.01 level, the result are significant. This means that the teachers ratings among students and among themselves are different.

Exercise 6.3

1. In an experimental study, the three subjects records the following scores on different trails

Subjects	Trails						
	1	2	3	4	5	6	7
A	5	9	3	7	9	3	7
B	6	8	4	5	2	4	3
C	5	7	3	5	9	3	7

Apply analysis of variance to test the difference between means of trails and experimenter.

6.3 Non-Parametric Tests

In previous section we discussed about parametric tests (*t* and *z*) under inferential statistics. In parametric tests the test of significance are based on the assumption that the samples are drawn from normal population. Therefore, the result of such tests can be used to make an assumption about population parameters. But in many situations it is not possible to make dependable assumption about parent population from which samples are drawn. This is due to small sample, lack of normality of parent population, and the data in discrete form (nominal and ordinal data). In such cases we use non-parametric tests namely *Chi* or *Ki-square* (χ^2), *Rho* (ρ) *test, sign test, median test, U-test, Run test, Kolmogrove-Simrnov (KS) test* to test the significance of statistics successfully.

6.3.1 Chi-Square Test

A statistics used to test the significance of the given discrete data is known as *Chi* or *Ki-square* distribution. This test was first introduced by *Helmert* in 1875 and rediscovered independently by *Karl Pearson* in 1900 as a test of goodness of fit.

Mathematically, χ^2-statistics is a measure of discrepancy between observed and expected frequencies. If we denote observed (actual) frequency by f_o and expected (theoretical) frequency by f_e then the formula for Chi-square is

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] \text{...} \quad (6.10)$$

It is a test of independence and used to estimate the likelihood that some factor other than chance accounts for the observed relationship. Therefore, it determines how well the experimentally obtained results fit with the results expected theoretically on some hypothesis.

Note: When observed and expected frequencies agree exactly, $\chi^2 = 0$. The greater the difference, the larger its value. The larger the χ^2 the greater the probability of real divergence of experimentally observed from expected results.

Degree of freedom (df): If the data are presented in a series or variant values in a row or column then the $df = n\text{-}1$ where n represents number of items in the given series. If the number of frequencies are presented in cells in a contingency table then $d.f. = (r - 1)\ (c - 1)$ where r and c are the number of rows and columns respectively.

Assumptions for χ^2-test: The Chi-square test under non parametric statistics can be used under following assumptions.

1. The parametric tests like z and t are based upon the assumption of normal distribution but χ^2 is free from such assumption. That is, we can use χ^2-test in any type of distribution (normal or non normal). Therefore, χ^2-test cannot be used for estimating the value of population parameter.

2. Usually the test is used with discrete data. If the continuous data is reduced to categories, then we can apply *Chi-squar*e test.
3. Chi-square test is used as a test of significance when we have data that are expressed in frequencies or in terms of percentages or proportions that can be reduced to frequencies.
4. The sum of the expected frequencies must always be equal to the sum of the observed frequencies in a χ^2 test.
5. The expected frequency of any item or cell must not be less than 5. If it is less than 5 the frequency of adjacent item or cell should be pooled together in order to make it 5 or more (for more than 1 *d.f.*). But in a 2 × 2 contingency table, the pooling method is fruitless since 1*df* is lost in pooling. In this case we apply correction given by *F Yates* (1934) and known as "*Yates correction for continuity*".
6. The frequency used in χ^2-test must be in absolute terms.
7. Each of the observation of the sample must be independent of each other.

6.3.2 Application of χ^2-test: Chi-square test is used for two broad purposes. Firstly, it is used *as a test of goodness of fit* and secondly, as a *test of independence*.

Case I: Chi-square test as a test of goodness of fit: As a test of goodness of fit, it tries to determine how well the observed results on some study/experiment fit with the results expected theoretically on some hypothesis (hypothesis of chance, hypothesis of equal probability and hypothesis of normal distribution).

The steps used in testing the goodness of fit are as follows:

1. Formulate null and alternative hypothesis using given condition.
2. Compute expected frequencies corresponding to observed frequencies under null hypothesis. If any item of expected frequency is less than 5 then that item must be combined with adjacent item until the combined item have an expected frequency of 5 or more. Compute χ^2 using formula.
3. Write down the critical value of χ^2 at certain level of significance (0.01 or 0.05) using table and compare that value with computed value of χ^2. If computed value of χ^2 is equal to or greater than the critical value (tabulated value) then the difference is taken to be significant and H_0 is rejected. Otherwise, H_0 is accepted.

Example 6.10: A multiple choice test of 50 items having 5 alternatives is administrated over a group of student. A student gets a score of 20 for his 20 correct answers. Find out whether or not his performance is mere guessing.

Solution:

i). Hypothesis formulation

H_0: There is no difference between the score obtained and the score from guessing.

$H_{0:}$: The obtained score is actually different from guessing.

ii). Computation

Since there are five alternatives to each item, the probability of right answer by guessing is 1/5 × 50 = 10 and probability of wrong answer is 4/5 × 50 = 40. To compute χ^2 the data are organized as given below.

Answer	f_o	f_e	f_o-f_e	$(f_o$-$f_e)^2$	$(f_o$-$f_e)^2/f_e$
Right	20	10	10	100	100/10 = 10
Wrong	30	40	-10	100	100/40 = 2.5
Total	50	50			χ^2 = 12.5

iii). Decision Making

The critical value of χ^2 = 3.841 at 0.05 level of significance and χ^2 = 6.635 at 0.01 level of significance. Since computed value of χ^2 = 12.5 is strictly greater than the tabulated value at 0.05 and 0.01 level of significance the null hypothesis is rejected. That is, the score 20 obtained by student was not by more guessing.

Example 6.11: Sixty college principles were asked to express their opinion in terms of yes, no or indifference for the implementation of new education policy from current year. The number of responses were as 13, 27 and 20 respectively. Do these result indicate a significant trend of opinion ?

Solution:

1. Hypothesis formulation

H_0 : The difference in response is by mere chance.

H_1 : There is actual difference in response.

2. Calculation: The data are organized for Chi-square value as below

Answer	f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$(f_o-f_e)^2/f_e$
Yes	13	20	-7	49	49/20 = 2.45
No	27	20	7	49	49/20 = 2.45
Indifference	20	20	0	0	$\chi^2 = 4.9$

Degree of Freedom (*d.f.)* = (*r* - 1) (*c* - 1) = (3-1) (2-1) = 2

3. Decision

Since critical value of $\chi^2 = 5.991$ at 0.05 level of significance is greater than calculated value at $\chi^2 = 4.9$ then H_0 is accepted. This means that the opinion are not significantly different.

Exercise 6.4

1. Forty lectures were rated into three groups: very effective, satisfactory, and poor, in terms of their professional competency by their principal as shown below.

Rating	Very effective	Satisfactory	Poor
Frequency	16	20	6

Does the distribution of ratings differ significantly from that to be expected if professional competency is normally distributed in our population of lecturers ?

2. A publisher is interested in knowing purchasing habit of student regarding particular publication of M.Ed. level and found the following data on a survey.

Publication	A	B	C	D
First choice	35	30	45	55

Test result hypothesis at $\alpha = 0.05$ that there are no difference among frequencies of first choice of tested publication ($\chi^2 = 7.815$ for 3*df*).

Case II. Chi-square test as a test of independence: As a test of independence, χ^2 is usually applied for testing the relationship between two variables in two ways. First, by testing the null hypothesis of independence saying that the two given variables are independent of each other and second, by computing the value of contingency coefficient, a measurement of relationship existing between the two variables.

In χ^2 test, as a test of goodness of fit, we use one way classification table of observed frequencies in a single row or column. When the observed frequencies occupy *r* rows and *c* columns, a two way classification table is formed and such table is called *contingency table*. Chi-square test is equally useful to test the relationship or association between variables. The **steps** used in testing the independence of attributes are as follows:

1. Formulate null and alternative hypothesis.
2. Computation
   i. Compute expected cell frequencies using the relation

$$\text{Element} = \frac{\text{Row total} \times \text{Column total}}{\text{Total number of observations}} = \frac{RT \times CT}{N}$$

   ii. Compute value for χ^2 using formula.

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

   iii. Determine degree of freedom (*df*) using

$$d.f. = (r\text{-}1)\ (c\text{-}1)$$

   iv. Find tabulated (critical) value of χ^2 at certain level of significance (0.05 or 0.01) using table at appendix.

3. Decision making

Make decision using calculated and tabulated value of χ^2. If calculated value is greater or equal to tabulated value, then reject H_0. Otherwise accept H_0 and give conclusion.

I: ***Testing null hypothesis of independence (2 × 2 contingency table)***

Example 6.12: There is general belief that high income families send their children to private schools and low income families often send their children to public schools. To verify this 2000 families in Kathmandu valley were selected at random and the following results were found.

Income	Schools		Total
	Private	Public	
Low	300	700	1000
High	500	500	1000
Total	800	1200	2000

Test which income and type of schooling are independent (χ^2 = 3.84 for 1 *df* at 5% level of significance).

Solution:

Step I: Hypothesis formulation

H_0 : The income and type of schooling are independent.

H_1 : The income and type of schooling are not independent.

Step II: Computation

i. Computation of expected frequencies (f_e) under H_o and χ^2.

f_o	$\frac{RT\times CT}{N} = f_e$	f_o-f_e	$(f_o\text{-}f_e)^2$	$(f_o\text{-}f_e)^2$/fe
300	1000×800/2000 = 400	-100	10000	25
700	1000×1200/2000 = 600	-100	10000	16.67
500	1000×800/2000 = 600	-100	10000	25
500	1000×1200/2000 = 600	-100	10000	16.67
2000	2000			χ^2 = 83.34

ii. Degree of freedom (*df*) = (*r* - 1) (*c* - 1) = (2 - 1) (2 - 1) = 1

iii. Given that tabulated value of χ^2 = 3.84 for 1*df* at 0.05 level of significance.

Step III: Decision making

Since the computed value of χ^2 is strictly greater than tabulated value at 5% level of significance for 1 *df*, the null hypothesis is rejected. That is, the income and type of schooling are not independent. This concluded that high income families tend to send their children to private school whereas low income families send their children to public schools.

Alternative Method

The above example can be solved to find the value of χ^2 using direct formula in the following way.

Put the given data with notations in 2 × 2 contingency table.

Income	Schools		Total
	Private	Public	
Low	300 (*A*)	700 (*B*)	1000 (*A* + *B)*
High	500 (*C*)	500 (*D)*	1000 (*C* + *D*)
Total	800 (*A* + *C*)	1200 (*B* + *D*)	2000 (*A*+*B*+*C*+*D*=*N)*

Now, substitute these values directly in formula for χ^2 statistics.

$$\chi^2 = \frac{N\ (AD - BC)^2}{(A+B)\ (C+D)\ (A+C)\ (B+D)} \text{..............................}(6.11)$$

$$= \frac{2000\ (300 \times 500 - 700 \times 500)^2}{1000 \times 1000 \times 800 \times 1200}$$

$$= 83.34$$

Now, follow other steps as in previous way.

Contingency coefficient (C): The contingency coefficient refers to the correlation between two variables / traits / attributes under study. This coefficient can be computed using following formula.

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}} \text{..}(6.12)$$

Therefore, contingency coefficient of previous example

$$= \sqrt{\frac{83.34}{2000 + 83.34}} = 0.2$$

Note:

1. Like other coefficient of correlation (r, ρ), c does not have certain limit of ±1.

3. *C* does not have negative value. Its value is dependent upon number of categories.

4. If the table have *K* rows and *K* columns then the upper limit

$$\text{C} = \sqrt{\frac{(K-1)}{2}} = 0.5$$

Therefore, for 2 × 2 table, upper limit of $C = \sqrt{\frac{(2-1)}{2}} = 0.5$

for 3 × 3 table, upper limit of $C = \sqrt{\frac{(3-1)}{3}} = 0.82$

for 4 × 4 table, upper limit of $C = \sqrt{\frac{(4-1)}{4}} = 0.87$

for 2 × 3 table, upper limit of $C = \sqrt{\frac{(2-1)}{2}} = 0.5$

(When the number of rows and columns differ in table, take smaller number as *K*).

Correction for small frequencies (in 2 × 2 table): If the number of items in any cell in 2×2 table (1 *df*) is less than 5, the computation of χ^2 in usual way gives an overestimate of χ^2 value. Consequently, we may reject null hypothesis which, in fact, should not be rejected. This problem can be avoided using following Yates correction formulae

Usual formula | Corrected formula

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right] \qquad \chi^2 = \Sigma\left[\frac{(|f_o - f_e| - 0.5)^2}{f_e}\right] \text{................(6.13)}$$

$$\chi^2 = \frac{N\,(AD - BC)^2}{(A+B)\,(C+D)\,(A+C)\,(B+D)} \qquad \chi^2 = \frac{N\,(|AD - BC| - N/2)^2}{(A+B)\,(C+D)\,(A+C)\,(B+D)}$$

Note: If *N* is large, the use of Yates correlation will make very little difference in the value of χ^2. If *N* is small, the application of Yates correction may overstate the probability. It is recommended, therefore, by many authors that Yates correction be applied to every 2×2 table, even if no theoretical cell frequency is less than 5 (Sancheti & Kapoor, 1983, p.18-21).

Example 6.13: A standard achievement test in arithmetic was conducted to all the sixth grade children in a public school. Among 40 boys selected randomly, 23 were at or above national norm in the test and 17 were below the national norm. Similarly, a random sample of 50 girls showed 22 at or above the national norm and 28 below. Are the boys really better than the girls in arithmetic ?

Solution:

1. Hypothesis formulation

H_0 : The achievement of boys and girls are not significantly different.

H_1: Boys are really better than the girls in arithmetic

2. Computation

I. Arrange obtained and expected frequency in fourfold table as follows:

Observed Frequency (f_o)				Expected frequency (f_o)			
Gender	Norm		Total	Gender	Norm		Total
	Below	At or above			Below	At or above	
Boys	17	23	40	Boys	20	25	40
Girls	28	22	50	Girls	25	25	50
Total	45	45	90	Total	45	45	90

II. Computation of χ^2

f_o	f_e	$f_o- f_e$	$(f_o-f_e)^2$	$(f_o-f_e)^2/f_e$
17	20	-3	6.25	0.31
23	20	3	6.25	0.31
28	25	3	6.25	0.31
22	25	-3	6.25	0.31
90	90	$\chi^2 = 1.25$		

III. Degree of freedom (df) = (r-1) (c-1) = (2-1) (2-1) = 1

IV. Critical value of χ^2 = 3.84 for 1 *d.f.* at 0.05 level of significance

4. Decision making

Since computed value of χ^2 is less than the tabulated value at 0.05 level of significance, the null hypothesis is accepted. That is, there is no true gender difference in arithmetic.

Exercise 6.5

1. A study related to preference for lecturer and discussion method was conducted over 30 teacher educators. The preference on different methods were as follows:

Teacher educators	Preference on method		Total
	Lecturer	Discussion	
Male	10	5	15
Female	12	3	15
Total	22	8	30

Is preference for certain type of method independent of gender at 1 percent level of significance ?

II: ***Testing of null hypothesis of independence (in any contingency table)***

Example 6.14: Students were asked to select one of the five elective subjects in grade XI at certain college. The choice of their selection were found as follows:

Students	Subjects					Total
	A	B	C	D	E	
Boys	25	30	10	25	10	100
Girls	10	15	5	15	15	60

Do you think that the choice of the subjects is dependent upon the gender of the students ?

Solution:

1. Hypothesis formulation

H_0 : The choice of the subject is independent of sex

H_1: Subject selection is influenced by sex.

2. Computation

i. Organization of data in contingency table with obtained and expected frequencies (expected frequencies are in brackets).

Students	Subjects					Total
	A	B	C	D	E	
Boys	25(21.9)	30(28.1)	10(9.4)	25(25.0)	10(15.6)	100
Girls	10(13.1)	15(16.9)	5(5.6)	15(25.0)	15(9.4)	60
Total	35	45	15	40	25	160

Since expected frequency = $\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$

(Other frequencies can be calculated in the same way)

ii. Computation of the value of χ^2

fo	*fe*	*fo-fe*	$(fo\text{-}fe)^2$	$(fo\text{-}fe)^2/fe$	
25	21.9	32.1	9.61	0.439	$\chi^2 = \sum\left[\frac{(f_o - f_e)^2}{f_e}\right]$
30	28.1	1.9	3.61	0.128	
10	9.4	0.6	0.36	0.038	
25	25.0	0	0	0.000	
10	15.6	-5.6	31.36	2.010	
10	13.1	-3.6	9.61	0.733	
15	16.9	-1.9	3.61	0.214	
5	5.6	0.6	0.36	0.064	
15	15.0	0	0	0.000	
15	9.4	5.6	31.36	3.336	
160	160			$\chi^2 = 6.962$	

iii. Degree of freedom (*df*) = (*r*-1) (*c*-1) = (2-1) (5-1) = 1

iv. Determining critical value

For 4 *df* the tabulated value of χ^2 (using table in Appendix E) at 5% level of significance is 9.488

at 1% level of significance is 13.277

2. Decision making

Since the calculated value of χ^2 = 6.962 is less than the tabulated value at 5% and 1% level of significance, the null hypothesis is accepted. That is, the choice of subject is independent of gender.

Example 6.15: Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students falling in different intelligence levels. The number of students in each category were as follows:

Students	Subjects				
	Below average	Average	Above average	Genius	Total
A	86	60	44	10	200
B	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different ? (χ^2 = 7.8 for 3*df* and 9.49 for 4 *df* at 5%)

Solution:

1. Hypothesis formulation

   H_0 : Two sampling techniques are not significantly different.

   H_1 : Two sampling techniques are different.

2. Computation

   i. Organization of data in contingency table.

   Since, expected frequency = $\frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$

   Expected frequency for 86 = $\frac{200\times126}{300}$ = 84

We can compute expected frequencies in similar manner. This gives following frequency table.

Researcher	Students				Total
	Below average	Average	Above average	Genus	
A	86 (84)	60 (62)	44 (46)	10 (8)	200
B	40 (42)	33 (31)	25 (23)	2 (4)	100
Total	126	93	69	12	300

ii. Computation of χ^2

f_o	f_e	f_o-f_e	$(f_o\text{-}f_e)^2$	$(f_o\text{-}f_e)^2/f_e$	$\chi^2 = \sum \left| \frac{(f_o - f_e)^2}{f_e} \right|$
86	84	2	4	0.047	
60	62	-2	4	0.064	
44	46	-2	4	0.086	
10	8	2	4	0.5	
40	42	-2	4	0.095	
33	31	2	4	0.129	
25	23	2	4	0.173	
2	4	-2	4	1	
$\Sigma(f_o\text{-}f_e)^2/f_e$ = 2.094					

iii. Degrees of freedom = (*r*-1) (*c*-1) = (2-1) (4-1) = 3

iv. Critical value

The tabulated value of χ^2 for 3 *df* at 0.05 = 7.815

The tabulated value of χ^2 for 3*df* at 0.05 = 11.345

2. Decision making

Since the computed value of χ^2 = 2.094 is less than the tabulated value at both 5% and 1% level of significance, the null hypothesis is accepted. That is, the sampling techniques adopted by two researchers are not significantly different.

Exercise 6.6

1. On a survey of 1000 people, the following number of vaccinated people and non-vaccinated people were found to be attacked by certain diseases.

People	Attacked	Not attacked	Total
Vaccinated	20	300	320
Not vaccinated	80	600	680
Total	100	900	1000

Test the effectiveness of inoculation at 0.05 and 0.01 level of significance.

2. At first day of admission opening, the following number of students on different major subjects of M. Ed. level were found to be admitted in Central Department of Education.

Gender	Subject			Total
	Math	English	Nepali	
Male	8	16	6	30
Female	2	8	10	20
Total	10	24	16	50

Decide whether the gender influence the selection of subjects (At 0.05 and 0.01 level of significance)

Exercise

Multiple choice questions

1. The degree of freedom (*df*) for two independent group is

   (a) $N_1 + N_2 - 1$ (b) $N_1 + N_2 - 2$ (c) $N_1 - N_2 + 2$ (d) $N - 1$

2. If we reject null hypothesis when it is actually true, then in such case we may commit an error called

   (a) α - error (b) β - error (c) γ - error (d) None

Short answer questions

1. Clarify the concept of inferential statistics and differentiate between parametric and non-parametric statistics.
2. Discuss the application of χ^2-test.

Long answer questions

1. Discuss basic assumption of ANOVA. Test the significance of the difference between the means of three groups on an experimental study to determine the effect of three different incentives on learning of particular skill.

Groups	Performance scores						
A	3	5	3	1	7	3	6
B	4	5	3	4	9	5	5
C	5	5	5	1	7	3	7

2. Define χ^2-test and state its basic assumptions. A multiple choice of 100 items having 4 alternatives is administrated over a group of students. A student gets a score of 35 for his 35 correct answers. Decide whether or not his performance is the effect of mere guessing at 0.05 and 0.01 level of significance.

References

Amos, J.R., Brown, F.L., & Mink, O.G. (1965). *Statistical concepts: A basic program.* USA: Harper & Row, New York.

Garrett, H.E. (2008). *Statistics in psychology and education.* New Delhi, India: Surjeet Publication.

Gupta, S.C., & Kapoor, V.K. (1983). *Fundamental of mathematical statistics.* (8th ed.). New Delhi, India: Sultan Chand & Sons.

Gupta, S.P. (1988). *An easy approach to statistics.* New Delhi, India: S. Chand & Company (Pvt.) Ltd.

Koul, L. (2009). *Methodology of educational research* (4th ed.). New Delhi, India. Vikash Publishing House, Pvt. Ltd.

Mangal, S.K. (2005). *Statistics in psychology and education.* (2nd ed.). New Delhi, India: Prentice Hall of India.

Manuel, H. T. ( ). *Elementary statistics for teachers.* Ram Nagar, Delhi: Eurasia Publishing House (Pvt.) Ltd.

Minium, E.W., King, B.M., & Bear, G. (2010). *Statistical reasoning in psychology and education* (3rd ed.). New Delhi, India: John Wiley & Sons Inc.

Sancheti, D.C & Kapoor, V.K. (1983). *Statistics : Theory methods & application.* Delhi, India: Sultan Chand & Son