ISBN - 978-81-949922-8-8

Pocket Book of Statistics in the Social Sciences

Dr. Dilip Kumar Ghosh, Dr. Tinni Dutta & Ms.Chilka Mukherjee

ISBN - 978-81-949922-8-8

Pocket book of Statistics in the Social Sciences

Dr. Dilip Kumar Ghosh
Dr. Tinni Dutta
Chilka Mukherjee

Published by
Global Learners Academy of Development
Korattur, Chennai – 600 080. Tamilnadu, India.
Email – editor@gladnetwork.in / www.gladnetwork.in

May 2021

ISBN - 978-81-949922-8-8

ISBN 978-81-949922-8-8

9 788194 992288 >

To cite this work:

Dilip Kumar Ghosh Dr. Tinni Dutta Chilka Mukherjee. (2021). ISBN - 978-81-949922-8-8 - Pocket book of Statistics in the Social Sciences. Chennai, India: Global Learners Academy of Development.

Published & Printed by

Global Learners Academy of Development
Korattur, Chennai – 600 080. Tamilnadu, India.
Email – editor@gladnetwork.in
Web - www.gladnetwork.in

Acknowledgement

Our heartfelt gratitude to Disha Halder, State Aided College Teacher, Bangabasi College, Kolkata, for going through each chapter and sums in the book and providing valuable suggestions.

Sincere thanks to our respected professors for their constant support and guidance and to our students who motivate us.

We are also thankful to our well-wishers for their countless blessings.

CONTENTS

BIOSKETCH OF AUTHORS

Professor D. K Ghosh

Dr. D. K. Ghosh did B.Sc and M.Sc. in Statistics from Bhagalpur University, Bihar and Ph,D. in Statistics specialized in Design of experiments from I. S. I. Delhi center under Professor M. N. Das. Ghosh was Professor and Head at Department of Statistics, Saurashtra University Rajkot since 1986. Presentably he is assigned as UGC BSR Faculty Fellow after his retirement.

Dr. Tinni Dutta

Dr. Tinni Dutta is a famous educationist and is a prolific and proficient writer since 1991. At present she is working as an Assistant Professor in the Department of Psychology, Muralidhar Girls' College, Kolkata, India. She has amazing credentials and has visited UK, USA, France, Germany, Switzerland, South Africa and different parts of Asia – China, Bangkok, Indonesia, and Singapore – in her professional capacity. She provides invited lectures, which are full of insights; value oriented and has motivational strategies. Her specialized areas are psychoanalysis and literature, clinical psychology, educational psychology, drug addiction and HIV/AIDS and Psychospirituality. She has been awarded Asian Fellowship by International Harm Reduction Association in the year 2001. She has been nominated as a member of prestigious IVP-

International Visitor Programmme in the arena of HIV/AIDS in USA in the year 2002. She was selected as a chief facilitator in Jakarta, Indonesia in the workshop entitled as Colour Therapy in the year 2013. She has achieved the honour of Research Fellow in Edinburg, Napier University in the year 2015. She has delivered a special lecture on Psychospirituality in Oxford University in the year 2016. She has presented her paper for several times in Asiatic Society and Ramkrishna Mission, Institute of Culture, Golpark, Kolkata. She has written quite a few books and a special UGC project on Tagore and is also recipient of several awards. In her private life, she enjoys poetry and songs.

Ms. Chilka Mukherjee

Ms. Chilka Mukherjee is working as a State Aided College Teacher in the Department of Psychology, Bangabasi College, Kolkata, India. She has qualified in the national Eligibility Test for Assistant Professor in December, 2019. She is also a RCI registered Clinical Psychologist. She has published two research articles. In her private life she enjoys reading books and listening to music.

Chapter 1

INTRODUCTION: STATISTICS IN THE SOCIAL SCIENCES

Psychology is concerned with the study of behavior which incorporates mainly the higher order organisms; both human and animal behavior. The term behavior includes both overt and covert behavior. The methods for studying behavior include direct observation, experiment, introspection, interview, case study, etc. Psychology has several branches such as Cognitive Psychology, Clinical Psychology, Industrial and Organizational Psychology, Social Psychology, Stress Management and Community Psychology and many more. Thus, it is evident that the scope of Psychology is manifold and it is present in every nook and corner of our life.

Probably all of us are consumers of Statistics because most of the information we get access to using the media use statistics to represent facts in a simplified manner to the public. Newspapers and Television channels often resort to graphs and pie charts to represent data. For instance, which newspaper is leading in the city in terms of sell may be represented using a bar graph. Thus it is evident that Statistics is indeed a part and parcel of life.

Research is an integral part of every discipline. The progress, nurture and expansion of any scientific endeavour rest on research. Both quantitative as well as qualitative research may be carried out to gain deeper insights in the study of behavior. Quantitative research involves collecting numerical data using standardized tools whereas qualitative research involves collecting qualitative data using in depth interviews, open-ended questions, etc. Quantification of variables is necessary while carrying out research where Statistics has an important role to play. Statistics deals with numerical facts. It is very essential to have a basic knowledge of statistical principles to pursue quantitative research as well as to understand research articles because quantitative research deals with numerical facts. This in turn helps to develop analytic and critical thinking. A scientific mind should always be skeptical.

Tate (1955) summarized the meaning of Statistics by commenting – It’s all perfectly clear; you compute statistics (mean, median, mode, etc.) from statistics (numerical facts) by statistics (statistical methods).

Scientific disciplines must rely on empirical methods for collection, organization, tabulation, analysis and interpretation of data. By ‘data’, we mean facts. For instance, the scores obtained by a group of students in an achievement test may be referred to as data. It is to be noted that a set of scores does not convey any meaningful information until and unless it is subjected to further analysis. In other words, researchers resort to statistical methods to make sense of the data collected while conducting research. For example, the mean score may be computed by summing up the scores obtained by all the students and dividing the sum by the number of students. Again by means of dispersion it may be understood how the data is distributed, whether the scores lie near to the mean value or far away from it. That is, how each of the data differ by a constant or a mean value. Going further, percentiles may be computed to get an idea about the position of each student based on the scores obtained by them in the achievement test and conclusion may be drawn. Categories such as Average, Below Average, Above Average and so on may also be made with the help of the data. So it is clear that a set of data can be utilized in various ways and the most suitable measures are to be selected by researchers depending on the nature of research and the data are to be utilized as effectively as possible.

The two types of statistical methods are:

1. **Descriptive Statistics** – It is the branch of Statistics used to summarize, describe and communicate numerical observations in an organized fashion.
   Measures of central tendency and measures of dispersion are included in descriptive statistics. Suppose a teacher is interested to know the average score of 50 students in a class test. For this purpose, she adds the scores obtained by all the students and divides the total by 50. This is known as the average or the arithmetic mean. Furthermore, she can also use the measures of dispersion to know the extent to which the scores are spread in the distribution, which will aid in the evaluation of the performance of the students.

2. **Inferential Statistics** – It is the branch of Statistics used to make inferences and draw conclusions based on the data obtained from a research.
   This statistical technique is exclusively used to test hypotheses. Two types of hypotheses are commonly encountered by social science researchers – the null hypothesis and the alternative hypothesis. Null hypothesis is a no-difference hypothesis, i.e., it states that there is no significant difference between the specified groups. Any observed difference may be attributed to chance or experimental error. Alternative hypothesis states that there is a significant difference between the groups under study.

Population and sample

These two concepts are very important in understanding statistical concepts and drawing inferences based on the same. A population may be defined as the totality of a particular characteristic for any specific group. In other words, population includes all the elements in a group of individuals or objects. For e.g., a population of graduates, a population if obese individuals, etc. Population includes all the past, present and future members of a group, which makes it infinitely large due to which it, cannot be studied empirically. Instead, samples drawn from populations may be used for measurement purposes. A sample may be defined as any selected number of individuals or objects from a population. In other words, a sample consists of one or more observations drawn from the pool of population, representing all the particular characteristics of the specific population. This selection is usually done according to some rule.

By studying the sample, some inferences can be drawn about the population.

A measure based upon a sample is known as a statistic and a measure based on the population and inferred from a statistic is known as a parameter.

Exercise

1. Define Statistics.
2. Why is it necessary for Social Scientists to study Statistics?
3. What is meant by ‘data’?
4. What do you understand by the terms population and sample?
5. What are the two types of statistical methods used in behavioural sciences?

Chapter 2

GRAPHICAL REPRESENTATION

Graphical representation refers to pictorial representation of Statistical data. There are several advantages of representing data graphically. They are eye-catching and are often successful in holding the attention of the viewer. It also provides a longer lasting impression on the brain. Moreover, it is easy to understand and a large set of data may be conveniently represented in a pictorial fashion. This in turn makes the data more concrete and understandable. Comparative analyses of several data sets may also be easily carried out with the help of graphs.

Usefulness of Graphical representation of data:

- Graphs are used to represent relationships between variables.
- It is a good alternative in case of representation of causal relationships.
- Graphs represent mathematical data in a picturesque form.
- Complex and huge data can be compressed into easy visual representation.
- Graphs make the statistical data easily understandable to lay persons.
- The attributable changes like growth, practice, training or learning of data can be shown through graphs.
- Graphs also help in seeing certain characteristics and trends in a set of data.

General principles of graphical representation of data

Plotting data on a graph paper is done with reference to two lines or coordinate axes in which the horizontal line is referred to as the X-axis and the vertical line is referred to as the Y-axis. These lines are perpendicular to each other. The point at which the two axes (plural of 'axis') intersect is known as the O, or Origin or the Zero point which is essentially considered as the starting point for representing any set of data. The distances measured along the X-axis to the right of the Origin (O) and along the Y-axis above the O are positive. The distances measured along the X-axis to the left of the O and along the Y-axis below the O are negative. The X and Y axes form four divisions known as quadrants with reference to their intersection at O. The directions on the four quadrants are as follows:

- The upper right division or the first quadrant – Both X and Y measures are positive (++)
- The upper left division or the second quadrant – X is negative and Y is positive (-+)
- The lower left division or the third quadrant – Both X and Y are negative (--)
- The lower right division or the fourth quadrant – X is positive and Y is negative (+-)

To locate a point whose coordinate axes are X=2 and Y=4, we have to move 2 units to the right of the Origin along the X-axis and 4 units above the Origin along the Y-axis. The point at which the perpendiculars drawn along these points intersect is the desired point or location on the graph.

Bar Graph

In Psychology, a bar graph or bar diagram is constructed to compare the relative amounts of traits owned by two or more groups. Ungrouped data may be conveniently represented using a bar diagram or bar graph. Horizontal or vertical bars are drawn to categorize the data. Comparative analyses of data may be easily carried out with the help of bar graphs. The bars may be drawn either vertically or horizontally.

The bar graph in the figure given below shows a student's performance in Psychology in three consecutive years which helps us understand the progress made by the student.

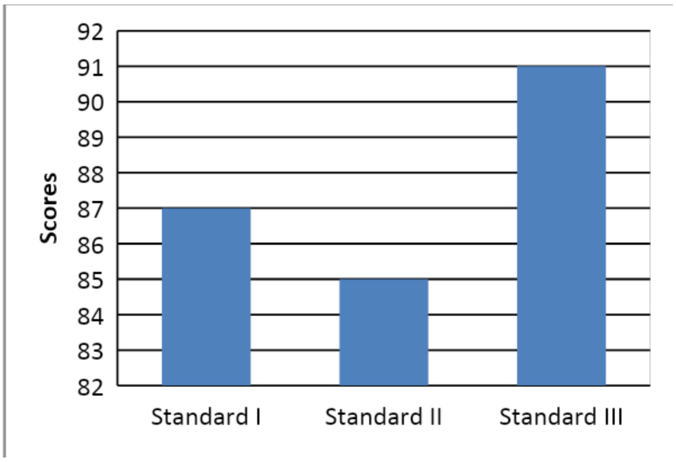


Fig 1: Bar graph showing the categories (Standard I, II and III) on X-axis and scores on Y-axis

1) A man with a monthly salary of 10,000rs plans his budget for a month as given below:

Item	Food	Clothing	Education	Monthly bills	Savings
Amount (in Rs.)	3000	1500	1400	2100	2000

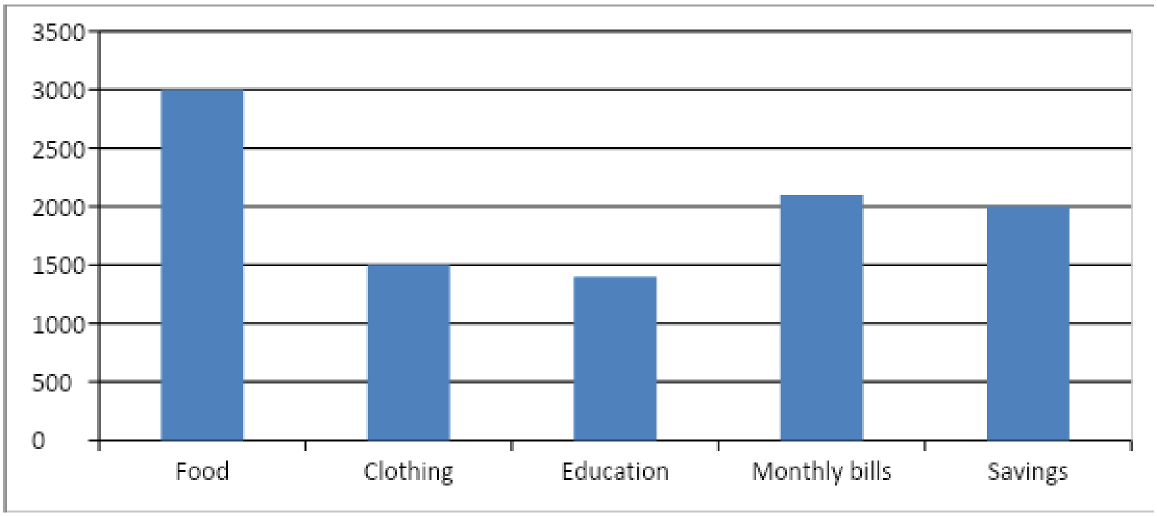


Fig 2: Bar graph showing the categories on X-axis and amount (in Rupees) on Y-axis

2) Given below is the data of scores obtained by boys of the 10th standard of a certain school in verbal reasoning and numerical reasoning.

Class	X A	X B	X C	X D	X E
Verbal reasoning	30	32	28	30	34
Numerical reasoning	41	43	40	42	44

Draw a bar graph to represent the above data.

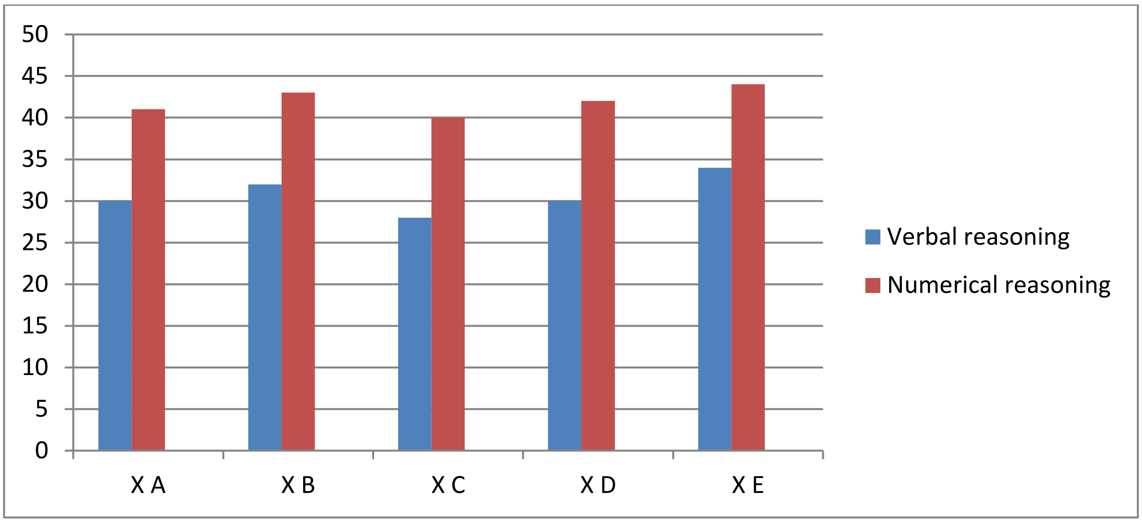


Fig 3: Bar graph showing classes on X-axis and scores on Y-axis

Histogram

Histogram is one of the most popularly used graphs in Psychology. It displays the frequency distribution of a given set of grouped data. The upper and lower limits of class intervals must be continuous in order to plot histogram. If the class intervals are not continuous, for instance 35-39, 40-44, then the upper and lower limits are to be converted to actual class limits, that is 34.5-39.5, 39.5-44.5 in order to make it continuous. The actual lower and upper limits are plotted on the graph paper along the X-axis maintaining continuity and the frequencies are plotted along the Y-axis. It is to be kept in mind that appropriate units must be selected for representation along the axes. Both the X-axis and the Y-axis should not be too long or too short. A general rule suggested by Garrett (1971) is that the X and Y units must be selected in such a way which will make the height of the figure approximately 75% of its width.

Histogram may be properly understood with the help of an example.

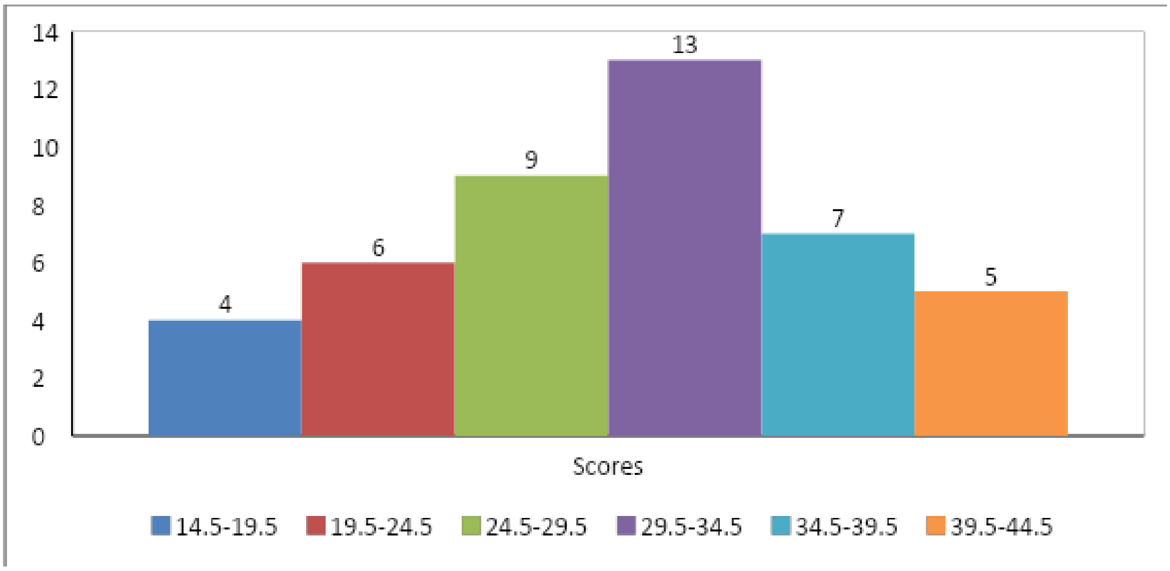


Fig 4: Histogram showing scores on the X-axis and number of students on the Y-axis

The graph given above shows the scores obtained by students in a particular subject and the number of students (frequency) scoring in a particular score range. For instance, 5 students score between 39.5 to 44.5, 7 students score between 34.5 to 39.5 and so on.

Sum 1: Draw a histogram for the following data:

Height (in cm)	140-150	150-160	160-170	170-180	180-190
No. of students	3	5	8	6	3

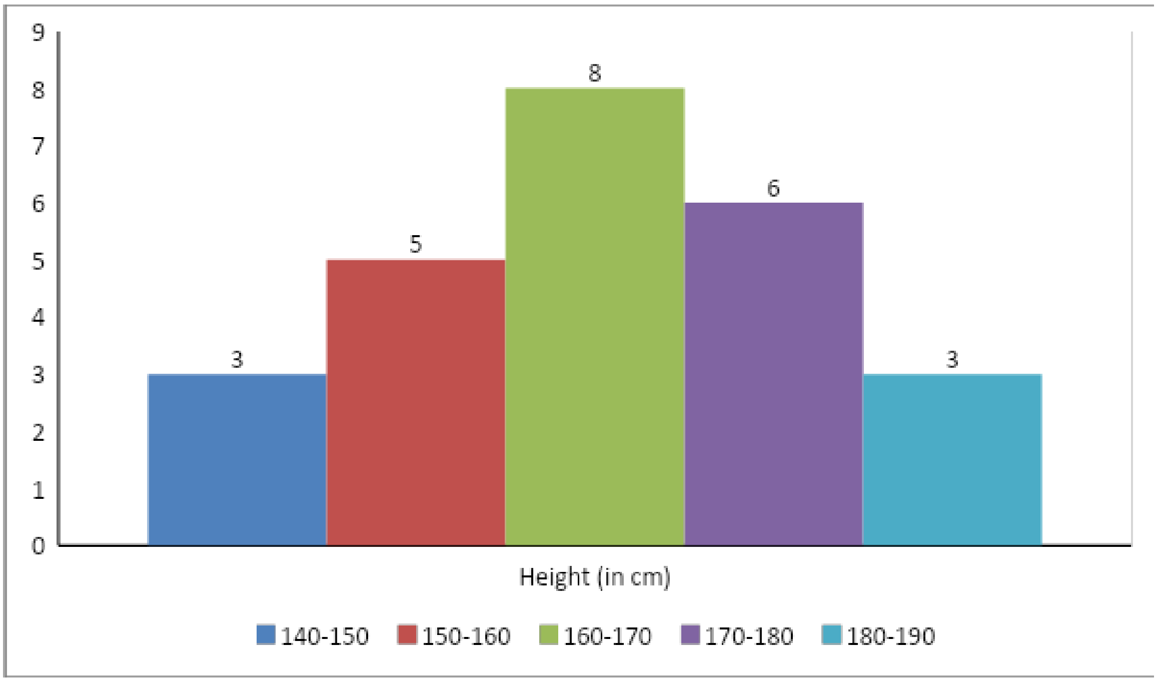


Fig 5: Histogram showing height (in cm) on X-axis and number of students (frequency) on Y-axis

Sum 2: Draw a histogram for the following data.

Marks obtained	51-55	56-60	61-65	66-70	71-75
No. of students	6	9	12	8	7

The given frequency distribution is discontinuous. It has to be converted into a continuous frequency distribution by means of adjustment factor.

Adjustment factor = (Lower limit of one class – Upper limit of previous class)/2

= (56-55)/2

= 0.5

The adjustment factor (0.5) is to be subtracted from all the lower limits and added to all the upper limits to bring about continuity in the frequency distribution.

Classes before adjustment	Classes after adjustment	Frequency
51-55	50.5-55.5	6
56-60	55.5-60.5	9
61-65	60.5-65.5	12
66-70	65.5-70.5	8
71-75	70.5-75.5	7

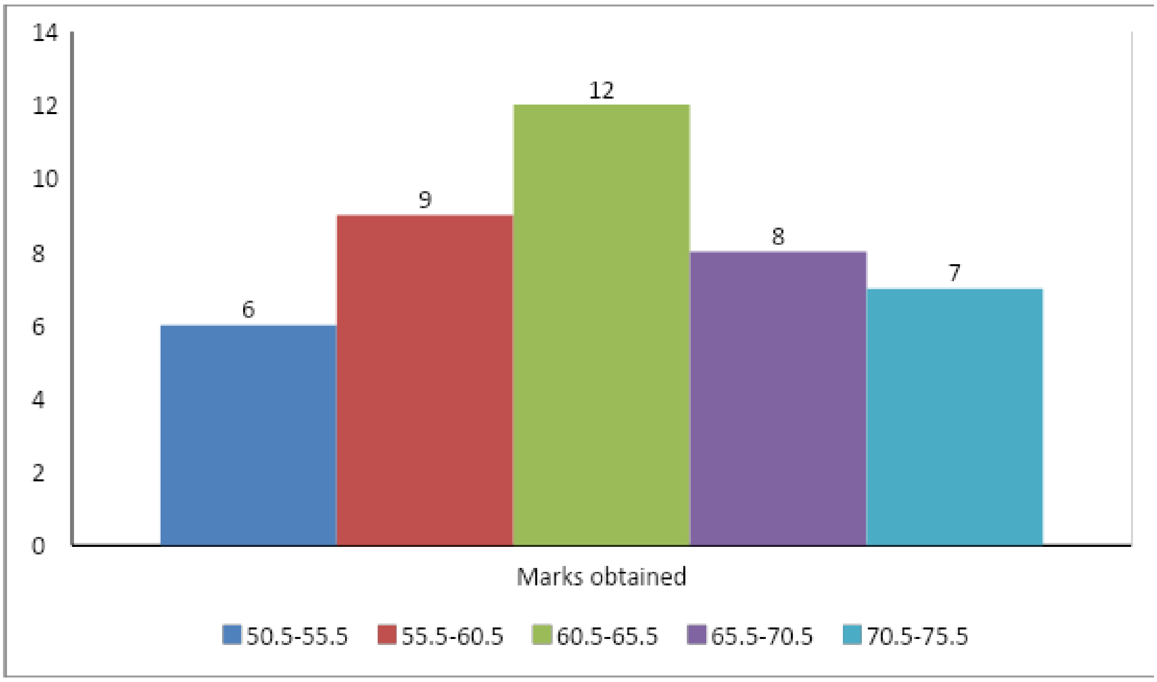


Fig 6: Histogram representing the marks obtained by students on X-axis and number of students (frequency) on Y-axis

Bar graph and Histogram: Comparison

Bar graph is based on categorical variables whereas histogram represents measures obtained on a continuous variable. In the bar graph each bar (horizontal or vertical) represents a specified category and there is a gap between two categories. On the other hand, there is no gap between the vertical rectangles in histogram as they represent the class intervals corresponding to a continuous measurement variable.

Frequency Polygon

A frequency polygon is a line graph constructed by joining the midpoints of each class interval of a given frequency distribution. Frequency polygon is an important way of representing a frequency distribution of quantitative data. An easy way to obtain a frequency polygon is to join the midpoints of the upper bases of the rectangles of the histogram by straight lines. However, it is not necessary to draw a histogram first before plotting a frequency polygon as it can be directly constructed from a frequency distribution.

Steps followed in drawing the frequency polygon:

1. Two extra class intervals, one below the lowest and the other above the highest, are to be taken.
2. The mid points of all the class intervals (including the two extra class intervals) are to be calculated.
3. The mid points are plotted along the X-axis and the corresponding frequencies are plotted along the Y-axis.
4. The points obtained by plotting the mid-points and frequencies are then joined by straight lines to obtain the frequency polygon.

Examples:

Sum 1: In a study of diabetic patients, the following data was obtained:

Age (in years)	30-40	40-50	50-60	60-70	70-80
No. of patients	6	9	18	13	8

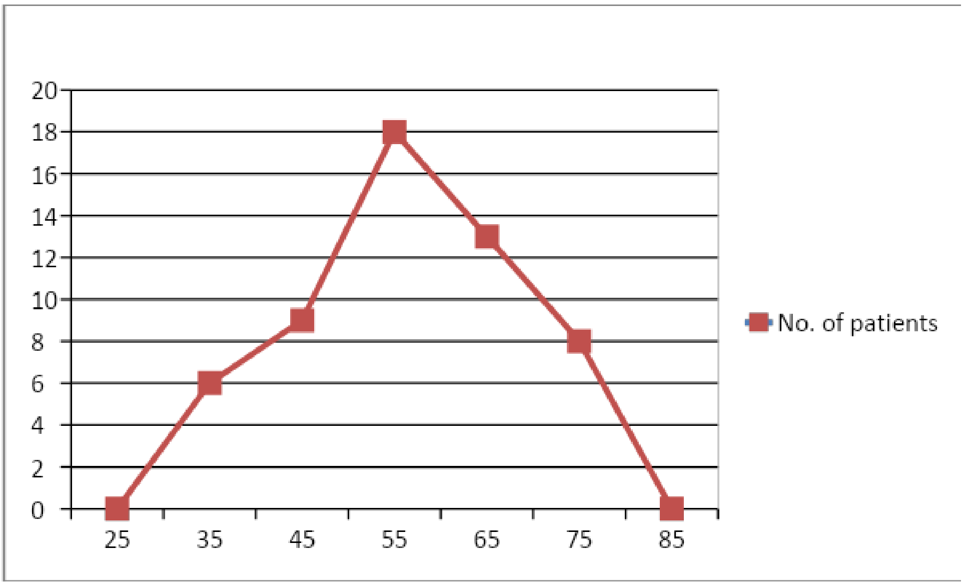


Fig 7: Frequency polygon showing age (in years) on the X-axis and number of patients on the Y-axis

Sum 2: Construct a frequency polygon for the following data.

Weekly earnings (in Rupees)	100-110	110-120	120-130	130-140	140-150
No. of workers	7	12	20	14	8

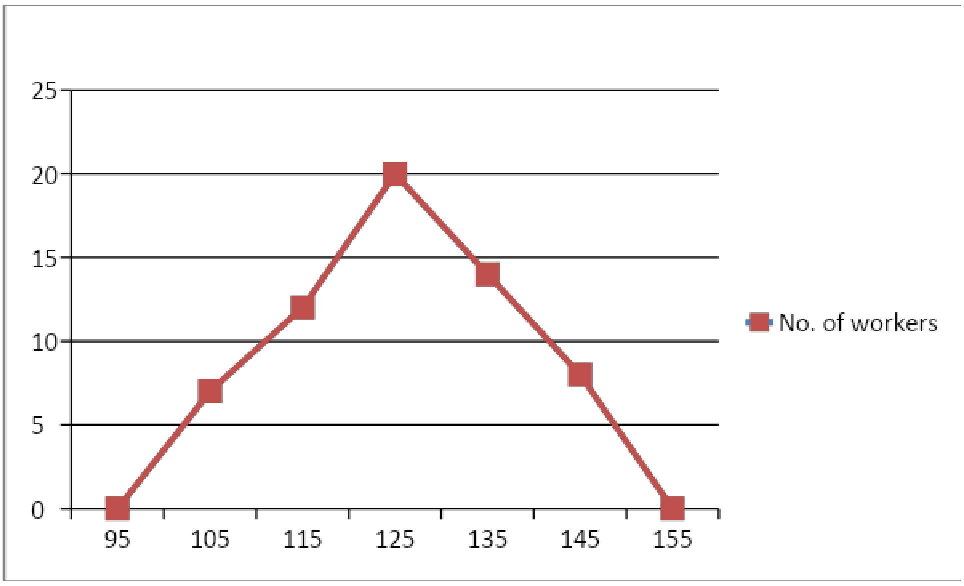


Fig 8: Frequency polygon showing weekly earnings on the X-axis and number of workers on the Y-axis

When to use a frequency polygon?

Frequency polygon is helpful when comparing two or more distributions. For instance, if a teacher wants to compare the progress of students in Half Yearly and Annual Examinations then it can be clearly shown with the help of frequency polygon. It is because the direction of each straight line in a frequency polygon is determined by the frequencies in two consecutive class intervals. Considering Fig. 2 it is seen that there is an increase in the number of workers from 7 (weekly earning – Rs. 100-110) to 12 (weekly earnings – Rs. 110-120). It is also seen that the highest number of workers (20) earn between Rs. 120-130 per week after which there is a decrease in the number of individuals (14) earning Rs. 130-140 per week. So the direct increase or decrease from one class interval to another can be shown more clearly with the help of a frequency polygon. On the other hand, the horizontal straight line on the top of each rectangle of a histogram represents the frequency and the vertical straight lines of the rectangle represent the class intervals.

Pie Chart

When we wish to represent a set of data in a simple but quite striking way, we use pie diagrams. In pie charts, data can be represented through portions of a circle. Pie charts always use relative frequencies to represent data. As we know the total angle of a circle is 360°, the data is to be represented in terms of proportions. An example is given below.

1. Draw a pie chart to represent the following data showing the monthly expenses of a man:

Food	Rs. 1000
Electricity	Rs. 2000
House rent	Rs. 4000
Miscellaneous	Rs. 3000
Total	Rs. 10,000

Solution: Each category of expenses must be converted to degrees in order to be represented in a pie chart. The conversion is as follows:

Category	Monthly expenses	Angle of the circle
Food	Rs. 1000	1000÷10000 × 360° = 36°
Electricity	Rs. 2000	2000÷10000 × 360° = 72°
House rent	Rs. 4000	4000÷10000 × 360° = 144°
Miscellaneous	Rs. 3000	3000÷10000 × 360° = 108°

Where 10000 denotes the total monthly expenses.

Remarks: Sum of all the angles of the circle must be 360°.

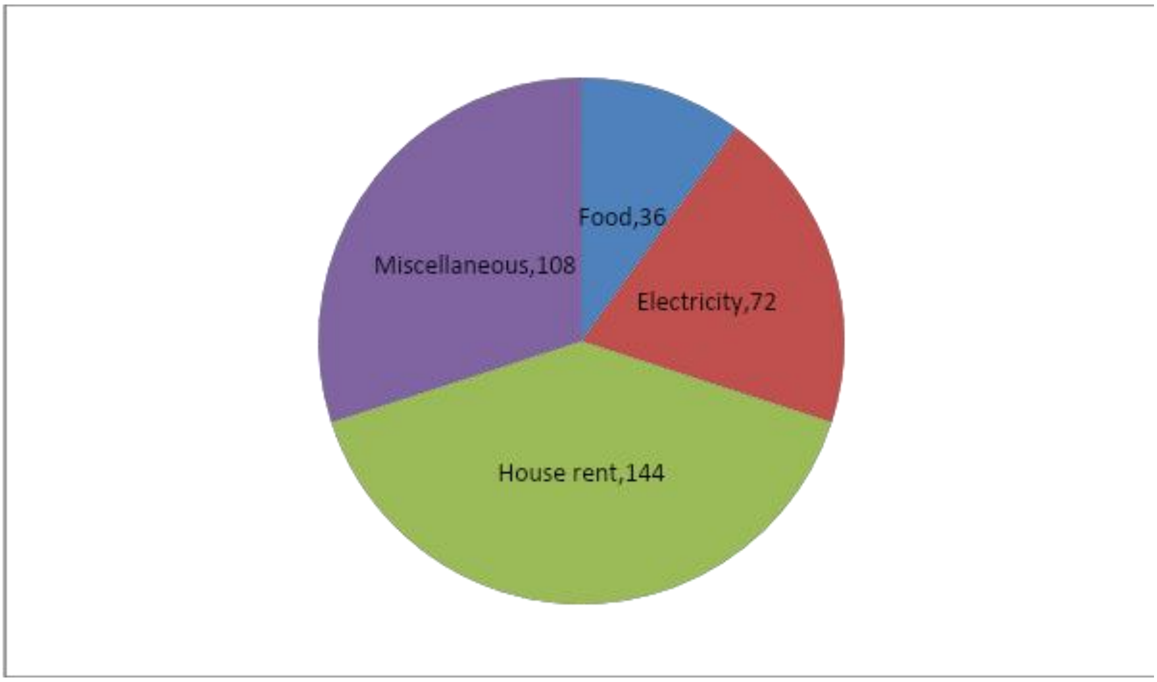


Fig 9: Pie diagram showing the monthly expenses in each category

2. Draw a pie chart to represent the preferences given by 200 students for stream selection.

Stream	No. of students
Science	45
Humanities	30
Commerce	50
Performing Arts	35
Sports	40

Solution: Each category of expenses must be converted to degrees in order to be represented in a pie chart. The conversion is as follows:

Stream	No. of students	Angle of the circle
Science	45	45÷200 × 360° = 81°

Humanities	30	30÷200 × 360° = 54°
Commerce	50	50÷200 × 360° = 90°
Performing Arts	35	35÷200 × 360° = 63°
Sports	40	40÷200 × 360° = 72°

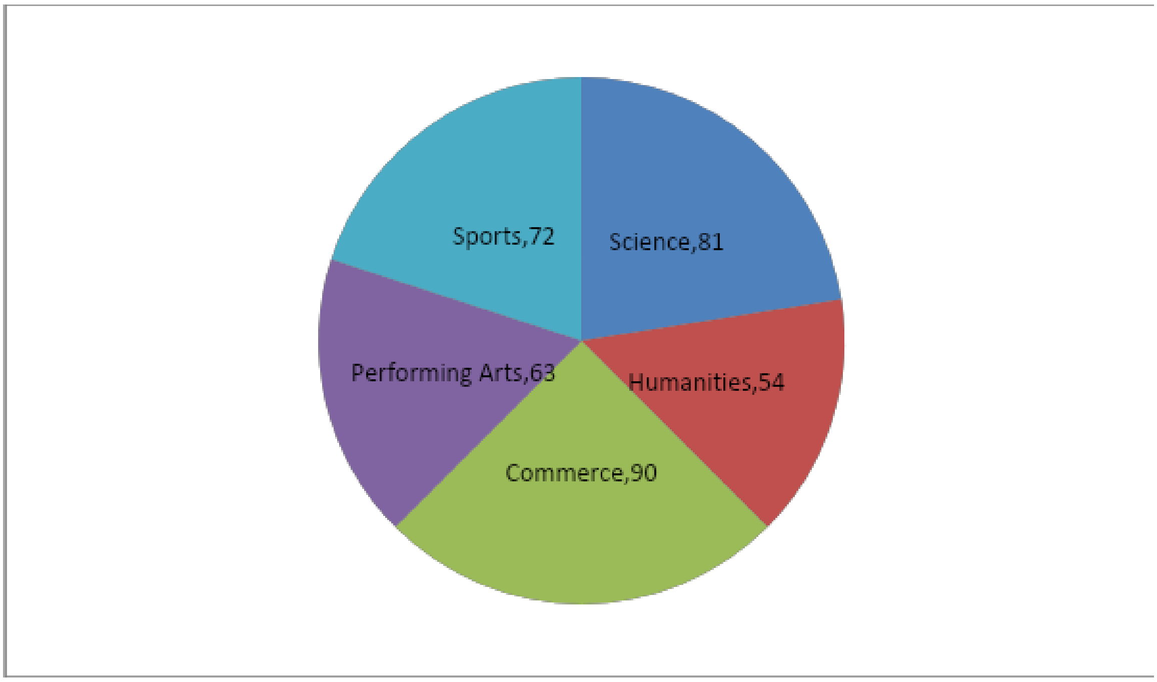


Fig 10: Pie diagram showing the preference of students

<u>Cumulative percentage frequency curve (ogive)</u>

When data is organized in the form of a cumulative frequency distribution and is represented with the help of a line graph then it is known as cumulative frequency graph. The graph is usually drawn by plotting the actual upper limits of the class intervals on the X-axis and the corresponding cumulative frequencies on the Y-axis.

The same procedure for plotting cumulative frequency graph is followed for plotting ogive (also known as cumulative percentage frequency curve) as well except that cumulative percentage frequencies are plotted on the Y-axis instead of cumulative frequencies. Examples are given below.

Example 1: The scores obtained by students in a test are as follows:

35, 28, 72, 68, 51, 17, 43, 56, 52, 26, 33, 37, 36, 41, 48, 46, 45, 44, 51, 57, 59, 53, 58, 56, 63, 61, 64, 69, 78, 77, 71, 82, 84, 91, 36, 47, 53, 57, 65, 87, 92, 67, 64, 74, 71, 53, 55, 58, 46, 43

Prepare a frequency distribution table and plot an ogive for the distribution.

Solution:

Class intervals	Actual limits of class intervals	Frequency	Cumulative Frequency	Cumulative Percentage Frequency
10-19	9.5-19.5	1	1	1÷50×100 = 2
20-29	19.5-29.5	2	3	3÷50×100 = 6
30-39	29.5-39.5	5	8	8÷50×100 = 16
40-49	39.5-49.5	9	17	17÷50×100 = 34
50-59	49.5-59.5	14	31	31÷50×100 = 62
60-69	59.5-69.5	8	39	39÷50×100 = 78
70-79	69.5-79.5	6	45	45÷50×100 = 90
80-89	79.5-89.5	3	48	48÷50×100 = 96
90-99	89.5-99.5	2	50	50÷50×100 = 100

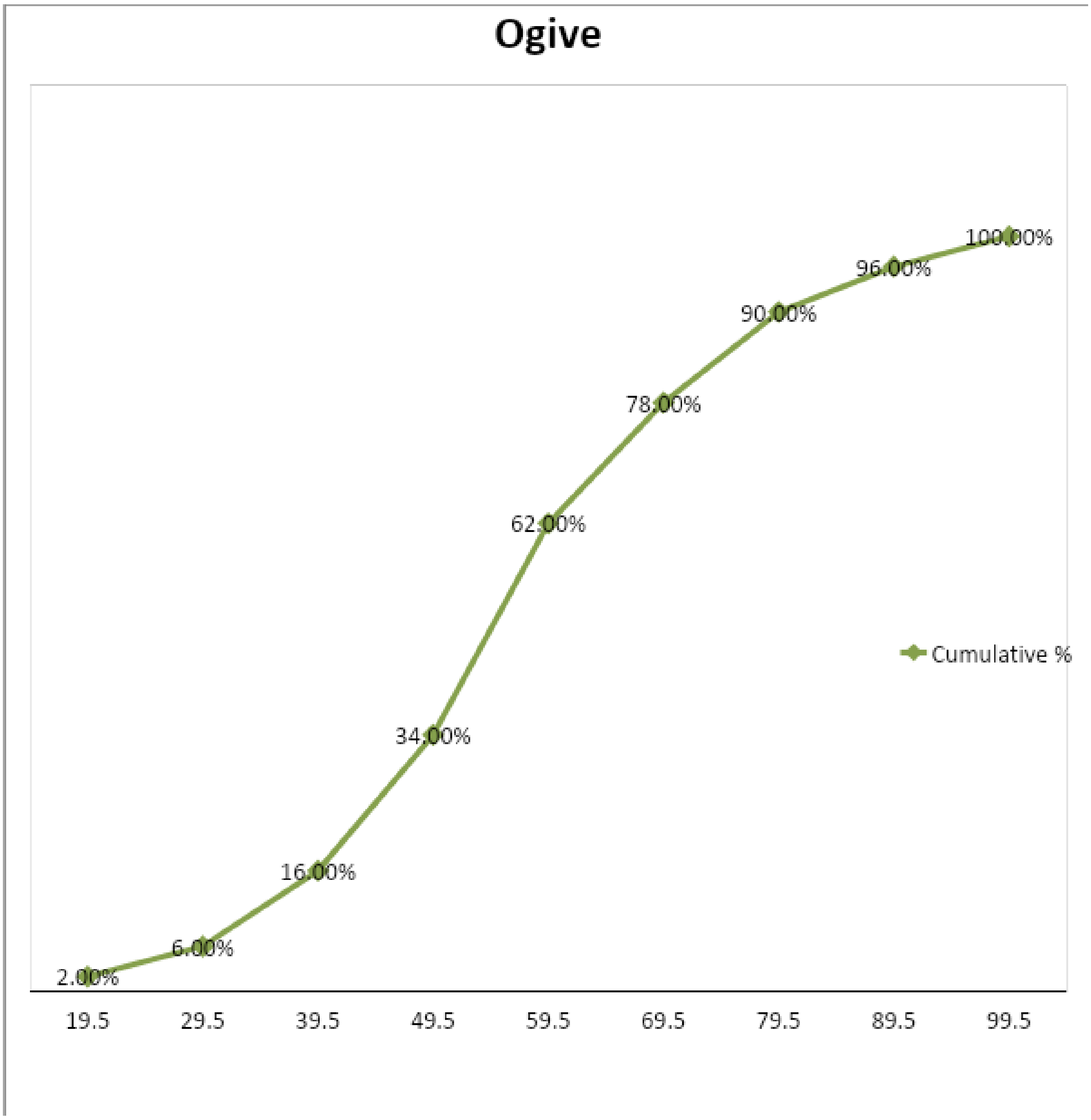


Fig 11: Ogive representing scores on X-axis and cumulative percentage frequencies on Y-axis

Example 2: Plot an ogive for the following frequency distribution:

Scores	Actual limits of class intervals	Frequency	Cumulative Frequency	Cumulative Percentage Frequency
30-34	29.5-34.5	2	2	3.08
35-39	34.5-39.5	4	6	9.23

40-44	39.5-44.5	7	13	20.00
45-49	44.5-49.5	8	21	32.31
50-54	49.5-54.5	10	31	47.69
55-59	54.5-59.5	12	43	66.15
60-64	59.5-64.5	9	52	80.00
65-69	64.5-69.5	6	58	89.23
70-74	69.5-74.5	4	62	95.38
75-79	74.5-79.5	2	64	98.46
80-84	79.5-84.5	1	65	100.00

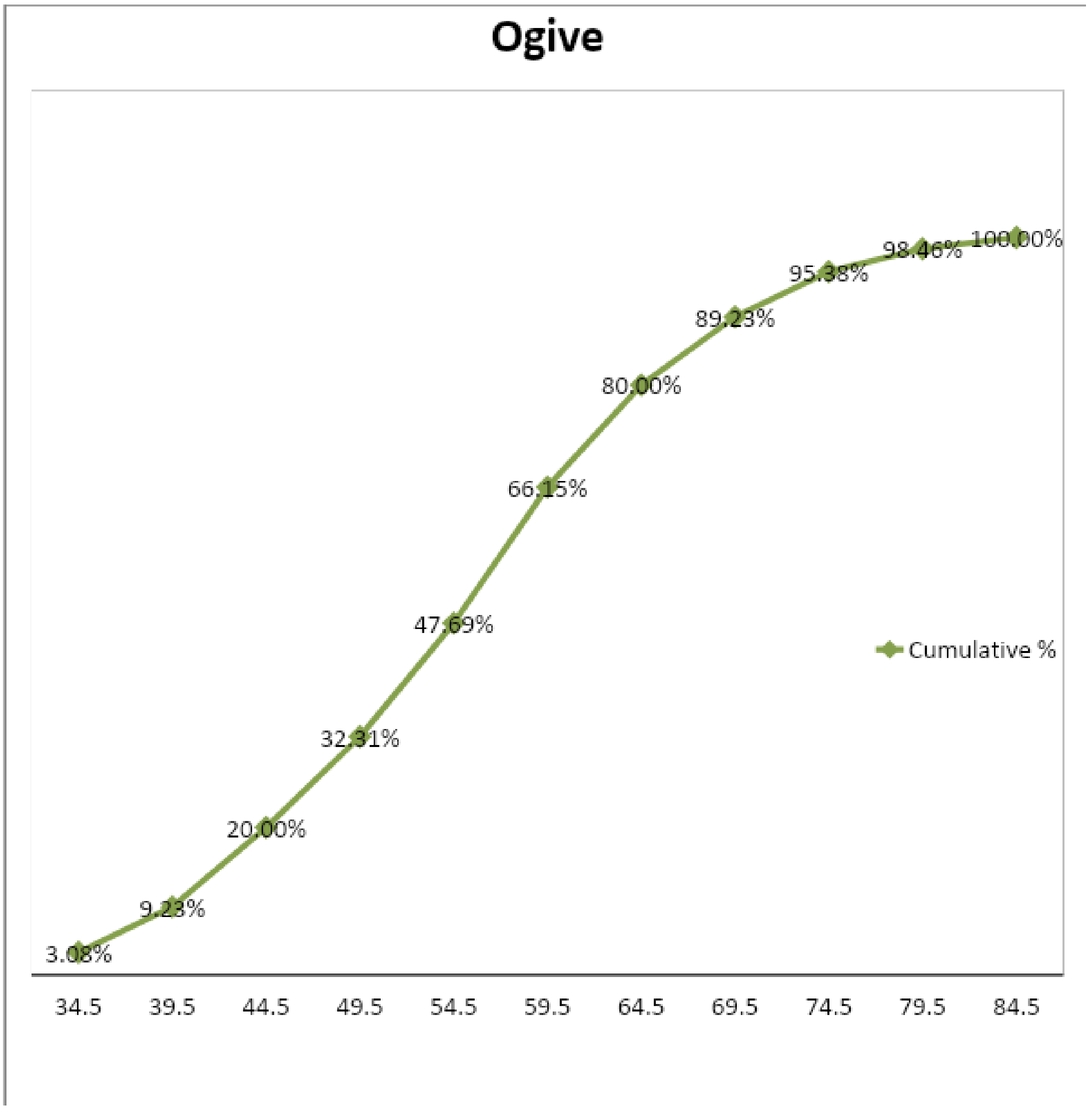


Fig 12: Ogive representing scores on X-axis and cumulative percentage frequencies on Y-axis

Applications

Bar Graph

- Bar diagrams are represented by vertical or horizontal bars
- Data in the form of raw scores, frequencies, percentages as well as computed statistics may be represented with the help of bar graphs.
- The lengths of the bars (represented on Y axis) are determined by the amount of the variable (height, weight, etc.), however, the width of the bars (represented on X axis) are not governed by any rules.

- It can be used to compare the relative amounts of some characteristic features possessed by two or more groups.

Histogram

- Histogram is a bar graph drawn on the basis of a frequency distribution.
- The horizontal or X-axis represents the class intervals and the vertical or Y-axis represents the frequencies.

Frequency Polygon

- Frequency polygon is a line graph drawn on the basis of a frequency distribution.
- It is represented by joining the mid points of class intervals with the help of straight lines.
- It is easier to compare two or more distributions using a frequency polygon than a histogram.

- It also helps in showing the continuity of a variable.

Pie Chart

- Using a Pie chart data can be represented by a circle of 360° divided into parts.

- Forming a pie chart or pie diagram is simple and “eye catching”.

Cumulative percentage frequency curve (ogive)

- It represents the cumulative percentage frequency distribution of a given set of scores
- It can also be used to compare two or more distributions
- Median, quartile, decile and percentile can be computed from ogive

Exercise

1. What are the advantages of representing data graphically?
2. What are the uses of bar graph?
3. How does a frequency polygon differ from histogram?
4. What is a pie chart? State its advantages.
5. What are the uses of ogive?

6. The scores obtained by students in English and Mathematics are given below. Draw a bar graph to represent the data.

Class	XA	XB	XC	XD	XE
English	72	68	81	69	73
Mathematics	89	87	88	75	82

7. The percentage of marks obtained by Ravi in four consecutive standards is given below. Graphically represent the data to show Ravi's progress.

Standard IV	Standard V	Standard VI	Standard VII
84%	88%	83%	85%

8. Draw a histogram and a frequency polygon for the following data:

Weight (in Kg)	40-50	50-60	60-70	70-80	80-90
No. of students	8	12	20	9	5

9. Plot histogram and frequency polygon for the following frequency distribution.

Scores	Frequency
10-14	4
15-19	6
20-24	9
25-29	14
30-34	11
35-39	8
40-44	7
45-49	5

10. Represent the data showing the career preferences of 500 students using a pie chart.

Career preferences	Number of individuals

Research	80
Doctor	120
Musician	50
Teacher	100
Business	150

11. Plot histogram, frequency polygon and ogive in separate graph sheets for the following distributions:

a.

Class intervals	Frequency
10-19	3
20-29	4
30-39	9
40-49	13
50-59	17
60-69	14
70-79	10
80-89	7
90-99	3
	N=80

b.

Scores	Frequency
35-39	3
40-44	4
45-49	7
50-54	12
55-59	15
60-64	20
65-69	14
70-74	11
75-79	8
80-84	4
85-89	2
	N=100

Chapter 3

MEASURES OF CENTRAL TENDENCY

It is generally said that most individuals possess average intelligence, weight, height and so on. Human attributes may be compared based on some predefined measures. For instance, an individual with an IQ score of 105 is said to fall in the "Average" category based on the predefined range, i.e., 90-110 in which his/her score falls. These ranges are standardized on the basis of scores obtained by members of large samples in a given time in a particular area with a particular socio-cultural background and norms are developed accordingly. The general trend is that most of the scores gravitate towards a central point and only a few deviate markedly from that point. This tendency of the majority of scores to gravitate towards the centre is known as central tendency. In a classroom, it is generally noticed that most of the students score near the average value and only a few deviate noticeably from the average, be it on the positive or the negative side. A similar trend may be observed in case of other human attributes too such as height, weight, blood pressure, etc. So, a measure of central tendency is advisable when there is a requirement of comparison between the performance of a group with that of a previously standardized reference group. The commonly used measures of central tendency include the mean, median and mode.

The Arithmetic Mean

The arithmetic mean or average refers to the sum total of a given set of scores divided by the number of scores. It is the most stable measure of central tendency as it takes all the scores in a given distribution into consideration. Thus every score contributes to the mean value in a given distribution. Mean is also required when there is a need for computation of other statistical measures such as standard deviation, correlation, etc.

The computational formula for arithmetic mean is

$$\bar{X} = \frac{\sum X}{N}$$

Where $\bar{X}$ (pronounced as X bar) = the arithmetic mean

$\sum X$ = the sum of all X values (scores)

N = the number of scores in a given data set

For example: the age (in years) of 8 weightlifters in Olympic Games are as follows:

25, 22, 26, 24, 21, 25, 23, 27

The mean age of the weightlifters may be computed by adding the ages of all the 8 weightlifters and dividing the obtained value by 8.

Therefore, the total age of weightlifters = (25+22+26+24+21+25+23+27) years = 193 years

The mean age of weightlifters = 193 ÷ 8 = 24.125 years

However, it is to be noted that the mean value may be affected if there are extreme scores in the series.

i) Calculation of the mean when data are ungrouped

Formula: M = $\frac{\sum X}{N}$
Where M = Mean
$\sum$X = Sum of scores or other measures
N = Number of scores / Measures in the series

Sum 1: The weights of 5 boys of age 12-year old are 40kg, 37kg, 42kg, 45kg and 39kg respectively. What is their mean weight?

Solution: Mean = $\frac{\sum X}{N}$

= (40+37+42+45+39)/5 kg

= 40.6 kg

Sum 2: If a man earns 60rs, 75rs, 65rs, 55rs, 85rs and x rs on 6 consecutive days and his mean daily wage is 70rs, then find the value of x.

Solution: M = $\frac{\sum X}{N}$

70 = (60+75+65+55+85+ x) / 6

Therefore, 420 = 340 + x

or, x = 420 – 340

or, x = 80

Therefore, x = 80rs

ii) Calculation of the mean from the data grouped into a frequency distribution:

Formula: M = $\frac{\sum fX}{N}$

Where M = Mean,

$\sum fX$ = sum of the midpoints weighted by their frequencies, that is, sum of the product of the observations and its respective frequency.

N = Number of scores (Measures in the series) = $\sum_{1}^{n} \quad f$

Sum 1: Find out the mean score of the following 50 students of the 10th standard in the Mathematics test.

Class Intervals	Midpoint (X)	Frequency (f)	fX
1-10	5.5	0	0
11-20	15.5	3	46.5
21-30	25.5	4	102.0
31-40	35.5	3	106.5
41-50	45.5	5	227.5
51-60	55.5	8	444.0
61-70	65.5	7	458.5
71-80	75.5	5	377.5
81-90	85.5	9	769.5
91-100	95.5	6	573.0
Total		50	3105.0

Solution: M = $\frac{\sum fX}{N}$

= 3105/50

= 62.1

iii) Calculation of the mean from combined samples or groups.

Formula: $M_{comb} = (N_1 M_1 + N_2 M_2 + ... + N_n M_n)/(N_1 + N_2 + ... + N_n)$

Sum 1: The mean mathematics score of students of the 10th standard in section A is 90 (N=50) and that in section B is 85 (N=70).

Find out the mean mathematics score for the sections combined.

Solution: $M_{comb} = (N_1\ M_1+N_2\ M_2)/(N_1 + N_2)$

$= [(50\times90)+(70\times85)]/(50+70)$

$= (4500+5950)/120$

$= 87.083$

Sum 2: The mean monthly salary of 30 TCS employees is Rs.90,000, that of 25 Cognizant employees is Rs.80,000, and that of 35 Microsoft employees is Rs.95,000. Find out the mean monthly salary of employees of three companies combined.

Solution: M_1 = Rs.90,000 , $N_1 = 30$

M_2 = Rs.80,000 , $N_2 = 25$

M_3 = Rs.95,000 , $N_3 = 35$

$M_{comb} = ((N_1\ M_1+N_2\ M_2+N_3\ M_3)/(N_1+N_2+N_3)$

$= (30\times90{,}000+25\times80{,}000+35\times95{,}000)/(30+25+35)$

$= (27{,}00{,}000+20{,}00{,}000+33{,}25{,}000)/90$

= Rs.89166.67

The Median

The median is the second measure of central tendency which is defined as the middle score or value score once all the scores are placed in rank order. It refers to the mid-value of a given set of scores where the scores are arranged in ascending or descending order of magnitude. However, it does not take the extreme scores into account. Basically, it is the value that lies in the middle of the sample.

1) **Calculation of the median when data are ungrouped.**
   First of all arrange the data either in ascending or in descending order then find the median from the following formula.
   Median = the (N+1)/2th measure in order of size.

i) When N is odd

Sum 1: The spelling test scores of 9 students of the 7th standard are 12,15,11,13,12,19,16,14,17. Find out the median.

Solution : The scores are arranged in ascending order – 11,12,12,13,(14),15,16,17,19.

The median is the midpoint of the series.

Therefore, the median is 14.

ii) When N is even

Sum 1: The spelling test scores of 6 students of the 7th standard are 15,18,17,14,16,19. Find out the median.

Solution: The scores are arranged in ascending order – 14,15,16,17,18,19.

Here the midpoint is in between 16 and 17. Here the median is (16 + 17) / 2 = 16.5.

2) Calculation of the median when the data are grouped into a frequency distribution.

Formula : Median = l +{(N/2-F)/fm} × i

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

N/2 = one half of the total number of scores

F = sum of the scores on all intervals below l

fm = frequency (number of scores) within the interval upon which the median falls

i = length of class interval

Sum 1 : Find out the median of the following frequency distribution.

Class Intervals (scores)	Actual limits of class intervals	Frequency	Cumulative Frequency
1-10	0.5-10.5	6	6
11-20	10.5-20.5	7	13
21-30	20.5-30.5	11	24
31-40	30.5-40.5	14	38
41-50	40.5-50.5	18	56

51-60	50.5-60.5	15	71
61-70	60.5-70.5	13	84
71-80	70.5-80.5	12	96
81-90	80.5-90.5	8	104
91-100	90.5-100.5	6	110
		N = 110	

The actual limits of class intervals have been stated taking into consideration that in a continuous series scores are thought of as distances along a continuum rather than as discrete points. In case of mental measurement, a score is a unit distance between two limits. Thus a score of 120 in an intelligence test might represent the interval between 119.5 to 120.5. This rule has been followed throughout the book to determine the exact limits of class intervals.

As stated before, median refers to the mid value. So the central value needs to be located. In the present distribution, N=110 is an even number. So by applying the previously mentioned rules it is implied that the median will fall somewhere between the 55th and 56th items in the given distribution. By referring to the column of cumulative frequency it can be seen that the score representing the median falls in the class interval of 41-50.

By applying the formula the median of the distribution may be found out.

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

$$= 40.5 + \{(55-38)/18\} \times 10$$

$$= 40.5 + 9.44$$

$$= 49.94$$

The Mode

The meaning of Mode in statistics is similar to that of in English - “fashionable”. So, mode is the score in an ungrouped distribution which is “in fashion” or appears with the greatest frequency. It is a “crude” or “empirical” form of measurement. It is the crudest measure of central tendency. In a given set of observations, the score which occurs the most number of times is the mode. Likewise, if two scores occur the most number of times in a given set of scores then the distribution is referred to as bimodal.

For example, in a Psychology test, the scores obtained by 10 college students are as follows:

12, 15, 19, 11, 12, 16, 11, 17, 12, 13

In this example, the modal value is 12 simply because it occurs the most number of times.

Sum 1 : In a series of scores – 12,10,11,17,11,14,19,16,11,14, the most often recurring measure, namely, 11, is the crude or empirical mode. So the mode is 11.

In a frequency distribution, the true mode is the point (or peak) of greatest concentration in the distribution; that is, the point at which more measures fall than at any other point.

Formula : Mode = 3Median – 2Mean

Calculation of the mean by the short method

Example:

Class Intervals (Scores)	Midpoint (X)	f	x'	fx'
95-99	97	2	4	8
90-94	92	3	3	9
85-89	87	4	2	8
80-84	82	6	1	6
75-79	77	9	0	0
70-74	72	8	-1	-8
65-69	67	6	-2	-12
60-64	62	5	-3	-15
55-59	57	4	-4	-16
50-54	52	3	-5	-15
		50		-35

Assumed Mean (A.M): Assumed mean is the value for which x' = $\frac{X-77}{5}$ is made 0. Here if we consider X = 77 then x' = $\frac{X-77}{5}$ = 0. Hence for this example assumed mean is 77.

Correction (c) = -35/50, = -0.7, x' = $\frac{X-77}{5}$

c i = -0.7 x 5 [Since size of intervals = 5]

= -3.5

Therefore, Mean = A.M + ci

= 77 + (-3.5)

= 73.5

Problems

1) Calculate the mean, median and mode for the following frequency distributions :

a)

Class Intervals	Frequency (f)	Midpoint (X)	fX	C.F.
35-39	2	37	74	2
40-44	4	42	168	6
45-49	6	47	282	12
50-54	7	52	364	19
55-59	5	57	285	24
60-64	4	62	248	28
65-69	3	67	201	31
70-74	3	72	216	34
	N = 34		∑fX = 1838	

Solution:

Mean = $\frac{\sum fX}{N}$

=1838/34

= 54.06

Median = l + {(N/2-F)/fm} × i *C.F. = Cumulative Frequency

Here l = 49.5, N/2 = 17, F = 12, fm = 7, i = 5

Median = 49.5 + {(17-12)/7} × 5

= 49.5 + 3.57

= 53.07

Mode = 3Median – 2Mean

= (3×53.07) – (2×54.06)

= 159.21 – 108.12

= 51.09

b)

Class Intervals	Frequency (f)	Mid Point (X)	fX	Cumulative Frequency
10-19	4	14.5	58.0	4
20-29	5	24.5	122.5	9
30-39	6	34.5	207.0	15
40-49	9	44.5	400.5	24
50-59	10	54.5	545.0	34
60-69	8	64.5	516.0	42
70-79	7	74.5	521.5	49
80-89	5	84.5	422.5	54
90-99	4	94.5	378.0	58
	N = 58		∑fX = 3171.0	

Solution: Mean = $\frac{\sum fX}{N}$

= 3171/58

= 54.67

Median = l+ {(N/2-F)/fm} × i

Here, l = 49.5, N/2 = 29, F = 24, fm = 10, i = 10

Therefore, Median = 49.5 + {(29-24)/10} × 10

= 49.5 + 5

= 54.5

Mode = 3Median - 2Mean

= (3×54.5) – (2×54.67)

= 163.5 – 109.34

= 54.16

c)

Class intervals	Frequency (f)	Midpoint	x’	fx’	Cumulative frequency
70-71	2	70.5	-5	-10	2
72-73	2	72.5	-4	-8	4
74-75	3	74.5	-3	-9	7
76-77	4	76.5	-2	-8	11
78-79	6	78.5	-1	-6	17
80-81	7	80.5	0	0	24
82-83	5	82.5	1	5	29
84-85	4	84.5	2	8	33
86-87	2	86.5	3	6	35
88-89	2	88.5	4	8	37
90-91	1	90.5	5	5	38
Total	N=38			-9	

Assumed Mean (A.M.) = 80.5, this is also called center point so x’ = (Mid point – 80.5)/2.

Correction (c) = -9/38 =-0.237

c × i = -0.237 × 2 = -0.474

Mean = A.M. + ci

= 80.5 + (-0.474) = 80.026

Median = l + {(N/2-F)/fm} × i

Here l = 79.5, N/2 = 19, F = 17, fm = 7, i = 2

Therefore, Median = 79.5 + {(19-17)/7} × 2

= 79.5 + 0.57 = 80.07

Mode = 3Median – 2Mean

= (3×80.07) – (2×80.026)

= 240.21 – 160.052 = 80.158

d)

Class intervals	Frequency (f)	Midpoint	x’	fx’	Cumulative frequency
61-65	2	63	-5	-10	2
66-70	2	68	-4	-8	4
71-75	5	73	-3	-15	9
76-80	6	78	-2	-12	15
81-85	8	83	-1	-8	23
86-90	12	88	0	0	35
91-95	9	93	1	9	44
96-100	7	98	2	14	51
101-105	4	103	3	12	55
106-110	3	108	4	12	58
111-115	2	113	5	10	60
Total	N=60			4	

Assumed Mean (A.M.) = 88

Correction (c) = 4/60 = 0.067

c × i = 0.067 × 5 = 0.335

Mean = A.M. + ci

= 88 + 0.335 = 88.335

Median = l + {(N/2-F)/fm} × i

Here, l = 85.5, N/2 = 30, F = 23, fm = 12, i = 5

Therefore, Median = 85.5 + {(30-23)/12} × 5

= 85.5 + 2.92 = 88.42

Mode = 3Median – 2Mean

= (3×88.42) – (2×88.335)

= 265.26 – 176.67 = 88.59

2) Test scores for a class of 30 students are as follows:

95, 82, 79, 58, 91, 87, 70, 58, 63, 67, 75, 92, 77, 65, 68, 59, 84, 80, 92, 71, 58, 69, 78, 86, 76, 67, 81, 73, 90, 82.

Test scores	Frequency
51-60	
61-70	
71-80	
81-90	
91-100	

a) Copy and complete the above table.
b) Compute the mean, median and mode of the distribution.

Solution:

Test scores	Frequency (f)	Midpoint (X)	fX	Cumulative frequency
51-60	4	55.5	222.0	4
61-70	7	65.5	458.5	11
71-80	8	75.5	604.0	19
81-90	7	85.5	598.5	26
91-100	4	95.5	382.0	30
Total	N=30		∑fX=2265.0	

Mean = $\frac{\sum fX}{N}$ = 2265/30 = 75.5

Median = l + {(N/2-F)/fm} × i

Here, l = 70.5, N/2 = 15, F = 11, fm = 8, i = 10

Therefore, Median = 70.5 + {(15-11)/8} × 10

= 70.5 + 5 = 75.5

Mode = 3Median – 2Mean

= (3×75.5) – (2×75.5) = 75.5

3) The spelling test scores of 15 students of the 5th standard are as follows:

9, 7, 17, 11, 8, 16, 13, 22, 15, 19, 10, 13, 18, 20, 13.

i) Calculate the mean, median and mode of the scores.
ii) Add 3 to each score and compute the mean, median, and mode of the new set of scores.
i) Mean = Sum of scores ÷ Total number of scores

Sum of scores = 9+7+17+11+8+16+13+22+15+19+10+13+18+20+13

= 211

Therefore, Mean = 211/15 = 14.067

To find out the median, the scores are to be arranged in ascending order.

7, 8, 9, 10, 11, 13, 13, (13), 15, 16, 17, 18, 19, 20, 22

The midpoint is 13. Hence it is the median.

The mode is 13.

ii) The new scores are:

9+3, 7+3, 17+3, 11+3, 8+3, 16+3, 13+3, 22+3, 15+3, 19+3, 10+3, 13+3, 18+3, 20+3, 13+3

12, 10, 20, 14, 11, 19, 16, 25, 18, 22, 13, 16, 21, 23, 16

Mean = Sum of scores/Total number of scores

Sum of scores = 12+10+20+14+11+19+16+25+18+22+13+16+21+23+16

= 256

Therefore, Mean = 256/15

= 17.067

To find out the median, the scores are to be arranged in ascending order.

10, 11, 12, 13, 14, 16, 16, (16), 18, 19, 20, 21, 22, 23, 25

The midpoint is 16. Hence, it is the Median.

The Mode is 16.

Another method for computation of mode for grouped data:

The mode of a distribution can be directly computed without calculating the mean and median with the help of the following formula:

$$M_o = L + \{f_1 \div (f_1+f_{-1}) \times i\}$$

Where

L = Lower limit of the class in which the mode is supposed to lie

i = Class interval

f_1 = Frequency of the class adjacent to the modal class for which lower limit is greater than that for the modal class

f_{-1} = Frequency of the class adjacent to the modal class for which the lower limit is less than that for the modal class

It may be illustrated with the help of the following example:

Example:

Scores	Frequency
25-29	1
30-34	3
35-39	6
40-44	10
45-49	14
50-54	11
55-59	7
60-64	5
65-69	3
	N=60

The mode supposedly lies in the interval 45-49.

By applying the formula

$M_o = L + \{f_1 \div (f_1+f_{-1}) \times i\}$

$= 44.5 + \{11 \div (11 + 10) \times 5\}$

$= 44.5 + (55/21)$

$= 44.5 + 2.62$

$= 47.12$

Application of the measures of Central Tendency

Mean:

- It is the most simple, stable and accurate measure of central tendency
- It takes all the scores in a given set of data into account and hence gives equal weightage to all the scores.
- The mean value represents the average value of a given set of data.
- The mean value is often required before proceeding to compute other statistical measures such as standard deviation.
- However, it is to be noted that the mean value may be affected by extreme scores. In such a case it may not be wise to compute the mean.

Median:

- It is the central value in a given distribution above and below which 50% of the cases lie.
- It does not take the extreme scores into account and thus may be considered as the most representative central measure.

Mode:

- It is the crudest measure of central tendency
- Mode may be computed when a quick and approximate measure of central tendency would suffice.
- Mode is that value or observation amongst a given set of scores which occurs most frequently.

Exercise

1. What do you understand by central tendency?
2. Define arithmetic mean. Delineate the different ways of computing the arithmetic mean.
3. What is median? State how the median can be computed from grouped and ungrouped data.
4. What do you understand by the term ‘mode’? State the different procedures for computing mode.
5. The scores obtained by 6 students in Mathematics are as follows:
   42, 38, 37, 41, 49, 43. Find out the mean score.
6. Compute the median for the following data:
   i. 18, 13, 17, 19, 12, 11, 16
   ii. 9, 7, 3, 6, 2, 8, 1, 4
7. The Spelling Test scores obtained by 10 students are as follows:

14, 12, 14, 11, 19, 12, 14, 17, 13, 16. Find out the mode.

8. Calculate the mean, median and mode for the following frequency distributions:

i.

Class intervals	Frequency
20-24	8
25-29	9
30-34	10
35-39	13
40-44	19
45-49	12
50-54	9
55-59	8
60-64	2

ii.

Class intervals	Frequency
45-49	2
50-54	3
55-59	6
60-64	10
65-69	13
70-74	16
75-79	15
80-84	9
85-89	8
90-94	5
95-99	3

9. Calculate the mode of the following distribution without computing the mean and median

Scores	Frequency
20-24	2
25-29	5
30-34	7
35-39	10
40-44	15
45-49	12
50-54	9
55-59	6
60-64	4

Chapter 4
MEASURES OF VARIABILITY

No two individuals are alike, not even identical twins. People differ in their mental abilities, dispositions, attitudes, personality characteristics, problem-solving abilities and so on. These unique variations are what make us human.

Central tendency gives us a sight of the typical score in a sample. And, generally the next step for us is to find out the measures of variability. Variation or variance is that characteristic of a data which represents how much "scattered" or "spread" the individual scores are. More constructively, measures of variability indicate how much variation is there in the sample or population. Measures of variability may be computed when it becomes necessary to know how individuals differ from each other based on certain characteristic features. For instance, a teacher may be interested to know the mean score obtained by students in a test along with the deviation of each student's score from the mean in order to evaluate individual performance. In other words, the teacher may be eager to know how near or far each student's score is located from the mean so that s/he could identify the good and poor performers and select appropriate intervention measures. Suppose a teacher administers a test of reasoning to a group of 100 individuals - 50 males and 50 females. The mean scores are as follows: Males = 55 (score ranges from 25 to 80) and females = 55.2 (score ranges from 39 to 76). Thus it can be clearly seen that the scores of males are more variable, i.e., covers a wider range than their female counterparts, considering the fact that there is almost negligible difference (0.2) between the two mean values. When a group is homogeneous, i.e., it consists of individuals of more or less similar ability then it is expected that the range of scores will be relatively short and the variability small. On the other hand, if the group consists of individuals of widely differing capacities (heterogeneous) then it is expected that the variability will be greater.

Some of the frequently used measures of variability are described below.

1) Range – The most simple way of obtaining an idea about the spread of scores is to compute the range. It is the interval between the highest and the lowest scores. It is calculated by subtracting the lowest score in a series from the highest. Range does not give us any clue of what is happening in between these scores.
   e.g.: Suppose the highest score obtained in English by students of the 10th standard is 82 and the lowest score is 58.
   Range = 82-58 = 24

2) Quartile Deviation
   A distribution may be divided into four quartiles. Q_1 or the first quartile is the point below which 1/4th of the scores lie and Q_3 or the third quartile is the point below which 3/4th of the scores lie. Q_2 is the median. Quartile deviation is generally computed when the researcher does not want the extreme scores to influence the measure of variation.

$Q = (Q_3 - Q_1)/2$
$Q_1 = l + i \times \{(N/4\text{-cum } f)/fq\}$
$Q_3 = l + i \times \{(3N/4\text{-cum } f)/fq\}$

l = the exact lower limit of the interval in which the quartile falls
i = the length of the interval
cum f = cumulative frequency up to the interval which contains the quartile
fq = the frequency on the interval containing the quartile

3) Average Deviation (A.D.) or Mean deviation (M.D.)

Garrett (1971) defined Average Deviation as the mean of deviations of all the separate scores in the series taken from their mean (occasionally from the median or mode). That is, in a sample every individual score will deviate from the mean. By computing mean deviation we can get an indication of how much the group as a whole differs from the sample mean.

$AD = \sum|x|/N$ [For ungrouped data]
$x = X\text{-}M$
$|x|$ signifies that in the deviation values the + or – signs are ignored.

$AD = \sum|fx|/N$ [For grouped data]

4) Standard Deviation (S.D./σ)

Standard deviation is a more informative measure of variability, defined as the degree to which the scores in a data deviate around the mean. It is the most stable measure of variability and it takes the mean as a reference point. Standard deviation tells us how much the scores deviate from the mean on either side.

$\sigma = \surd \sum\{(X\text{-}M)^2 /N\} = \surd (\sum x^2 /N)$ [In case of ungrouped data)
X = Individual score
M = Mean of the given set of scores
N = Total no. of scores
x = Deviation of each score from the mean

$\sigma = \surd \{\sum fx^2)/N\}$ [In case of grouped data]

Examples:

Sum 1: Find the A.D. and S.D. of the following set of scores – 8, 12, 6, 7, 9, 14.

Scores (X)	x=X-M	|x|	x^2
8	-1.3	1.3	1.69
12	2.7	2.7	7.29
6	-3.3	3.3	10.89
7	-2.3	2.3	5.29
9	-0.3	0.3	0.09
14	4.7	4.7	22.09
Total = 56		14.6	47.34

Mean (M) = 56/6 = 9.3
A.D. = $\sum|x|/N$ = 14.6/6 = 2.43
S.D. = $\sqrt{\ } \sum x^2 /N = \sqrt{\ } (47.34/6)$ = 2.81

Sum 2: The scores obtained by 8 students in a mathematics test are as follows:

42, 36, 45, 28, 33, 30, 29, 38. Calculate the A.D. and S.D.

Scores (X)	x=X-M	|x|	x^2
42	6.9	6.9	47.61
36	0.9	0.9	0.81
45	9.9	9.9	98.01
28	7.1	7.1	50.41
33	2.1	2.1	4.41
30	5.1	5.1	26.01
29	6.1	6.1	37.21
38	2.9	2.9	8.41
Total = 281		41.0	272.88

Mean (M) = 281/8 = 35.1
A.D. = $\sum|x|/N$ = 41/8 = 5.125
S.D. = $\sqrt{\ } \sum x^2 /N = \sqrt{\ } (272.88/8)$ = 5.84

Sum 3 : Calculate Q from the following frequency distribution.

Class intervals	Frequency (f)	Cumulative frequency
90-99	7	50
80-89	6	43
70-79	8	37
60-69	11	29
50-59	9	18
40-49	5	9
30-39	4	4
	N = 50	

Q_1 = l + i × {(N/4 – cum f) ÷ fq}

= 49.5 + 10 × {(12.5-9)/9}
= 49.5 + (10 × 0.389)
= 49.5 + 3.89 = 53.39

Q_3 = l + i × {(3N/4 – cum f) ÷ fq}

= 79.5 + 10 × {(37.5-37)/6}
= 79.5 + (10 × 0.083)
= 79.5 + 0.83 = 80.33

Q = $(Q_3 – Q_1)/2$

= (80.33-53.39)/2 = 13.47

Sum 4: Calculate Q from the following frequency distribution

Class intervals	Frequency (f)	Cumulative frequency
80-84	9	90
75-79	11	81
70-74	10	70
65-69	12	60
60-64	15	48
55-59	13	33
50-54	8	20
45-49	6	12
40-44	6	6
	N = 90	

$Q_1 = l + i \times \{(N/4 – \text{cum } f) \div fq\}$

$= 54.5 + 5 \times \{(22.5-20)/13\}$

$= 54.5 + (5\times0.19) = 55.45$

$Q_3 = l + i \times \{(3N/4 – \text{cum } f) \div fq\}$

$= 69.5 + 5 \times \{(67.5-60)/10\}$

$= 69.5 + (5\times0.75) = 73.25$

$Q = (Q_3 – Q_1)/2$

$= (73.25-55.45)/2 = 8.9$

Sum 5: Calculate Q from the following frequency distribution

Class intervals	Frequency (f)	Cumulative frequency
90-99	3	60
80-89	5	57
70-79	6	52
60-69	9	46
50-59	14	37
40-49	11	23
30-39	7	12
20-29	3	5
10-19	2	2
	N = 60	

$Q_1 = l + i \times \{(N/4 - \text{cum f}) \div fq\}$

$= 39.5 + 10 \times \{(15\text{-}12)/11\}$

$= 39.5 + (10 \times 0.27) = 42.2$

$Q_3 = l + i \times \{(3N/4 - \text{cum f}) \div fq\}$

$= 59.5 + 10 \times \{(45\text{-}37)/9\}$

$= 59.5 + (10\times0.89) = 68.4$

$Q = (Q_3 - Q_1)/2$

$= (68.4\text{-}42.2)/2 = 13.1$

Sum 6: Calculate Q, A.D. and S.D. from the following frequency distribution

Class intervals	Frequency (f)	Cumulative frequency	Midpoint (X)	fX	x=X-M	fx	|fx|	f x^2
90-99	2	40	94.5	189	29.25	58.5	58.5	1711.1
80-89	5	38	84.5	422.5	19.25	96.25	96.25	1852.8
70-79	8	33	74.5	596	9.25	74	74	684.5
60-69	12	25	64.5	774	-0.75	-9	9	6.75
50-59	7	13	54.5	381.5	-10.75	-75.25	75.25	808.9
40-49	4	6	44.5	178	-20.75	-83	83	1722.2
30-39	2	2	34.5	69	-30.75	-61.5	61.5	1891.1
Total	N=40			2610			457.5	8677.35

Mean (M) = 2610/40 = 65.25

A.D. = ∑|fx|/N = 457.5/40 = 11.44

S.D. = $\sqrt{\{(\sum fx^2)/N\}} = \sqrt{(8677.35/40)} = 14.73$

$Q_1 = l + i \times \{(N/4 – \text{cum f}) \div fq\}$

$= 49.5 + 10 \times \{(10-6)/7\}$

$= 49.5 + 5.71 = 55.21$

$Q_3 = l + i \times \{(3N/4 – \text{cum f}) \div fq\}$

$= 69.5 + 10 \times \{(30-25)/8\}$

$= 69.5 + 6.25 = 75.75$

$Q = (Q_3-Q_1)/2$

$= (75.75-55.21)/2 = 10.27$

Sum 7: Calculate Q, A.D. and S.D. from the following frequency distribution

Class intervals	Frequency (f)	Cumulative frequency	Midpoint (X)	fX	x=X-M	fx	|fx|	fx^2
80-84	6	70	82	492	18.2	109.2	109.2	1987.44
75-79	7	64	77	539	13.2	92.4	92.4	1219.68
70-74	9	57	72	648	8.2	73.8	73.8	605.16
65-69	10	48	67	670	3.2	32	32	102.4
60-64	13	38	62	806	-1.8	-23.4	23.4	42.12
55-59	12	25	57	684	-6.8	-81.6	81.6	554.88
50-54	6	13	52	312	-11.8	-70.8	70.8	835.44
45-49	4	7	47	188	-16.8	-67.2	67.2	1128.96
40-44	3	3	42	126	-21.8	-65.4	65.4	1425.72
Total	N=70			4465			615.8	7901.8

Mean (M) = $\sum fX/N = 4465/70 = 63.8$

A.D. = $\sum|fx|/N = 615.8/70 = 8.80$

S.D. = $\sqrt{\{(\sum fx^2)/N\}} = \sqrt{(7901.8/70)} = 10.62$

$Q_1 = l + i \times \{(N/4 - \text{cum } f) \div fq\}$

$= 54.5 + 5 \times \{(17.5-13)/12\}$

$= 54.5 + 1.875 = 56.375$

$Q_3 = l + i \times \{(3N/4 - \text{cum } f) \div fq\}$

$= 69.5 + 5 \times \{(52.5-48)/9\}$

$= 69.5 + 2.5 = 72$

$Q = (Q_3-Q_1)/2$

$= (72-56.375)/2 = 7.81$

Sum 8: Calculate Q, A.D. and S.D. from the following frequency distribution

Class intervals	Frequency (f)	Cumulative frequency	Midpoint (X)	fX	x=X-M	fx	|fx|	fx^2
91-100	6	80	95.5	573	37.6	225.6	225.6	8482.56
81-90	8	74	85.5	684	27.6	220.8	220.8	6094.08
71-80	10	66	75.5	755	17.6	176	176	3097.6
61-70	11	56	65.5	720.5	7.6	83.6	83.6	635.36
51-60	15	45	55.5	832.5	-2.4	-36	36	86.4
41-50	12	30	45.5	546	-12.4	-148.8	148.8	1845.12
31-40	9	18	35.5	319.5	-22.4	-201.6	201.6	4515.84
21-30	6	9	25.5	153	-32.4	-194.4	194.4	6298.56
11-20	3	3	15.5	46.5	-42.4	-127.2	127.2	5393.28
Total	N=80			4630			1414	36448.8

Mean = $\sum fX/N = 4630/80 = 57.9$

A.D. = $\sum|fx|/N = 1414/80 = 17.675$

S.D. = $\sqrt{\{(\sum fx^2)/N\}} = \sqrt{(36448.8/80)} = 21.345$

$Q_1 = l + i \times \{(N/4 – \text{cum } f) \div fq\}$

$= 40.5 + 10 \times \{(20-18)/12\}$

$= 40.5 + (0.167\times10) = 42.17$

$Q_3 = l + i \times \{(3N/4 – \text{cum } f) \div fq\}$

$= 70.5 + 10 \times \{(60-56)/10\}$

$= 70.5 + 4 = 74.5$

$Q = (Q_3-Q_1)/2$

$= (74.5-42.17)/2 = 16.165$

Sum 9: Calculate S.D. from the following frequency distribution using the short method

Scores	Midpoint (X)	f	x’	fx’	fx'^{2}
145-149	147	4	4	16	64
140-144	142	7	3	21	63
135-139	137	9	2	18	36
130-134	132	10	1	10	10
125-129	127	15	0	0	0
120-124	122	11	-1	-11	11
115-119	117	7	-2	-14	28
110-114	112	5	-3	-15	45
105-109	107	2	-4	-8	32
		N=70		17	289

A.M. = 127

c = 17/70 = 0.24

$c^2 = (0.24)^2 = 0.0576$

Mean = A.M + ci

= 127 + (0.24×5) = 128.2

S.D. = $i \times \sqrt{[\{(\sum fx'^{2})/N\} - c^2]}$

= $5\times \sqrt{\{(289/70) - 0.0576\}}$

= 5 × 2.018 = 10.09

Advantages of Measures of Variability

Range

- Range is the simplest measure of variability which is computed by subtracting the lower extreme score from the upper extreme score. For instance in a class test, the highest score obtained by a pupil is 80 out of 100 and the lowest score obtained by a pupil is 50 out of 100. So the range is computed by subtracting 50 from 80, i.e., 30.
- It is computed when the distribution is small and a rough and quick measure of variability is required.

Quartile Deviation

- Quartile Deviation is computed when the distribution is skewed, i.e., when there are extreme scores in either or both ends.
- It is computed when the median is computed as the measure of central tendency.
- When a measure emphasizing on the middle 50 per cent of the cases is required and would suffice, quartile deviation may be used.

Average Deviation

- Average Deviation may be computed when the distribution of scores is normal or near normal.
- It takes into account the deviations of all the values in a given set of data from the mean.
- It is a less stable measure of variability as compared with the standard deviation.
- It indicates how much the group as a whole differs from the sample mean.

Standard Deviation

- Standard Deviation is computed when the distribution is normal or near normal.
- It is the most reliable and stable measure of variability.
- When the mean is computed as the measure of central tendency, standard deviation may be used as a measure of variability.
- It is computed when other statistical measures such as coefficient of correlation and significance of difference between the means are to be computed.
- Application of standard deviation gives us the knowledge about the scores present between the highest and lowest value.

Exercise

1. What do you understand by measures of variability? What are the different procedures by which variability can be measured?
2. Discuss the needs for computation of measures of variability in the social sciences.
3. What is quartile deviation?
4. What do you understand by the term 'standard deviation'? What are its applications?
5. Calculate the Quartile Deviation for the frequency distributions given below:

i.

Scores	Frequency
35-44	3
45-54	7
55-64	13
65-74	18
75-84	26
85-94	16
95-104	11
105-114	4
115-124	2

ii.

Scores	Frequency
20-24	2
25-29	4
30-34	6
35-39	9
40-44	11
45-49	13

50-54	10
55-59	8
60-64	4
65-69	2
70-74	1

6. Find out the Average Deviation and Standard Deviation from the following set of scores:
   9, 13, 11, 6, 17, 8
7. The scores obtained by 10 students in a Spelling test are as follows:
   38, 47, 26, 29, 41, 35, 33, 28, 40, 31. Calculate the A.D. and S.D.
8. Compute the Standard Deviation from the following frequency distributions:
   i.

Scores	Frequency
25-26	2
27-28	4
29-30	5
31-32	7
33-34	9
35-36	8
37-38	6
39-40	4
41-42	3
43-44	1

ii.

Scores	Frequency
40-44	5
45-49	8
50-54	9
55-59	13
60-64	15
65-69	12
70-74	10
75-79	6
80-84	3

Chapter 5

CORRELATION

Correlation is used to find out the relationship between two variables. For instance, the researcher may be interested to know whether there exists any relationship between intelligence and academic achievement of individuals. Likewise, he may also be interested to find out whether children who score high in Mathematics score higher in Physics as well. In other words, whether there exists any relationship between the scores in Mathematics and Physics. To solve such problems, correlation may be used.

Correlation may be of several types such as linear, curvilinear, partial or multiple. Linear correlation is the simplest type of correlation found between two variables in which the relationship between the two sets of scores of the two variables can be represented graphically by a straight line. It shows how a change in one variable is accompanied by a change in the other. Correlation may be positive, negative or zero. The correlation between two variables is said to be positive when an increase or decrease in the scores of one variable accompanies an increase or decrease in the other set of scores as well. The correlation is said to be negative when an increase in the scores of one variable accompanies a decrease in the scores of the other variable and vice versa. That, is an inverse relationship exists between the variables. In case of zero correlation, there exists no relationship between the two sets of scores. In order to express the relationship quantitatively between two variables, an index which includes the magnitude (numerical value) as well as direction (positive or negative) is used which is known as the coefficient of correlation. It is to be noted that correlation may not indicate any cause and effect relationship between the variables. It merely indicates an association between their changes without inferring whether or not the change in one has caused the change in the other. In addition, correlation cannot be used to predict the value of one variable from that of another.

In order to express the degree of relationship quantitatively between two variables, an index known as coefficient of correlation is used. It is the ratio which expresses the extent to which change in one variable is accompanied by change in the other. It does not involve any unit and varies from -1 to +1.

The value of correlation ranges from -1 to +1 where -1 indicates perfect negative correlation, +1 indicates perfect positive correlation and 0 indicates no relationship between two variables.

A descriptive label of the coefficient of correlation is given below:

r value	Descriptive label
± 0.00 – 0.20	Indifferent or negligible relationship
± 0.20 – 0.40	Low correlation; present but slight

± 0.40 – 0.70	Substantial or marked relationship
± 0.70 – 1.00	High to very high relationship

USES OF CORRELATION ANALYSIS:

After going through the above introduction about correlation, we can summarize the following uses or functions;

- The main purpose of computing correlation is to observe whether there is a significant relationship present between the variables.
- It helps in determining the direction of the relationship - that is positive, negative or zero.
- The magnitude or strength of the relationship.

The calculation of the coefficient of correlation using the Product Moment Method:

1) PRODUCT MOMENT CORRELATION (PEARSON r)

The product moment correlation coefficient, also known as Pearson r, was formulated by the British mathematician Karl Pearson. It is a measure of the magnitude and direction of the relation between two variables in a sample when their relationship can be described by a straight line. It can thus be called as a coefficient of simple linear correlation.

Product Moment coefficient of correlation can be computed when the following assumptions are met:

a. The relationship between the two variables must be linear, i.e., described by a straight line.
b. The standard deviation of the scores in the two variables should be equal and fairly homogeneous. This property is known as homoscedaticity.
c. The two variables should be continuous measurement variables.
d. The distributions should be unimodal and fairly symmetrical.

The standard formula used for computing Pearson's product moment correlation coefficient (r) is:

$$r = \frac{\sum xy}{\surd\,(\sum x2.\sum y2)}$$

where r = Correlation between the variables X and Y

x = Deviation of any X-score from the mean in variable X y = Deviation of the corresponding Y-score from the mean in variable Y

∑xy = Sum of all the products of deviation (each x deviation multiplied by its corresponding y deviation)

σ_x = Standard Deviation of the distribution of scores in the variable X

σ_y = Standard Deviation of the distribution of scores in the variable Y

N = Total number of cases or scores

Pearson's Product Moment Correlation Coefficient(r) can also be calculated directly from the raw scores, that is, without calculating deviation from the means. For this purpose, the following formula is used:

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

where X = Raw scores in variable X

Y = Raw scores in variable Y

∑XY = Sum of the products of each X score multiplied with its corresponding Y score

N = Total number of cases or scores

Examples:

Sum 1: Find the correlation coefficient between the two sets of scores given below.

Individuals	Math Test (X)	History Test (Y)	x ($X-M_x$)	y ($Y-M_y$)	x^2	y^2	xy
1	45	41	7.5	8.4	56.25	70.56	63
2	38	32	0.5	-0.6	0.25	0.36	-0.3
3	28	27	-9.5	-5.6	90.25	31.36	53.2
4	42	39	4.5	6.4	20.25	40.96	28.8
5	35	28	-2.5	-4.6	6.25	21.16	11.5
6	39	31	1.5	-1.6	2.25	2.56	-2.4

7	40	34	2.5	1.4	6.25	1.96	3.5
8	33	29	-4.5	-3.6	20.25	12.96	16.2
Total	300	261			202	181.88	173.5

Mean of Math scores (M_x) = 300/8 = 37.5

Mean of History scores (M_y) = 261/8 = 32.6

$$r = \frac{\sum xy}{\sqrt{(\sum x2.\sum y2)}}$$

= 173.5/√ (202×181.88)

= 173.5/191.676

= 0.9 (High positive correlation)

That is, X and Y are highly positive correlated.

Sum 2: Find the correlation between the two sets of scores given below.

Individuals	Spelling Test (X)	Comprehension Test (Y)	x ($X-M_x$)	y ($Y-M_y$)	x^2	y^2	xy
1	62	58	-3.3	-5.2	10.89	27.04	17.16
2	73	67	7.7	3.8	59.29	14.44	29.26
3	78	70	12.7	6.8	161.29	46.24	86.36
4	57	63	-8.3	-0.2	68.89	0.04	1.66
5	65	65	-0.3	1.8	0.09	3.24	-0.54
6	58	57	-7.3	-6.2	53.29	38.44	45.26
7	66	64	0.7	0.8	0.49	0.64	0.56
8	64	60	-1.3	-3.2	1.69	10.24	4.16
9	70	67	4.7	3.8	22.09	14.44	17.86

10	60	61	-5.3	-2.2	28.09	4.84	11.66
Total	653	632			406.1	159.6	213.4

Mean of spelling test (M_x) = 653/10 = 65.3

Mean of comprehension test (M_y) = 632/10 = 63.2

$$r = \frac{\sum xy}{\surd\,(\sum x2.\sum y2)}$$

= 213.4/√ (406.1×159.6)

= 213.4/254.585

= 0.84 (High positive correlation)

Sum 3: Compute the correlation between the two sets of scores given below.

Individuals	Verbal Test (X)	Arithmetic Test (Y)	X^2	Y^2	XY
1	42	28	1764	784	1176
2	34	40	1156	1600	1360
3	29	38	841	1444	1102
4	41	31	1681	961	1271
5	36	30	1296	900	1080
6	39	27	1521	729	1053
7	28	37	784	1369	1036
8	31	39	961	1521	1209
9	40	29	1600	841	1160
10	33	36	1089	1296	1188
Total	353	335	12,693	11,445	11,635

$$r = \frac{N.\sum XY - \sum X.\sum Y}{\surd\,[N.\sum X^2 - (\sum X)^2]\,[N.\sum Y^2 - (\sum Y)^2]}$$

= (10×11,635 - 353×335) / √ [10×12,693 – $(353)^2$] [10×11,445 – $(335)^2$]

= (1,16,350 – 1,18,255) / √ [(1,26,930-1,24,609)×(1,14,450-1,12,225]

= - 1905 / √ (2321×2225)

= - 1905/2272.493

= -0.84 (High negative correlation)

Sum 4: Compute the correlation between the two sets of scores given below.

Individuals	Test 1 (X)	Test 2 (Y)	X^2	Y^2	XY
1	58	70	3364	4900	4060
2	72	69	5184	4761	4968
3	63	65	3969	4225	4095
4	62	68	3844	4624	4216
5	71	66	5041	4356	4686
6	64	67	4096	4489	4288
7	59	62	3481	3844	3658
8	68	71	4624	5041	4828
9	61	58	3721	3364	3538
10	60	60	3600	3600	3600
Total	638	656	40,924	43,204	41,937

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

= (10×41,937 - 638×656) / √ [10×40,924 – (638) ^{2}] [10×43,204 – $(656)^2$]

= (4,19,370 - 4,18,528) / √ [(4,09,240-4,07,044)×(4,32,040-4,30,336)]

= 842 / √ (2196×1704)

= 842/1934.42

= 0.435 (Moderate positive correlation)

Sum 5: Compute the coefficient of correlation for the following continuous scores assuming normality.

Performance Test Scores (X)	Clerical Test Scores (Y)	x (X-M_x)	y (Y-M_y)	x^2	y^2	xy
72	55	0	-2.7	0	7.29	0
80	61	8	3.3	64	10.89	26.4
64	47	-8	-10.7	64	114.49	85.6
68	50	-4	-7.7	16	59.29	30.8
73	60	1	2.3	1	5.29	2.3
85	68	13	10.3	169	106.09	133.9
76	62	4	4.3	16	18.49	17.2
65	54	-7	-3.7	49	13.69	25.9
70	59	-2	1.3	4	1.69	-2.6
67	61	-5	3.3	25	10.89	-16.5
720	577			408	348.1	303

Mean of performance test scores (M_x) = 720/10 = 72

Mean of clerical test scores (M_y) = 577/10 = 57.7

$$r = \frac{\sum xy}{\surd\,(\sum x2.\sum y2)}$$

= 303/ √ (408×348.1)

= 303/376.86

= 0.804 (High positive correlation)

Sum 6: Compute the coefficient of correlation for the following continuous scores assuming normality.

Logical reasoning scores (X)	Clerical test scores (Y)	x= $(X\text{-}M_x)$	y = $(Y\text{-}M_y)$	x^2	y^2	xy
48	30	8.5	-2.8	72.25	7.84	-23.8
42	27	2.5	-5.8	6.25	33.64	-14.5
31	39	-8.5	6.2	72.25	38.44	-52.7
35	40	-4.5	7.2	20.25	51.84	-32.4
46	29	6.5	-3.8	42.25	14.44	-24.7
44	28	4.5	-4.8	20.25	23.04	-21.6
38	33	-1.5	0.2	2.25	0.04	-0.3
32	36	-7.5	3.2	56.25	10.24	-24
316	262			292	179.52	-194

Mean of logical reasoning scores (M_x) = 316/8 = 39.5

Mean of clerical test scores (M_y) = 262/8 = 32.8

$$r = \frac{\sum xy}{\surd\,(\sum x2.\sum y2)}$$

= -194 / √ (292×179.52)

= -194/228.95

= -0.85 (High negative correlation)

Sum 7: Compute the correlation between the two sets of scores given below.

Test 1 (X)	Test 2 (Y)	x= $(X-M_x)$	y = $(Y-M_y)$	x^2	y^2	xy
50	45	10.1	4.5	102.01	20.25	45.45
33	36	-6.9	-4.5	47.61	20.25	31.05
42	44	2.1	3.5	4.41	12.25	7.35
35	31	-4.9	-9.5	24.01	90.25	46.55
46	50	6.1	9.5	37.21	90.25	57.95
39	37	-0.9	-3.5	0.81	12.25	3.15
34	38	-5.9	-2.5	34.81	6.25	14.75
40	43	0.1	2.5	0.01	6.25	0.25
319	324			250.88	258	206.5

Mean of test 1 scores (M_x) = 319/8 = 39.9

Mean of test 2 scores (M_y) = 324/8 = 40.5

$$r = \frac{\sum xy}{\surd\,(\sum x2.\sum y2)}$$

= 206.5 / √ (250.88×258)

= 206.5/254.4

= 0.81 (High positive correlation)

Sum 8: Compute the coefficient of correlation between the two sets of scores given below.

Verbal reasoning scores (X)	Numerical reasoning scores (Y)	x = $(X-M_x)$	y = $(Y-M_y)$	x^2	y^2	xy
38	46	3.4	11.7	11.56	136.89	39.78
42	30	7.4	-4.3	54.76	18.49	-31.82

34	37	-0.6	2.7	0.36	7.29	-1.62
39	31	4.4	-3.3	19.36	10.89	-14.52
27	39	-7.6	4.7	57.76	22.09	-35.72
40	29	5.4	-5.3	29.16	28.09	-28.62
36	36	1.4	1.7	1.96	2.89	2.38
33	28	-1.6	-6.3	2.56	39.69	-10.08
26	35	-8.6	0.7	73.96	0.49	-6.02
31	32	-3.6	-2.3	12.96	5.29	8.28
346	343			264.4	272.1	-77.96

Mean of verbal reasoning scores (M_x) = 346/10 = 34.6

Mean of numerical reasoning scores (M_y) = 343/10 = 34.3

$$r = \frac{\sum xy}{\sqrt{(\sum x2 . \sum y2)}}$$

= -77.96 /√ (264.4×272.1)

= -77.96/268.22

= -0.29 (Low negative correlation)

Other Methods of Correlation:

2) SPEARMAN'S RANK CORRELATION (ρ-rho)

In distributions where it is more convenient to express the scores in terms of ranks, Spearman's Rank Difference Correlation may be used. For instance, if the scores obtained by 5 students are 37, 43, 39, 35 and 41 then their ranks would be 4, 1, 3, 5 and 2 respectively. However, it is to be remembered that in certain cases, say, for example when measures of reaction time of individuals are given then the individual with the least reaction time should be ranked 1 and so on. The rank correlation coefficient may be computed using the following formula:

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

where, ρ = Spearman's rank correlation coefficient

$\sum D^2$ = Summation of the squares of the rank difference values

n = Total number of scores

Sum 1: Compute the correlation between the following two series of test scores by the rank difference method.

Individuals	Score on X	Score on Y	Rank 1 (R_1)	Rank 2 (R_2)	$D=R_1-R_2$	D^2
1	15	19	4	2	2	4
2	9	14	9	6	3	9
3	17	10	2	9	-7	49
4	13	12	6	8	-2	4
5	11	20	8	1	7	49
6	14	8	5	10	-5	25
7	8	13	10	7	3	9
8	12	16	7	4	3	9
9	18	17	1	3	-2	4
10	16	15	3	5	-2	4
Total						166

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

= 1 - [(6×166) / (10×99)]

= 1- (996/990)

= 1-1.006

= -0.006 (Very low negative correlation, that is, almost negligible relationship, i.e. zero correlation)

Sum 2: Compute the correlation between the following two series of test scores by the rank difference method.

Students	Marks in History	Marks in Civics	R_1	R_2	$D=R_1-R_2$	D^2
1	78	82	2	2	0	0
2	46	57	10	10	0	0
3	83	85	1	1	0	0
4	65	69	7	6	1	1
5	72	70	3	5	-2	4
6	59	63	8	8	0	0
7	55	61	9	9	0	0
8	71	68	4	7	-3	9
9	66	71	6	4	2	4
10	67	73	5	3	2	4
Total						21

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

= 1- [(6×21) / (10×99)]

= 1- (126/990)

= 1-0.13

= 0.87 (High positive correlation)

Sum 3: Compute the correlation between the following two series of test scores by the rank difference method.

Students	History scores	Math scores	R_1	R_2	$D=R_1-R_2$	D^2
1	82	64	1	10	-9	81
2	79	70	3	7.5	-4.5	20.25
3	54	86	10	1	9	81
4	63	72	8	6	2	4
5	70	70	5	7.5	-2.5	6.25
6	68	77	6	5	1	1
7	59	81	9	3	6	36
8	78	66	4	9	-5	25
9	80	78	2	4	-2	4
10	65	83	7	2	5	25
Total						283.5

Tied rank in math scores = (7+8)/2 = 15/2 = 7.5 each for the score of 70

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

= 1- [(6×283.5) / (10×99)]

= 1- (1701/990)

= 1- 1.72

= -0.72 (High negative correlation)

Sum 4: Compute the correlation between the following two series of test scores by the rank difference method.

Individuals	Abstract reasoning scores	Mechanical reasoning scores	R_1	R_2	$D=R_1-R_2$	D^2
1	43	32	1	4.5	-3.5	12.25
2	39	28	4	7	-3	9
3	41	25	2.5	9	-6.5	42.25
4	29	37	9	1	8	64
5	23	36	10	2	8	64
6	41	28	2.5	7	-4.5	20.25
7	34	28	7	7	0	0
8	35	22	6	10	-4	16
9	31	32	8	4.5	3.5	12.25
10	36	33	5	3	2	4
Total						244

Tied rank in abstract reasoning= (2+3)/2 = 2.5 each for the score of 41

Tied rank in mechanical reasoning= (4+5)/2 = 4.5 each for the score of 32

(6+7+8)/3 = 7 each for the score of 28

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

= 1- [(6×244) / (10×99)]

= 1- 1.48

= -0.48 (Moderate negative correlation)

Sum 5: Compute the correlation between the following two series of test scores by the rank difference method.

Individuals	Intelligence test scores	Cancellation score	R_1	R_2	$D=R_1-R_2$	D^2
1	119	98	7	8.5	-1.5	2.25
2	131	87	3	2	1	1
3	129	99	4	10	-6	36
4	136	91	1	5	-4	16
5	118	85	8	1	7	49
6	112	89	9	3	6	36
7	133	93	2	6	-4	16
8	126	98	5	8.5	-3.5	12.25
9	120	95	6	7	-1	1
10	110	90	10	4	6	36
Total						205.5

[The cancellation scores are in seconds; hence the smallest score numerically, i.e. 85, is highest and is ranked 1.]

Tied rank in cancellation score = (8+9)/2 =8.5 each for the score of 98

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

$= 1- [(6\times205.5) / (10\times99)]$

$= 1- (1233/990)$

$= 1-1.25$

$= -0.25$ (Low negative correlation)

Sum 6: Compute the correlation between the following two series of test scores by the rank difference method.

Individuals	Digit span	Letter span	R_1	R_2	$D=R_1-R_2$	D^2
1	8	7	11	11.5	-0.5	0.25
2	11	13	6	1	5	25
3	9	9	9	8	1	1
4	7	8	12	9.5	2.5	6.25
5	10	11	7	4.5	2.5	6.25
6	13	12	2.5	2.5	0	0
7	9	8	9	9.5	-0.5	0.25
8	12	10	4.5	6.5	-2	4
9	9	7	9	11.5	-2.5	6.25
10	13	11	2.5	4.5	-2	4
11	14	10	1	6.5	-5.5	30.25
12	12	12	4.5	2.5	2	4
Total						87.5

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

= 1- [(6×87.5) / (12×143)]

= 1- (525/1716)

= 1-0.31

= 0.69 (High positive correlation)

Sum 7: Compute the correlation between the following two series of test scores by the rank difference method.

Individuals	Scores in verbal comprehension	Scores in reasoning	R_1	R_2	$D=R_1-R_2$	D^2
1	41	37	1	6.5	-5.5	30.25
2	33	35	8.5	9	-0.5	0.25
3	39	40	3	3.5	-0.5	0.25
4	26	30	10	11	-1	1
5	33	39	8.5	5	3.5	12.25
6	36	37	6	6.5	-0.5	0.25
7	25	28	11.5	12	-0.5	0.25
8	38	41	4	2	2	4
9	25	31	11.5	10	1.5	2.25
10	34	36	7	8	-1	1
11	37	42	5	1	4	16
12	40	40	2	3.5	-1.5	2.25
Total						70

$$\rho \quad 1 - \frac{6\sum D^2}{n\,(n^2 - 1)}$$

= 1- [(6×70) / (12×143)]

= 1- (420/1716)

= 0.76 (High positive correlation)

Sum 8: Compute the correlation between the following two series of test scores by the rank difference method.

Individuals	Test 1	Test 2	R_1	R_2	$D=R_1-R_2$	D^2
1	86	80	2	2	0	0
2	77	71	4	4	0	0
3	52	51	14.5	14.5	0	0
4	68	57	8.5	11	-2.5	6.25
5	54	53	13	13	0	0
6	65	59	11	9.5	1.5	2.25
7	52	51	14.5	14.5	0	0
8	83	72	3	3	0	0
9	90	91	1	1	0	0
10	71	62	7	6	1	1
11	75	65	5	5	0	0
12	68	59	8.5	9.5	-1	1
13	66	61	10	7	3	9
14	61	56	12	12	0	0
15	73	60	6	8	-2	4
Total						23.5

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

= 1- [(6×23.5) / (15×224)]

= 1- (141/3360)

= 0.96 (High positive correlation)

3) BISERIAL CORRELATION

In social sciences, we often come across in regards to correlated variables in which one is artificially reduced to dichotomy. The term dichotomy means “split into”. The variables which we split into halves as per our convenience, such as successful-unsuccessful, rich-poor, moral-immoral, etc. are referred to as artificially dichotomized variables. When one of the variables is a continuous measurement variable and the other is artificially dichotomized, biserial correlation seems more appropriate.

Sum 1: The following data shows the scores obtained by individuals in a neuroticism-screening test and they are classified into two groups- neurotic and non-neurotic. Compute the coefficient of biserial correlation.

Scores	Neurotic (p)	Non-neurotic (q)	Total f
70-79	18	4	22
60-69	13	5	18
50-59	9	8	17
40-49	11	2	13
30-39	6	6	12
20-29	5	5	10
Total	N_1=62	N_2=30	N=92

Scores	Midpoint (X)	p	pX
70-79	74.5	18	1341
60-69	64.5	13	838.5
50-59	54.5	9	490.5
40-49	44.5	11	489.5
30-39	34.5	6	207
20-29	24.5	5	122.5
Total		N_1=62	3489

Mean (M_p) = $\sum pX/N_1$ = 3489/62 = 56.27

Scores	Midpoint (X)	q	qX
70-79	74.5	4	298
60-69	64.5	5	322.5
50-59	54.5	8	436
40-49	44.5	2	89
30-39	34.5	6	207
20-29	24.5	5	122.5
Total		N_2=30	1475

Mean (M_q) = $\sum qX/N_2$ = 1475/30 = 49.17

Scores	Midpoint (X)	Total f	fX	x=X-M	fx	fx^2
70-79	74.5	22	1639	20.54	451.88	9281.62
60-69	64.5	18	1161	10.54	189.72	1999.65
50-59	54.5	17	926.5	0.54	9.18	4.96
40-49	44.5	13	578.5	-9.46	-122.98	1163.40
30-39	34.5	12	414	-19.46	-233.52	4544.30
20-29	24.5	10	245	-29.46	-294.6	8678.92
Total		N=92	4964			25,672.85

Mean (M) = $\sum fX/N$ = 4964/92 = 53.96

Standard Deviation (S.D.) = $\sqrt{(\sum fx^2/N)}$ = $\sqrt{(25,672.85/92)}$ = $\sqrt{279.053}$ = 16.7

P= N_1/N = 62/92 =0.67

Q= 1-P = 1-0.67 = 0.33

From the table given at the end of the chapter, y=0.362

r bis = [(Mp-Mq)/S.D.] × (PQ/y)

= [(56.27-49.17)/16.7] × [(0.67×0.33)/0.362]

= 0.425 × 0.611

= 0.26 (Low positive correlation)

Sum 2: The table below shows the distribution of scores on a music aptitude test by individuals who are trained in music and those who are untrained. Compute r bis.

Scores	Trained (p)	Untrained (q)	Total f
90-99	9	4	13
80-89	13	7	20
70-79	12	8	20
60-69	34	14	48
50-59	22	6	28
40-49	14	5	19
30-39	11	2	13
Total	N_1=115	N_2=46	N=161

Scores	Trained (p)	x’	px’
90-99	9	3	27
80-89	13	2	26
70-79	12	1	12
60-69	34	0	0
50-59	22	-1	-22
40-49	14	-2	-28

30-39	11	-3	-33
Total	N_1=115		-18

Assumed Mean (A.M.) = 64.5

Correction (c) = -18/115 =-0.16

Mean (Mp) = A.M. + ci [i=size of class interval]

= 64.5+ (-0.16×10) = 64.5-1.6 = 62.9

Scores	Untrained (q)	x’	qx’
90-99	4	3	12
80-89	7	2	14
70-79	8	1	8
60-69	14	0	0
50-59	6	-1	-6
40-49	5	-2	-10
30-39	2	-3	-6
Total	N_2=46		12

Assumed Mean (A.M.) = 64.5

Correction (c) = 12/46 = 0.26

Mean (Mq) = A.M. + ci

= 64.5+ (0.26×10) = 64.5+2.6 = 67.1

Scores	Midpoint (X)	Total f	fX	x=X-M	fx	fx^2
90-99	94.5	13	1228.5	30.37	394.81	11,990.38
80-89	84.5	20	1690	20.37	407.4	8298.74
70-79	74.5	20	1490	10.37	207.4	2150.74

60-69	64.5	48	3096	0.37	17.76	6.57
50-59	54.5	28	1526	-9.63	-269.64	2596.63
40-49	44.5	19	845.5	-19.63	-372.97	7321.4
30-39	34.5	13	448.5	-29.63	-385.19	11,413.18
Total		N=161	10,324.5			43,777.64

Mean (M) = $\sum fX/N$ = 10,324.5/161 = 64.13

S.D. = $\sqrt{(\sum fx^2/N)} = \sqrt{(43{,}777.64/161)}$ = 16.49

$P=N_1/N$ = 115/161 = 0.71

Q=1-P = 1-0.71 = 0.29

From table given at the end of the chapter, y=0.342

r bis = [(Mp-Mq)/S.D.] × (PQ/y)

= [(62.9-67.1)/16.49] × [(0.71×0.29)/0.342]

= -0.255 × 0.602

= -0.15 (Low negative correlation)

Sum 3: The following data gives the distribution of marks obtained in mathematics in the CBSE examination for those who scored more than 50% and those who scored 50% or less in the selection test. Compute the correlation between the marks obtained in mathematics in CBSE examination and performance in selection test.

Mathematics scores	Those who scored more than 50% in selection test (p)	Those who scored 50% and less in selection test (q)	Total f
95-99	3	0	3
90-94	10	2	12
85-89	12	4	16
80-84	21	6	27
75-79	13	5	18

70-74	6	5	11
65-69	5	3	8
Total	N_1=70	N_2=25	N=95

Scores	p	x’	px’
95-99	3	3	9
90-94	10	2	20
85-89	12	1	12
80-84	21	0	0
75-79	13	-1	-13
70-74	6	-2	-12
65-69	5	-3	-15
Total	N_1=70		1

Assumed Mean (A.M.) = 82

Correction (c) = 1/70 = 0.014

Mean (Mp) = A.M. + ci

= 82 + (0.014×5) = 82 + 0.07 = 82.07

Scores	q	x’	qx’
95-99	0	3	0
90-94	2	2	4
85-89	4	1	4
80-84	6	0	0
75-79	5	-1	-5
70-74	5	-2	-10

65-69	3	-3	-9
Total	N_2=25		-16

Assumed Mean (A.M.) = 82

Correction (c) = -16/25 = -0.64

Mean (Mq)= A.M. + ci

= 82 + (-0.64×5) = 82-3.2 = 78.8

Scores	Midpoint (X)	Total f	fX	x=X-M	fx	fx^2
95-99	97	3	291	15.79	47.37	747.97
90-94	92	12	1104	10.79	129.48	1397.09
85-89	87	16	1392	5.79	92.64	536.39
80-84	82	27	2214	0.79	21.33	16.85
75-79	77	18	1386	-4.21	-75.78	319.03
70-74	72	11	792	-9.21	-101.31	933.07
65-69	67	8	536	-14.21	-113.68	1615.39
Total		N=95	7715			5565.79

Mean (M) = ∑fX/N = 7715/95 = 81.21

S.D. = √ ($\sum fx^2$/N) = √ (5565.79/95) = 7.65

P = N_1/N = 70/95 = 0.74

Q = 1-P = 1-0.0.74 = 0.26

From table given at the end of the chapter, y = 0.324

r bis = [(Mp-Mq)/S.D.] × (PQ/y)

= [(82.07-78.8)/7.65] × [(0.74×0.26)/0.324]

= 0.427 × 0.594 = 0.25 (Low positive correlation)

Sum 4: The table below shows the distribution of scores on an achievement test earned by students who answered more than 50% and those who answered 50% or less of the items in a numerical reasoning test correctly. Compute r bis.

Achievement test scores	Students answering more than 50% of the items on numerical reasoning test correctly (p)	Students answering 50% or less of the items on numerical reasoning test correctly (q)	Total f
95-99	7	4	11
90-94	12	8	20
85-89	21	18	39
80-84	36	31	67
75-79	19	9	28
70-74	10	7	17
65-69	5	3	8
Total	N_1=110	N_2=80	N=190

Scores	p	x’	px’
95-99	7	3	21
90-94	12	2	24
85-89	21	1	21
80-84	36	0	0
75-79	19	-1	-19
70-74	10	-2	-20
65-69	5	-3	-15
Total	N_1=110		12

Assumed Mean (A.M.) = 82

Correction (c) = 12/110 = 0.11

Mean (Mp) = A.M. + ci

= 82 + (0.11×5) = 82.55

Scores	q	x’	qx’
95-99	4	3	12
90-94	8	2	16
85-89	18	1	18
80-84	31	0	0
75-79	9	-1	-9
70-74	7	-2	-14
65-69	3	-3	-9
Total	N_2=80		14

Assumed Mean (A.M.) = 82

Correction (c) = 14/80 = 0.175

Mean (Mq) = A.M. + ci

= 82 + (0.175×5) = 82.88

Scores	Midpoint (X)	Total f	fX	x=X-M	fx	fx^2
95-99	97	11	1067	14.32	157.52	2255.69
90-94	92	20	1840	9.32	186.4	1737.25
85-89	87	39	3393	4.32	168.48	727.83
80-84	82	67	5494	-0.68	-45.56	30.98
75-79	77	28	2156	-5.68	-159.04	903.35
70-74	72	17	1224	-10.68	-181.56	1939.06
65-69	67	8	536	-15.68	-125.44	1966.90
Total		N=190	15,710			9561.06

Mean (M) = ∑fX/N = 15,710/190 = 82.68

S.D. = √ (∑fx^2/N)

= √ (9561.06/190) = 7.09

P= N$_1$/N = 110/190 = 0.58

Q= 1-P = 1-0.58 = 0.42

From table given at the end of the chapter, y=0.391

r bis = [(Mp-Mq)/S.D.] × (PQ/y)

= [(82.55-82.88)/7.09] × [(0.58×0.42)/0.391]

= -0.047 × 0.623 = -0.03 (Very low negative correlation; almost negligible relationship, i.e. zero correlation)

4) Point Biserial r

In contrast to artificial dichotomy, variables may be dichotomized naturally in several ways such as right-wrong, male-female, alive-dead, etc. In these categories, the division is clear that makes these naturally or genuinely dichotomized variables. Point biserial correlation seems more appropriate when one of the variables is a continuous measurement variable and there exists genuine dichotomy in the other variable.

Sum 1: Compute r_p bis from the following data.

Serum cholesterol values (mg/dL)	Males (p)	Females (q)	Total f
215-219	2	1	3
210-214	4	3	7
205-209	7	8	15
200-204	14	9	23
195-199	22	19	41

190-194	18	11	29
185-189	13	6	19
180-184	8	5	13
175-179	3	3	6
Total	N_1=91	N_2=65	N=156

Serum cholesterol values (mg/dL)	p	x’	px’
215-219	2	4	8
210-214	4	3	12
205-209	7	2	14
200-204	14	1	14
195-199	22	0	0
190-194	18	-1	-18
185-189	13	-2	-26
180-184	8	-3	-24
175-179	3	-4	12
Total	N_1=91		-32

Assumed Mean (A.M) = 197

Correction (c) = -32/91 = -0.35

Mean (Mp) = A.M. + ci

$$= 197 + (-0.35 \times 5) = 197-1.75 =195.25$$

Serum cholesterol values (mg/dL)	Females (q)	x'	qx'
215-219	1	4	4
210-214	3	3	9
205-209	8	2	16
200-204	9	1	9
195-199	19	0	0
190-194	11	-1	-11
185-189	6	-2	-12
180-184	5	-3	-15
175-179	3	-4	-12
Total	N_2=65		-12

Assumed Mean = 197

Correction (c) = -12/65 = -0.18

Mean (Mq) = A.M. + ci

= 197 + (-0.18 × 5) = 197-0.9 = 196.1

Serum cholesterol values (mg/dL)	Midpoint (X)	Total f	fX	x=X-M	fx	fx^2
215-219	217	3	651	21.4	64.2	1373.88
210-214	212	7	1484	16.4	114.8	1882.72
205-209	207	15	3105	11.4	171	1949.4
200-204	202	23	4646	6.4	147.2	942.08
195-199	197	41	8077	1.4	57.4	80.36
190-194	192	29	5568	-3.6	-104.4	375.84

185-189	187	19	3553	-8.6	-163.4	1405.24
180-184	182	13	2366	-13.6	-176.8	2404.48
175-179	177	6	1062	-18.6	-111.6	2075.76
Total		N=156	30,512			12,489.76

Mean (M) = 30,512/156 = 195.6

S.D. = $\sqrt{}$ ($\sum fx^2/N$) = $\sqrt{}$ (12,489.76/156) = 8.95

P= N_1/N = 91/156 = 0.58

Q= 1-P = 1-0.58 = 0.42

r_p bis = [(Mp-Mq)/S.D.] × $\sqrt{}$ (PQ)

= [(195.25-196.1)/8.95] × $\sqrt{}$ (0.58×0.42) = -0.095 × 0.49 = -0.05 (Very low negative correlation; almost negligible relationship, i.e. zero correlation)

Sum 2: Compute r_p bis from the following data.

Prolactin values (ng/mL)	Pregnant females (p)	Non-pregnant females (q)	Total f
81-90	7	0	7
71-80	11	0	11
61-70	19	3	22
51-60	21	5	26
41-50	36	12	48
31-40	23	6	29
21-30	15	5	20
11-20	9	7	19
1-10	5	3	8
Total	N_1=146	N_2=41	N=187

Prolactin values (ng/mL)	Pregnant females (p)	x’	px’
81-90	7	4	28
71-80	11	3	33
61-70	19	2	38
51-60	21	1	21
41-50	36	0	0
31-40	23	-1	-23
21-30	15	-2	-30
11-20	9	-3	-27
1-10	5	-4	-20
Total	N1=146		20

Assumed Mean (A.M.) = 45.5

Correction (c) = 20/146 = 0.14

Mean (Mp) = A.M. + ci

= 45.5 + (0.14×10) = 46.9

Prolactin values (ng/mL)	Non-pregnant females (q)	x’	qx’
81-90	0	4	0
71-80	0	3	0
61-70	3	2	6
51-60	5	1	5
41-50	12	0	0
31-40	6	-1	-6

21-30	5	-2	-10
11-20	7	-3	-21
1-10	3	-4	-12
Total	N_2=41		-38

Assumed Mean = 45.5

Correction (c) = -38/41 = -0.93

Mean (Mq) = A.M. + ci

= 45.5 + (-0.93×10) = 45.5 – 9.3 = 36.2

Prolactin values (ng/mL)	Midpoint (X)	f	fX	x=X-M	fx	fx^2
81-90	85.5	7	598.5	40.96	286.72	11744.05
71-80	75.5	11	830.5	30.96	340.56	10543.74
61-70	65.5	22	1441	20.96	461.12	9665.08
51-60	55.5	26	1443	10.96	284.96	3123.16
41-50	45.5	48	2184	0.96	46.08	44.24
31-40	35.5	29	1029.5	-9.04	-262.16	2369.93
21-30	25.5	20	510	-19.04	-380.8	7250.43
11-20	15.5	19	248	-29.04	-464.64	13493.15
1-10	5.5	8	44	-39.04	-312.32	12192.97
Total		N=187	8328.5			70,426.75

Mean (M) = 8328.5/187 = 44.54

S.D. = $\sqrt{}$ ($\sum fx^2$/N) = $\sqrt{}$ (70,426.75/187) = 19.41

P=N_1/N = 146/187 = 0.78

Q=1-P = 1-0.78 = 0.22

r_p bis = [(Mp-Mq)/S.D.] × √ (PQ)

= [(46.9-36.2)/19.41] × √ (0.78 × 0.22) = 0.55×0.414 = 0.23 (Low positive correlation)

Sum 3: Compute r_p bis from the following data.

Logical reasoning test scores	Those who responded rightly (p)	Those who responded wrongly (q)	Total f
75-79	5	1	6
70-74	8	3	11
65-69	9	4	13
60-64	8	6	14
55-59	9	5	14
50-54	10	7	17
45-49	9	11	20
40-44	6	10	16
35-39	5	12	17
30-34	2	7	9
25-29	1	13	14
Total	N_1=72	N_2=79	N=151

Logical reasoning test scores	Those who responded rightly (p)	x’	px’
75-79	5	5	25
70-74	8	4	32
65-69	9	3	27

60-64	8	2	16
55-59	9	1	9
50-54	10	0	0
45-49	9	-1	-9
40-44	6	-2	-12
35-39	5	-3	-15
30-34	2	-4	-8
25-29	1	-5	-5
Total	N_1=72		60

Assumed Mean = 52

Correction (c) = 60/72 = 0.83

Mean (Mp) = 52 + (0.83×5) = 56.15

Logical reasoning test scores	Those who responded wrongly (q)	x’	qx’
75-79	1	5	5
70-74	3	4	12
65-69	4	3	12
60-64	6	2	12
55-59	5	1	5
50-54	7	0	0
45-49	11	-1	-11
40-44	10	-2	-20

35-39	12	-3	-36
30-34	7	-4	-28
25-29	13	-5	-65
Total	N_2=79		-114

Assumed Mean (A.M.) = 52

Correction (c) = -114/79 = -1.44

Mean (Mq) = A.M + ci = 52 + (-1.44×5) = 44.8

Logical reasoning test scores	Midpoint (X)	f	fX	x=X-M	fx	fx^2
75-79	77	6	462	26.79	160.74	4306.22
70-74	72	11	792	21.79	239.69	5222.85
65-69	67	13	871	16.79	218.27	3664.75
60-64	62	14	868	11.79	165.06	1946.06
55-59	57	14	798	6.79	95.06	645.46
50-54	52	17	884	1.79	30.43	54.47
45-49	47	20	940	-3.21	-64.2	206.08
40-44	42	16	672	-8.21	-131.36	1078.47
35-39	37	17	629	-13.21	-224.57	2966.57
30-34	32	9	288	-18.21	-163.89	2984.44
25-29	27	14	378	-23.21	-324.94	7541.86
Total		N=151	7582			30,617.23

Mean = ∑fX/N = 7582/151 = 50.21

S.D. = √ ($\sum fx^2$/N) = √ (30,617.23/151) = 14.24

P=N_1/N = 72/151 = 0.48

Q=1-P = 1-0.48 = 0.52

r_p bis = [(Mp-Mq)/S.D.] × √ (PQ) = [(56.15-44.8)/14.24] × √ (0.48×0.52) = 0.797×0.4996 = 0.4 (Moderate positive correlation)

Sum 4: Compute r_p bis from the following data.

Memory test scores	Females (p)	Males (q)	Total f
70-74	8	5	13
65-69	7	7	14
60-64	10	6	16
55-59	11	11	22
50-54	12	11	23
45-49	9	7	16
40-44	5	3	8
35-39	2	3	5
30-34	0	2	2
Total	N_1=64	N_2=55	N=119

Memory test scores	Females (p)	x’	px’
70-74	8	4	32
65-69	7	3	21
60-64	10	2	20
55-59	11	1	11
50-54	12	0	0
45-49	9	-1	-9
40-44	5	-2	-10

35-39	2	-3	-6
30-34	0	-4	0
Total	N_1=64		59

Assumed Mean (A.M.) = 52

Correction (c) = 59/64 = 0.92

Mean (Mp) = A.M. + ci = 52 + (0.92×5) = 56.6

Memory test scores	Males (q)	x'	qx'
70-74	5	4	20
65-69	7	3	21
60-64	6	2	12
55-59	11	1	11
50-54	11	0	0
45-49	7	-1	-7
40-44	3	-2	-6
35-39	3	-3	-9
30-34	2	-4	-8
Total	N_2=55		34

Assumed Mean (A.M.) = 52

Correction (c) = 34/55 = 0.63

Mean (Mq) = A.M. + ci = 52 + (0.62×5) = 55.1

Memory test scores	Midpoint (X)	f	fX	x=X-M	Fx	fx^2
70-74	72	13	936	16.1	209.3	3369.73

65-69	67	14	938	11.1	155.4	1724.94
60-64	62	16	992	6.1	97.6	595.36
55-59	57	22	1254	1.1	24.2	26.62
50-54	52	23	1196	-3.9	-89.7	349.83
45-49	27	16	752	-8.9	-142.4	1267.36
40-44	42	8	336	-13.9	-111.2	1545.68
35-39	37	5	185	-18.9	-94.5	1786.05
30-34	32	2	64	-23.9	-47.8	1142.42
Total		N=119	6653			11,807.99

Mean (M) = $\sum fX/N = 6653/119 = 55.9$

S.D. = $\sqrt{(\sum fx^2/N)} = \sqrt{(11,807.99/119)} = 9.96$

$P=N_1/N = 64/119 = 0.54$

$Q=1-P = 1-0.54 = 0.46$

r_p bis = $[(Mp-Mq)/S.D.] \times \sqrt{(PQ)}$

$= [(56.6-55.1)/9.96] \times \sqrt{(0.54 \times 0.46)}$

$= 0.15 \times 0.50 = 0.075$ (Low positive correlation)

5) TETRACHORIC CORRELATION

Tetrachoric correlation (r_t) is used when both the variables are artificially reduced to dichotomies and none of them can be expressed in scores. For instance, if we want to study the relationship between emotional intelligence and adjustment then the variables may be dichotomized as: high emotional intelligence - low emotional intelligence and adjusted – maladjusted. However, tetrachoric correlation assumes that both the variables under study are continuous measurement variables and they would be normally distributed if scores could be obtained, thus classifying both the variables into frequency distributions.

Sum 1: Compute the tetrachoric correlation coefficient from the following table.

	Non-hypertensive	Hypertensive	Total
Neurotic	35 (B)	45 (A)	80
Normal	30 (D)	50 (C)	80
Total	65	95	160

$r_t = \cos\left[\frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})}\right]$

$= \cos\left[\frac{(180° \times \sqrt{35\times50})}{(\sqrt{45\times30} + \sqrt{35\times50})}\right]$

$= \cos\left[\frac{(180°\times41.83)}{(36.74+41.83)}\right]$

= cos (7529.4/78.57)

= cos 96° = -0.10 (Low negative correlation)

Sum 2: Compute the tetrachoric correlation coefficient from the following table

	Normal	Neurotic	Total
Adjusted	35 (B)	50 (A)	85
Maladjusted	45 (D)	25 (C)	70
Total	80	75	155

$r_t = \cos\left[\frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})}\right]$

$= \cos\left[\frac{(180° \times \sqrt{35\times25})}{(\sqrt{50\times45} + \sqrt{35\times25})}\right]$

$= \cos\left[\frac{(180°\times29.58)}{(47.43+29.58)}\right]$

= cos (5324.4/77.01)

= cos 69° = 0.36 (Moderate positive correlation)

Sum 3: 180 individuals (diabetic and non-diabetic) were tested on a test of anxiety and the results were tabulated as follows. Find the tetrachoric correlation coefficient between diabetes and anxiety.

	High anxiety level	Low anxiety level	Total
Diabetic	35 (A)	45 (B)	80
Non-diabetic	12 (C)	88 (D)	100
Total	47	133	180

$$r_t = \cos\left[\ \frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})}\ \right]$$

$$= \cos\left[\frac{(180° \times \sqrt{45{\times}12})}{(\sqrt{35{\times}88} + \sqrt{45{\times}12})}\right]$$

$$= \cos\left[\frac{(180° {\times} 23.24}{(55.50{+}23.24)}\right]$$

= cos (4183.2/78.74)

= cos 53° = 0.60 (High positive correlation)

Sum 4: Compute the tetrachoric correlation coefficient from the given data to find the relationship between emotional maturity and anxiety level

	High anxiety level	Low anxiety level	Total
Emotionally mature	45 (A)	55 (B)	100
Emotionally immature	75 (C)	35 (D)	110
Total	120	90	210

$$r_t = \cos\left[\ \frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})}\ \right]$$

$$= \cos\left[\frac{(180° \times \sqrt{55{\times}75})}{(\sqrt{45{\times}35} + \sqrt{55{\times}75}}\right]$$

$$= \cos\left[\frac{(180° {\times} 64.23)}{(39.69{+}64.23)}\right]$$

= cos (11561.4/103.92)

= cos 111° = -0.36 (Moderate negative correlation)

Sum 5: Compute the tetrachoric correlation coefficient from the given data to find the relationship between emotional maturity and adjustment

	Adjusted	Maladjusted	Total
Emotionally mature	55 (A)	40 (B)	95
Emotionally immature	20 (C)	70 (D)	90
Total	75	110	185

$r_t = \cos\left[\frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})}\right]$

$= \cos\left[\frac{(180° \times \sqrt{40{\times}20})}{(\sqrt{55{\times}70} + \sqrt{40{\times}20})}\right]$

$= \cos\left[\frac{(180° {\times} 28.28)}{(62.05{+}28.28)}\right]$

$= \cos(5090.4/90.33)$

$= \cos 56° = 0.56$ (High positive correlation)

Sum 6: Compute the tetrachoric correlation coefficient from the following table

	Athlete	Non-athlete	Total
Practiced	95 (A)	60 (B)	155
Unpracticed	50 (C)	85 (D)	135
Total	145	145	290

$r_t = \cos\left[\frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})}\right]$

$= \cos\left[\frac{(180° \times \sqrt{60{\times}50})}{(\sqrt{95{\times}85} + \sqrt{60{\times}50})}\right]$

$= \cos\left[\frac{(180° {\times} 54.77)}{(89.86{+}54.77)}\right]$

$= \cos(9858.6/144.63)$

$= \cos 68° = 0.37$ (Moderate positive correlation)

Sum 7: Compute the tetrachoric correlation coefficient from the following table

	Athlete	Non-athlete	Total
High achievement motivation	60 (A)	25 (B)	85
Low achievement motivation	20 (C)	45 (D)	65
Total	80	70	150

$$r_t = \cos \left[\frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})} \right]$$

$$= \cos \left[\frac{(180° \times \sqrt{25 \times 20})}{(\sqrt{60 \times 45} + \sqrt{25 \times 20})} \right]$$

$$= \cos \left[\frac{(180° \times 22.36)}{(51.96 + 22.36)} \right]$$

$$= \cos (4024.8/74.32)$$

$$= \cos 54° = 0.59 \text{ (High positive correlation)}$$

Sum 8: Compute the tetrachoric correlation coefficient from the following table

	Normal	Hypertensive	Total
High achievement motivation	47 (B)	63 (A)	110
Low achievement motivation	33 (D)	37 (C)	70
Total	80	100	180

$$r_t = \cos \left[\frac{(180° \times \sqrt{BC})}{(\sqrt{AD} + \sqrt{BC})} \right]$$

$$= \cos \left[\frac{(180° \times \sqrt{47 \times 37})}{(\sqrt{63 \times 33} + \sqrt{47 \times 37})} \right]$$

$$= \cos \left[\frac{(180° \times 41.70)}{(45.60 + 41.70)} \right]$$

$$= \cos (7506/87.3)$$

$$= \cos 86° = 0.07 \text{ (Very low positive correlation)}$$

6) YULE’S PHI COEFFICIENT (Φ)

When both the variables are genuinely dichotomized and not more than two categories (for e.g.: alive-dead, agree-disagree) are allowed then Phi coefficient of correlation may be computed. The relationship of phi coefficient to tetrachoric correlation coefficient is the same as that of point biserial r to biserial r.

Sum 1: Compute phi correlation coefficient from the following table.

	Question 1 Right	Question 1 Wrong	Total
Question 2 Right	75 (A)	60 (B)	135
Question 2 Wrong	55 (C)	80 (D)	135
Total	130	140	270

$$\Phi = \frac{(AD-BC)}{\sqrt{[(A+B)(C+D)(B+D)(A+C)]}}$$

$$= \frac{[(75\times80)-(60\times55)]}{\sqrt{(135\times135\times140\times130)}}$$

= (6000-3300) / √ 331695000

= 2700/18212.50 = 0.15 (Low positive correlation)

Sum 2: Compute phi correlation coefficient from the following table

	Item 1 No	Item 1 Yes	Total
Item 2 Yes	20 (B)	60 (A)	80
Item 2 No	30 (D)	25 (C)	55
Total	50	85	135

$\Phi = \frac{(AD-BC)}{\sqrt{[(A+B)(C+D)(B+D)(A+C)]}}$

$= \frac{[(60\times30)-(20\times25)]}{\sqrt{(80\times55\times85\times50)}}$

= (1800-500) / √ 18700000

= 1300/4324.35 = 0.3 (Moderate positive correlation)

Sum 3: 420 individuals responded to a test in which there were two items, X and Y, given in the 2×2 fold table. Compute Φ correlation coefficient.

	Item X Agree	Item X Disagree	Total
Item Y Agree	167 (A)	73 (B)	240
Item Y Disagree	83 (C)	97 (D)	180
Total	250	170	420

$\Phi = \frac{(AD-BC)}{\sqrt{[(A+B)(C+D)(B+D)(A+C)]}}$

$= \frac{[(167\times97)-(73\times83)]}{\sqrt{(240\times180\times170\times250)}}$

= (16199-6059) / √ 1836000000

= 10140/42848.57 = 0.24 (Low positive correlation)

Sum 4: Compute phi correlation coefficient from the following table

	Item 1 Pass	Item 1 Fail	Total
Item 2	100 (A)	65 (B)	165

Pass			
Item 2 Fail	35 (C)	60 (D)	95
Total	135	125	260

$$\Phi = \frac{(AD-BC)}{\sqrt{}\,[(A+B)(C+D)(B+D)(A+C)]}$$

$$= \frac{[(100\times60)-(65\times35)]}{\sqrt{}\,(165\times95\times125\times135)}$$

= (6000-2275) / √ 264515625

= 3725/16263.94 = 0.23 (Low positive correlation)

Sum 5: Compute phi correlation coefficient from the following table

	HIV-negative	HIV-positive	Total
Rh+	20 (B)	45 (A)	65
Rh-	35 (D)	15 (C)	50
Total	55	60	115

$$\Phi = \frac{(AD-BC)}{\sqrt{}\,[(A+B)(C+D)(B+D)(A+C)]}$$

$$= \frac{(45\times35)-(20\times15)]}{\sqrt{}\,(65\times50\times55\times60)}$$

= (1575-300) / √ 10725000

= 1275/3274.9 = 0.4 (Moderate positive correlation)

Sum 6: Compute phi correlation coefficient from the following table

	Living	Dead	Total
Male	60 (A)	20 (B)	80

Female	35 (C)	30 (D)	65
Total	95	50	145

$$\Phi = \frac{(AD-BC)}{\sqrt{[(A+B)(C+D)(B+D)(A+C)]}}$$

$$= \frac{[(60\times30)-(20\times35)]}{\sqrt{(80\times65\times50\times95)}}$$

$$= \frac{(1800-700)}{\sqrt{24700000}}$$

= 1100/4969.91 = 0.22 (Low positive correlation)

Sum 7: 180 individuals responded to a test in which there were two questions, 1 and 2, the response categories being agree and disagree, as shown in the 2×2fold table. Compute Φ coefficient.

	Question 1 Agree	Question 1 Disagree	Total
Question 2 Agree	75 (A)	35 (B)	110
Question 2 Disagree	25 (C)	45 (D)	70
Total	100	80	180

$$\Phi = \frac{(AD-BC)}{\sqrt{[(A+B)(C+D)(B+D)(A+C)]}}$$

$$= \frac{[(75\times45)-(35\times25)]}{\sqrt{(110\times70\times80\times100)}}$$

= (3375-875) / √ 61600000

= 2500/7848.57 = 0.32 (Moderate positive correlation)

Sum 8: Compute the coefficient of correlation from the following table, using a suitable measure.

	Question 1 Fail	Question 1 Pass	Total
Question 2 Pass	67 (B)	53 (A)	120
Question 2 Fail	33 (D)	47 (C)	80
Total	100	100	200

$$\Phi = \frac{(AD-BC)}{\sqrt{[(A+B)(C+D)(B+D)(A+C)]}}$$

$$= \frac{[(53\times 33)-(67\times 47)]}{\sqrt{(120\times 80\times 100\times 100)}}$$

$= (1749-3149)\ /\ \sqrt{96000000}$

$= -1400/9797.96 = -0.14$ (Low negative correlation)

7) CONTINGENCY COEFFICIENT

When each of the variables involved are divided into two or more categories then contingency coefficient of correlation may be computed. It is to be noted that like the other coefficients of correlation, contingency coefficient of correlation does not have limits (i.e., -1 to +1). The coefficient of correlation is converted to chi square to check its significance.

Sum 1: Compute the coefficient of contingency for the following table.

	Black hair	Brown hair	Golden hair	Total (f_r)
Mongoloid race	18 (18.3)	22 (23.2)	15 (13.4)	55
Caucasian race	12 (11.7)	16 (14.8)	7 (8.6)	35
Total (f_c)	30	38	22	90 (n)

$f_e = (f_r \times f_c)/n$ [f_e=expected frequency]

f_e values are following:

Mongoloid black: (30×55)/90 = 18.3

Mongoloid brown: (38×55)/90 = 23.2

Mongoloid golden: (22×55)/90 = 13.4

Caucasian black: ((30×35)/90 = 11.7

Caucasian brown: (38×35)/90 = 14.8

Caucasian golden: (22×35)/90 = 8.6

The statistics S and C are computed, using f_o^2 and f_e values of each cell.

[f_o = observed frequency]

$S=\sum(f_o^2 /f_e)$

$= 18^2 /18.3 + 22^2 /23.2 + 15^2 /13.4 + 12^2 /11.7 + 16^2 /14.8 + 7^2 /8.6$

$= 17.70+20.86+16.79+12.31+17.30+5.70$

$= 90.66$

$C = \surd\,[1-(n/S)]$

$= \surd\,[1-(90/90.66)]$

$= 0.09$

C is converted to χ^2 (chi square) for testing its significance.

$\chi^2 = (n\times C^2) \div (1 – C^2)$

$= [90\times(0.09)^2] \div [1-(0.09)^2]$

$= 0.729/0.9919$

$= 0.73$

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

Tabulated value of χ^2 at 0.05 level = 5.99

Tabulated value of χ^2 at 0.01 level = 9.21

The computed value of χ^2 is 0.73 which is smaller than the tabular value of χ^2 at 0.05 level of significance. Thus, the null hypothesis is accepted. This shows that there is no statistically significant correlation between race and hair colour.

Sum 2: Compute the coefficient of contingency for the table given below.

	Classical music	Rock music	Pop music	Total (f_r)
Religion-Hindu	23 (21.1)	21 (20.4)	16 (18.5)	60
Religion-Christian	13 (19.3)	24 (18.7)	18 (17)	55
Religion-Buddhist	22 (17.6)	11 (17)	17 (15.5)	50
Total (f_c)	58	56	51	165 (n)

$f_e = (f_r \times f_c)/n$ [f_e=expected frequency]

f_e values are following

Hindu classical: (58×60)/165 = 21.1

Hindu rock: (56×60)/165 = 20.4

Hindu pop: (51×60)/165 = 18.5

Christian classical: (58×55)/165 = 19.3

Christian rock: (56×55)/165 = 18.7

Christian pop: (51×55)/165 = 17

Buddhist classical: (58×50)/165 = 17.6

Buddhist rock: (56×50)/165 = 17

Buddhist pop: (51×50)/165 = 15.5

The statistics S and C are computed, using f_o^2 and f_e values of each cell.

[f_o = observed frequency]

$S=\sum(f_o^2 / f_e)$

$= 23^2 /21.1 + 21^2 /20.4 + 16^2 /18.5 + 13^2 /19.3 + 24^2 /18.7 + 18^2 /17 + 22^2 /17.6 + 11^2 /17 + 17^2 /15.5$

= 25.07+21.62+13.84+8.76+30.80+19.06+27.5+7.12+18.65

= 172.42

C = √ [1-(n/S)]

= √ [1-(165/172.42)]

= 0.21

C is converted to χ^2 (chi square) for testing its significance.

χ^2 = (n×C^2) / (1-C^2)

= [165×(0.21)^2 / [1 - (0.21)^2]

= 7.2765/0.9559 = 7.61

df=(r-1)(c-1) = (3-1)(3-1) = 4

Tabulated value of χ^2 at 0.05 level=9.49

Tabulated value of χ^2 at 0.01 level=13.28

The computed value of χ^2 is 7.61which is smaller than the tabulated value of χ^2 at 0.05 level of significance. Thus, the null hypothesis is accepted. This shows that there is no statistically significant correlation between music preference and religious group.

Sum 3: The responses given by three groups on an interest inventory are given below. Find out if there is any relationship between nationality and interest.

	Literature	Sports	Politics	Science and technology	Total (f_r)
Indian	20 (14.7)	35 (45.3)	40 (33)	18 (20)	113
Chinese	13 (17)	57 (52.6)	38 (38.3)	23 (23.2)	131
Japanese	11 (12.3)	44 (38.1)	21 (27.7)	19 (16.8)	95
Total (f_c)	44	136	99	60	339 (n)

f_e = (f_r×f_c)/n , where f_e=expected frequency

f_e values

Indian literature: (44×113)/339 = 14.7

Indian sports: (136×113)/339 = 45.3

Indian politics: (99×113)/339 = 33

Indian science and technology: (60×113)/339 = 20

Chinese literature: (44×131)/339 = 17

Chinese sports: (136×131)/339 = 52.6

Chinese politics: (99×131)/339 = 38.3

Chinese science and technology: (60×131)/339 = 23.2

Japanese literature: (44×95)/339 = 12.3

Japanese sports: (136×95)/339 = 38.1

Japanese politics: (99×95)/339 = 27.7

Japanese science and technology: (60×95)/339 = 16.8

The statistics S and C are computed, using f_o^2 and f_e values of each cell.

[f_o = observed frequency]

$S=\sum(f_o^2 /f_e)$

$= 20^2 /14.7 + 35^2 /45.3 + 40^2 /33 + 18^2 /20 + 13^2 /17 + 57^2 /52.6 + 38^2 /38.3 + 23^2 /23.2 + 11^2 /12.3 + 44^2 /38.1 + 21^2 /27.7 + 19^2 /16.8$

$= 27.21+27.04+48.48+16.2+9.94+61.77+37.70+22.80+9.84+50.81+15.92+21.49$

$= 349.2$

$C = \sqrt{[1-(n/S)]}$

$= \sqrt{[1-(339/349.2)]}$

$= 0.17$

C is converted to χ^2 (chi square) for testing its significance.

$\chi^2 = (n\times C^2)/(1 – C^2)$

$= [339\times(0.17)^2] / [1-(0.17)^2]$

$= 9.7971/0.9711 = 10.09$

df= (r-1)(c-1) = (3-1)(4-1) = 6

Tabulated value of χ^2 at 0.05 level= 12.59

Tabulated value of χ^2 at 0.01 level= 16.81

The computed value of χ^2 is 10.09 which is smaller than the tabulated value of χ^2 at 0.05 level of significance. Thus, the null hypothesis is accepted. This shows that there is no statistically significant correlation between nationality and interest.

Sum 4: Compute contingency coefficient from the following table and find out if there is a relationship between socioeconomic status and chosen sport

	Upper class	Middle class	Lower class	Total (f_r)
Cricket	70 (64.94)	82 (72.81)	31 (45.26)	183
Football	42 (58.55)	63 (65.65)	60 (40.81)	165
Hockey	53 (41.52)	40 (46.55)	24 (28.94)	117
Total (f_c)	165	185	115	465 (n)

$f_e = (f_r \times f_c)/n$, where f_e = expected frequency

f_e values are following

Cricket upper class: (165×183)/465 = 64.94

Cricket middle class: (185×183)/465 = 72.81

Cricket lower class: (115×183)/465 = 45.26

Football upper class: (165×165)/465 = 58.55

Football middle class: (185×165)/465 = 65.65

Football lower class: (115×165)/465 = 40.81

Hockey upper class: (165×117)/465 = 41.52

Hockey middle class: (185×117)/465 = 46.55

Hockey lower class: (115×117)/465 = 28.94

The statistics S and C are computed, using f_o^2 and f_e values of each cell., where f_o = observed frequency.

$S=\sum(f_o^2/f_e)$

$= 70^2/64.94 + 82^2/72.81 + 31^2/45.26 + 42^2/58.55 + 63^2/65.65 + 60^2/40.81 + 53^2/41.52 + 40^2/46.55 + 24^2/28.94$

$= 75.45+92.35+21.23+30.13+60.46+88.21+67.65+34.37+19.90$

$= 489.75$

$C = \sqrt{[1-(n/S)]}$

$= \sqrt{[1-(465/489.75)]}$

$= 0.22$

C is converted to χ^2 (chi square) for testing its significance.

χ^2 = (n×C square) / $(1-C^2)$

= [465×(0.22)square] / [1-(0.22)square]

= 22.506/0.9516 = 23.65

df= (r-1)(c-1) = (3-1)(3-1) = 4

Tabulated value of χ^2 at 0.05 level = 9.49

Tabulated value of χ^2 at 0.01 level = 13.28

The computed value of χ^2 is 23.65 which is greater than the tabulated value of χ^2 at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that there is a statistically significant correlation between socioeconomic status and chosen sport.

Different types of Correlation and their uses

Following are the different types of correlations.

Pearson's Product Moment Correlation (Pearson r)

The following considerations are to be made for computing Pearson r:

- The two variables for which the relationship between which is to be measured must be continuous variables.
- The relationship between the two variables must follow a straight line or trajectory. In other words, the relationship should be linear.
- The distribution must be unimodal and fairly symmetrical.

- The measures of variance or standard deviation of the scores must be homogeneous.

Spearman’s Rank Difference Correlation

- For the two variables, the relationship between which is to be measured must be given in terms of rank (ordinal variable).
- The distributions in which the data is expressed should be in terms of ranks only, Spearman’s coefficient of correlation is to be computed.

Biserial Correlation

- Biserial correlation is computed between two variables when one is a continuous measurement variable and the other one is apparently dichotomized, that is, divided into two categories. For instance, the behavioural science researcher might be interested in knowing the relationship between intelligence and success. The variable ‘success’ in turn may be dichotomized into successful and unsuccessful. This is an artificial dichotomy because there is no clear-cut point or criteria for such a division.
- The dichotomous variable must be dichotomized at a point near the median.
- The variables involved must be continuous as well as normal or near normal.

Point Biserial Correlation

- Point Biserial Correlation is computed between two variables when one is a continuous measurement variable and the other one is genuinely dichotomized. For instance, the behavioural science researcher might be interested in knowing the relationship between sex and memory test scores. The variable ‘sex’ is dichotomized into male and female, which is a genuine dichotomy because the point or criteria for such a division is clear-cut.
- Assumptions regarding the form of distribution of the dichotomous variable are not as stringent as that in biserial correlation which has too many assumptions such as normality and continuity.
- It is safe to compute point biserial correlation when the researcher is not sure whether the dichotomy is artificial or genuine.

Tetrachoric Correlation

- Tetrachoric correlation is computed when both the variables involved are artificially dichotomous and none of them are expressed in terms of scores.
- If the dichotomy is artificial then tetrachoric correlation may be computed. For instance, the behavioural science researcher might be interested in knowing the relationship between neurosis and adjustment. The variable ‘neurosis’ may be dichotomized into neurotic and non-neurotic and the variable ‘adjustment’ may be dichotomized into adjusted and maladjusted. The data is then arranged in a 2×2 contingency table.

- Both the variables must fulfil the assumptions of normality, continuity and linearity of relationship with each other.

Phi Correlation

- Phi coefficient of correlation is computed when both the variables involved are genuinely dichotomous. For instance, when the classification of variables is in terms of yes-no, true-false, phi coefficient of correlation may be computed. The observations are arranged in a 2×2 contingency table.
- No assumptions are made regarding the form of distribution of the dichotomized variables.
- It may be used in item analysis.
- Phi coefficient of correlation has a relationship with chi square.
- It is safe to compute the phi coefficient of correlation when the researcher is not sure whether the dichotomy is artificial or genuine.

Contingency Coefficient

- The Contingency coefficient may be computed between two variables when both are classified into two or more categories. The categories may be 3×3, 3×4, 4×4 and so on. For instance, if the researcher is interested to find out the relationship between race (Mongoloid, Caucasian) and hair color (black, brown, golden), Contingency coefficient may be used.

Exercise

1. What do you mean by correlation? Why is it used in social science research?
2. What conditions must be satisfied before computing product moment correlation?
3. How is point biserial correlation different from biserial correlation?
4. Discuss the similarities and differences between tetrachoric and phi coefficient of correlation.
5. What is contingency coefficient?
6. Compute the coefficient of correlation between the two sets of scores given below:
   i.

Subject	Numerical Reasoning	Logical Reasoning
1	83	92
2	67	73
3	79	76
4	78	84

5	81	85
6	80	79
7	86	93
8	82	89
9	78	90
10	77	91

ii.

Subject	IQ values	Numerical Reasoning
1	109	107
2	93	100
3	95	98
4	101	103
5	100	99
6	98	103
7	87	90
8	85	86
9	96	101
10	102	105

7. Compute the coefficient of correlation between the following sets of scores using the rank-difference method.

   i.

Individuals	Spelling	Clerical Speed
1	87	52
2	76	54

3	83	51
4	72	67
5	68	63
6	61	74
7	79	72
8	82	64
9	69	61
10	74	62

ii.

Individuals	Mechanical Reasoning	Spatial Reasoning
1	58	51
2	64	62
3	52	57
4	68	63
5	71	65
6	57	53
7	54	50
8	65	62
9	59	56
10	63	60

8. The following data shows the scores obtained by individuals in an adjustment inventory and they are classified into 2 groups – adjusted and maladjusted. Compute the coefficient of correlation between the two groups.
   i.

Scores	Adjusted	Maladjusted
70-79	11	2
60-69	16	3
50-59	20	5
40-49	18	6
30-39	13	4
20-29	9	2
10-19	5	1

9. Compute point biserial coefficient of correlation from the following data.
   i.

Scores on opinion scale	Agree	Disagree
55-59	0	2
50-54	2	6
45-49	3	7
40-44	7	11
35-39	5	9
30-34	4	8
25-29	3	5
20-24	1	2

10. Compute the tetrachoric correlation coefficient from the given data to find out the relationship between emotional intelligence and adjustment.
    i.

	Adjusted	Maladjusted
High emotional intelligence	70	40
Low emotional intelligence	30	60

ii.

	Adjusted	Maladjusted
High achievement motivation	65	35
Low achievement motivation	45	75

11. Compute the phi correlation coefficient from the following tables:

i.

	Item	1	
Item		Yes	No
2	Yes	80	30
	No	20	70

ii.

	Question	1	
Question		Yes	No
2	Yes	90	70
	No	30	60

12. Compute the contingency coefficient of correlation using the following data:

			Interest			
		Business	Gaming	IT	Literature	Music
	Science	11	26	30	18	13
Stream	Humanities	12	11	8	25	29
	Commerce	27	13	12	7	8

Table: Ordinates corresponding to divisions of the area under the normal curve into a larger proportion (p) and smaller proportion (q)

The larger area p	The smaller area q	Ordinate (y)
0.50	0.50	0.399
0.58	0.42	0.391
0.60	0.40	0.386
0.67	0.33	0.362
0.70	0.30	0.348
0.71	0.29	0.342
0.74	0.26	0.324
0.80	0.20	0.280
0.90	0.10	0.175

Chapter 6

SIGNIFICANCE OF THE DIFFERENCE BETWEEN MEANS

Following tests are used for testing the statistically significant difference between small sample means.

t test

When the researcher is required to find out whether the difference between the means of two groups is significant or not then t test may be computed. The groups may be dependent or independent. By dependent, we mean that the two sets of scores are related to each other. For instance, the Beck Depression Inventory scores of depressed individuals before and after Cognitive Behaviour Therapy (CBT) are said to be related or dependent because the post-therapy scores are hypothesized to depend on the CBT.

A t test is a type of inferential statistic, which is essentially used as a hypothesis testing tool, and helps in testing an assumption applicable to a population. Computing a t test needs three data values, which include the mean differences of each data set, the standard deviation of each group and the number of data values of each group.

The concepts of degrees of freedom, hypothesis and levels of significance are important in this context.

Degree of freedom means freedom to vary and it is abbreviated as df. The degree of freedom is the number of observations that are independent of each other. It is essential for determining the importance and validity of the null hypothesis.

After formulation of a research problem and review of literature it is necessary to state a hypothesis which is a tentative solution to the problem. By testing hypotheses, we test an assumption regarding a population parameter. It is used to assess the probability of a hypothesis by using the sample data. The hypothesis may be stated in terms of null or alternative hypotheses. Null hypothesis signifies no difference. For instance, there is no significant difference between the means of males and females in terms of their smartphone use. On the other hand, the alternative hypothesis states that there exists a significant difference between the means of the two groups. After data collection following standard procedures statistical analyses are carried out and on the basis of the findings the hypotheses are accepted or rejected based on levels of significance. The commonly used levels of significance are the 0.05 or 5% level and the 0.01 or 1% level. If the null hypothesis is rejected at 0.05 level it means that in 95 out 100 replications of the experiment or study the null hypothesis would be false and 5 times it would be true. Similarly, if the null hypothesis is rejected at 0.01 level of significance it means that in 99 times out of 100 replications of the experiment or study the null hypothesis would be false and once it would be true.

Uses of t test

- t test can be used to determine whether the means of two independent or related groups differ significantly or not on the basis of certain predetermined characteristics. For example, a researcher may be interested to find out whether there is a significant difference between males and females in terms of their smartphone use. Here, sex is the independent variable, the levels of which are males and females and smart phone use is the dependent variable.

  In another instance the researcher may be interested to find out whether the achievement test scores of a group of individuals differ before and after practice.

  The first one is an example of independent samples t test and the second one is an example of related samples t test.
- t test being a parametric test, it is more powerful than the nonparametric tests and the statistical inference drawn on the basis of t test is more stable than its nonparametric counterparts.
- It is one of the important statistics used to analyze a hypothesis.

List of formulae

Sample	Formula	Degrees of freedom
1. Small independent equal	$t = \dfrac{\lvert(M1\ -\ M2)\rvert}{\surd\,[\{\sum(X1 - M1)^2\ +\sum(X2 - M2\,)^2\,\}\,/\,N(N-1)]}$	2(N-1)
2. Small independent unequal	$t = \dfrac{\lvert\,(M1-M2)\rvert}{\surd\,[\{(\sum x1^2 +\ \sum x2^2\,)/(N1+N2-2)\}\times\{(N1+N2)/N1N2\}]}$	N_1+N_2-2
3. Large independent equal	$t = \lvert(M_1 - M_2)\rvert\ /\ \surd\ \{(S_1^2 + S_2^{2)}/N\}$	2(N-1)
4. Large independent unequal	$t = \lvert(M_1 - M_2)\rvert\ /\ \surd\ [(S_1^2\ /N_1)+(S_2^2/N_2)]$	N_1+N_2-2
5. Dependent small	$t =[\ \sum D]\ /\ \surd\ [\{N\sum D^2\ - (\sum D)^2\ \}/(N-1)]$	N-1

	D = Difference in score between initial and final testing (in case of single group) or between pairs of matched subjects (in case of equivalent groups)	
6. Dependent large	$t = \|(M_1 – M_2)\| / \sqrt{} [(S_{\bar{X}_1})^2 + (S_{\bar{X}_2})^2 – 2.r.(S_{\bar{X}_1})(S_{\bar{X}_2})]$ $S_{\bar{X}_1} = S_1/\sqrt{} N$ $S_{\bar{X}_2} = S_2/\sqrt{} N$	N-1

Examples:

Sum 1: The scores obtained by males and females on a neuroticism inventory are given below. Test the significance of the difference between means of the two groups using an appropriate statistical technique.

Males (X_1)	Females (X_2)	$x_1=X_1-M_1$	$x_2=X_2-M_2$	x_1^2	x_2^2
50	53	16.6	16.4	275.56	268.96
33	37	-0.4	0.4	0.16	0.16
41	42	7.6	5.4	57.76	29.16
28	34	-5.4	-2.6	29.16	6.76
35	36	1.6	-0.6	2.56	0.36
31	35	-2.4	-1.6	5.76	2.56
25	28	-8.4	-8.6	70.56	73.96
29	29	-4.4	-7.6	19.36	57.76
32	33	-1.4	-3.6	1.96	12.96
30	39	-3.4	2.4	11.56	5.76
Total=334	366			474.4	458.4

Mean (M_1) = 334/10 = 33.4

Mean (M_2) = 366/10 = 36.6

$$t = \frac{|(M1\ -\ M2)|}{\sqrt{[\{\sum(X1 - M1)^2\ + \sum(X2 - M2\ )^2\ \}\ /\ N(N-1)]}}$$

$$= \frac{(36.6-33.4)}{\sqrt{[(474.4+458.4)\ /\ (10\times 9)]}}$$

$= 3.2/\sqrt{10.364} = 3.2/3.2194 = 0.99$

df= 2(N-1) = 2×9 = 18

Tabulated value of t at 0.05 level = 2.101

Tabulated value of t at 0.01 level = 2.878

The computed value of t is 0.99 which is smaller than the tabulated value of t at 0.05 level of significance. Thus, the null hypothesis is accepted. This shows that there is no statistically significant difference between males and females in terms of their neuroticism scores.

Sum 2: The body weights of 12 adult males and 8 adult females are given below. Find whether the mean weight of males is significantly higher than that of females.

Males (X_1)	Females (X_2)	x_1=X_1-M_1	x_2=X_2-M_2	x_1^2	x_2^2
68	58	3.2	-3.3	10.24	10.89
56	57	-8.8	-4.3	77.44	18.49
72	63	7.2	1.7	51.84	2.89
63	59	-1.8	-2.3	3.24	5.29
55	47	-9.8	-14.3	96.04	204.49
75	71	10.2	9.7	104.04	94.09
80	70	15.2	8.7	231.04	75.69
67	65	2.2	3.7	4.84	13.69
59		-5.8		33.64	

65		0.2		0.04	
58		-6.8		46.24	
60		-4.8		23.04	
Total=778	490			681.68	425.52

M_1= 778/12 = 64.8

M_2= 490/8 = 61.3

$$t = \frac{|(M1-M2)|}{\sqrt{[\{(\sum x1^2 + \sum x2^2)/(N1+N2-2)\}\times\{(N1+N2)/N1N2\}]}}$$

$$= \frac{(64.8-61.3)}{\sqrt{[\{(681.68+425.52)\div(12+8-2)\}\times\{(12+8)\div(12\times8)\}]}}$$

$= 3.5 / \sqrt{(61.51\times0.208)} = 3.5/3.58 = 0.98$

df= N_1+N_2-2 = 12+8-2 = 18

Tabulated value of t at 0.05 level = 1.734

Tabulated value of t at 0.01 level = 2.552

The calculated value of t is 0.98 which is smaller than the tabulated value of t at 0.05 level of significance. Thus, the null hypothesis is accepted. This shows that there is no statistically significant difference between males and females in terms of their weight.

Sum 3: The mean and S.D. of the heights of 568 boys of the 10th standard were found to be 5.5 feet and 0.6 feet and that of 450 girls were found to be 5.1 feet and 0.4 feet respectively. Is the mean height of the 10th standard boys significantly higher than that of the girls?

$t = |(M_1 - M_2)| / \sqrt{[(S_1^2/N_1)+(S_2^2/N_2)]}$

$= (5.5-5.1) / \sqrt{[(0.6)^2/568)+(0.4)^2/450)]}$

$= 0.4 / \sqrt{(0.0006338+0.000356)}$

= 0.4/0.03146 = 12.7

df= N_1+N_2-2 = 568+450-2 =1016

Since df is very large and hence it should be considered as df=∞ (one-tailed)

Tabulated value of t at 0.05 level = 1.645

Tabulated value of t at 0.01 level = 2.326

The calculated value of t is 12.7 which is greater than the tabulated value of t at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that there is a statistically significant difference between males and females in terms of their height.

Sum 4: The mean and S.D. of depression scores of 120 males aged 60 years were found to be 16 and 5.5 and that of 120 females aged 60 years were found to be 21 and 3.5 respectively. Apply a suitable statistical technique for testing the significance of difference between the means of the two groups and comment on the results.

$t = |(M_1 – M_2)| / \sqrt{[(S_1^2 /N_1)+(S_2^2/N_2)]}$

$= (21-16) / \sqrt{[(30.25+12.25)/120]}$

$= 5/0.595 = 8.4$

df= 2(N-1) = 2×119 = 238 as $N_1 = N_2$

Since df is very large and hence it should be considered as df=∞ (two-tailed)

Tabulated value of t at 0.05 level = 1.960

Tabulated value of t at 0.01 level = 2.576

The calculated value of t is 8.4 which is greater than the tabulated value of t at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that there is a statistically significant difference between 60-year old males and females in terms of their depression scores.

Paired t test: Sometimes we are interested to test the average significance difference between two samples when data of one sample is collected before conducting the experiment while data of other sample is collected after conducting the experiment. In such cases we use the pair t test.

Sum 5: The achievement test scores of 12 students before and after practice are given below. Does practice make a significant difference in achievement test scores?

Students	Before practice (X_1)	After practice (X_2)	$D=\|X_2-X_1\|$	D^2
1	72	78	6	36

2	63	74	11	121
3	70	82	12	144
4	58	73	15	225
5	75	88	13	169
6	62	71	9	81
7	54	68	14	196
8	67	81	14	196
9	59	75	16	256
10	66	70	4	16
11	60	69	9	81
12	71	79	8	64
Total			131	1585

$t = [\sum D] / \surd\ [\{N\sum D^2 - (\sum D)^2\}/(N-1)]$

$= 131 / \surd\ [\{12\times1585 - (131)^2\}/(12-1)]$

$= 131 / \surd\ [(19020-17161)/11]$

$= 131/13 = 10.08$

$df = N-1 = 12-1 = 11$

Tabulated value of t at 0.05 level = 2.201

Tabulated value of t at 0.01 level = 3.106

The calculated value of t is 10.08 which is greater than the tabulated value of t at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that there is a statistically significant difference in achievement test scores before and after practice.

Sum 6: The mean and S.D. of scores obtained by consisting of 55 individuals before receiving experimental treatment and after receiving experimental treatment is as follows:

	Mean	S.D.
Before treatment	56.3 (M_1)	8.17 (S_1)
After treatment	71.2 (M_2)	9.33 (S_2)

Here N=55. The correlation coefficient between the test scores, before and after experimental treatment, was found to be 0.58. Find whether there is a significant difference between the mean test scores before and after receiving experimental treatment.

$t = |(M_1 - M_2)| \,/\, \sqrt{}\ [(\ S_{\bar{X}_1})^2 + (S_{\bar{X}_2})^2 - 2.r.(\ S_{\bar{X}_1}))(S_{\bar{X}_2})]$

$S_{\bar{X}_1} = S_1/\sqrt{}\ N = 8.17/\sqrt{}\ 55 = 1.10$

$S_{\bar{X}_2} = S_2/\sqrt{}\ N = 9.33/\sqrt{}\ 55 = 1.26$

$t = (71.2\text{-}56.3) / \sqrt{}\ [(1.10)^2 + (1.26)^2 - 2\times0.58\times1.10\times1.26)]$

$= 14.9 / \sqrt{}\ (1.21+1.59-1.608)$

$= 14.9/\sqrt{}\ 1.192 = 14.9/1.09 = 13.67$

df = N-1 = 55-1 = 54

Tabulated value of t at 0.05 level = 2.005

Tabulated value of t at 0.01 level = 2.670

The calculated value of t is 13.67 which is greater than the tabulated value of t at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that there is a statistically significant difference between the mean test scores before and after receiving experimental treatment.

Sum 7: The attitude test scores of 10 individuals before and after experimental treatment are given below. Is the difference between the means of two groups significant at 0.01 level ?

Individuals	Before treatment (X_1)	After treatment (X_2)	D=$\lvert X_2-X_1\rvert$	D ^{2}
1	32	34	2	4
2	21	30	9	81
3	26	29	3	9
4	30	36	6	36
5	18	28	10	100
6	23	31	8	64
7	17	29	12	144
8	20	32	12	144
9	22	27	5	25
10	28	35	7	49
Total			74	656

$t = \sum D / \sqrt{[\{N\sum D^2 - (\sum D)^2\}/(N-1)]}$

$= 74 / \sqrt{[\{(10\times656)-(74)^2\}/9]}$

$= 74/\sqrt{120.44} = 6.75$

df=N-1 = 10-1 = 9

Tabulated value of t at 0.05 level = 2.262

Tabulated value of t at 0.01 level = 3.250

The calculated value of t is 6.75 which is greater than the tabulated value of t at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that there is a statistically significant difference between the means of attitude test scores before and after experimental treatment.

Remarks: Such t test is also called paired t test.

Sum 8: The vocabulary test scores obtained by a random sample of 10 male students and 10 female students of the 8th standard are given below. Test the significance of the difference between means of the two groups using the appropriate statistical technique.

Scores obtained by males (X_1)	Scores obtained by females (X_2)	x_1=X_1-M_1	x_2=X_2-M_2	x_1^2	x_2^2
35	37	3	4.3	9	18.49
27	26	-5	-6.7	25	44.89
24	29	-8	-3.7	64	13.69
42	40	10	7.3	100	53.29
36	38	4	5.3	16	28.09
39	41	7	8.3	49	68.89
25	24	-7	-8.7	49	75.69
31	30	-1	-2.7	1	7.29
33	35	1	2.3	1	5.29
28	27	-4	-5.7	16	32.49
Total=320	327			330	348.1

$M_1 = 320/10 = 32$

$M_2 = 327/10 = 32.7$

$t= |(M_1 – M_2) / \sqrt{} [\{\sum(X_1\text{-}M_1)^2 + \sum(X_2\text{-}M_2)^2 \} / N(N\text{-}1)]$

$= (32.7\text{-}32) / \sqrt{} [(330+348.1)/(10\times9)]$

$= 0.7/2.74 = 0.26$

df=2(N-1) = 2×9 = 18 as $N_1 = N_2$

Tabulated value of t at 0.05 level = 2.101

Tabulated value of t at 0.01 level = 2.878

The calculated value of t is 0.26 which is smaller than the tabulated value of t at 0.05 level of significance. Thus, the null hypothesis is accepted. This shows that there is no statistically significant difference between the mean vocabulary test scores of the two groups.

Exercise

1. When is t test computed?
2. State the uses of t test.
3. The mean and Standard Deviation of the weights of 160 girls and 180 boys are as follows:

	Mean	S.D.
Girls	60	7.30
Boys	72	9.80

Test the significance of the difference between the mean of the two groups at 0.05 level of significance.

4. The scores obtained by students in a Spelling test are as follows:
XA – 14, 17, 12, 9, 8, 11, 13, 15, 7
XB – 18, 15, 11, 8, 9, 16, 17, 13, 12, 14
Is the difference between the means of two groups significant at 0.05 level?
5. The performance (in seconds) of athletes before and after consuming energy drinks is given below.
Before – 42, 37, 43, 33, 46, 41, 35, 40, 45, 39
After – 33, 29, 36, 27, 30, 31, 27, 32, 34, 28
Does consuming energy drinks bring about a significant difference in the performance of the athletes?
6. A teacher wanted to know whether teaching method has any effect on student's achievement scores. For this purpose she divided her class into two random groups (A and B) and taught group A by the lecture method and group B by the demonstration method. After teaching for a given period of time she administered an achievement test to evaluate the students' performance. The obtained data is as follows:

	Group A	Group B
Mean	62	74
SD	8	9
Number of students	54	58

7. A researcher wanted to test the efficacy of a drug for reducing anxiety. For this purpose he selected two groups (experimental and control) of individuals who were matched in pairs. The drug was administered to the subjects in the experimental group only and not the control group. He assessed the level of anxiety of the subjects in both the groups by administering a self-report inventory meant for assessing anxiety. The obtained scores are given below:

Experimental group	Control group
51	57
48	47
47	52
45	50
43	48
42	49
40	39
37	46
35	38
34	37

Test the null hypothesis at the 0.05 level of significance.

8. Two groups of students are matched for mean and SD in an intelligence test. The scores of these two groups of students on an aptitude battery is as follows:

	Group A	Group B
Mean	35	43
SD	5.50	6.20
N	50	50

The correlation between the intelligence test scores and scores obtained in aptitude battery by the entire group of students is 0.55. Test the significance of the difference between the two groups A and B at the 0.05 level of significance.

Critical values of t

Degrees of freedom	P 0.05	P 0.01
1	t = 12.71	t = 63.66
2	4.30	9.92
3	3.18	5.84
4	2.78	4.60
5	2.57	4.03
6	2.45	3.71
7	2.36	3.50
8	2.31	3.36
9	2.26	3.25
10	2.23	3.17
11	2.20	3.11
12	2.18	3.06
13	2.16	3.01
14	2.14	2.98
15	2.13	2.95
16	2.12	2.92
17	2.11	2.90
18	2.10	2.88
19	2.09	2.86
20	2.09	2.84
21	2.08	2.83
22	2.07	2.82
23	2.07	2.81
24	2.06	2.80
25	2.06	2.79
26	2.06	2.78
27	2.05	2.77
28	2.05	2.76
29	2.04	2.76
30	2.04	2.75
35	2.03	2.72
40	2.02	2.71
45	2.02	2.69
50	2.01	2.68
60	2.00	2.66
70	2.00	2.65
80	1.99	2.64
90	1.99	2.63
100	1.98	2.63
125	1.98	2.62
150	1.98	2.61
200	1.97	2.60
300	1.97	2.59
400	1.97	2.59
500	1.96	2.59
1000	1.96	2.58
∞	1.96	2.58

CHAPTER 7

CHI SQUARE TEST (χ^2)

The Chi square test was formulated to compute categorical data, it is mainly required for data comprising qualitative categories; for example colour of eyeball, gender etc. A chi square is a test which measures how expectations are compared to actual observed data. It is used when the scores are on a nominal variable and they are represented in terms of frequencies. , that is, the number of observations falling into different categories. Chi square may be computed to find out whether there exists any difference between observed and expected frequencies, that is; this statistic imparts a computation of the difference between observed frequencies and expected frequencies. It is used as a test of goodness of fit to determine whether the observed results of an experiment or study differ from the theoretically obtained results based on some hypothesis. Goodness-of-fit tests are statistical measurements aiming to determine how well sample data fit a distribution from a population with a normal distribution. For example, if an unbiased coin is tossed 10 times it is expected that 5 times out of 10 it will come out to be heads and the remaining 5 times it will come out to be tail. However, in reality when a coin is tossed several different combinations may be possible such as 6 heads-4 tails, 7 heads-3 tails and so on. Chi square test may be used to compare an observed frequency distribution to an expected frequency distribution. Then it is to be checked whether the obtained value of chi square indicates a greater mismatch than we would expect by chance. This is to be accomplished by comparing the obtained value of chi square to the critical value of chi square based on the degrees of freedom.The degrees of freedom refers to the highest number of logically independent values, which are values that have the freedom to vary. It is essential when we are computing chi-square and the validity of null hypothesis.

Hypothesis of Equal Probability – Suppose a teacher wants to introduce a new method of teaching for which she seeks opinion from the students. The response categories are agree, disagree and neutral. According to the equal probability hypothesis, if there are 90 individuals then the chances of getting agree, disagree and neutral will be equal, that is, 30 in each category.

The equation for chi square (χ^2) is given below:

$\chi^2 = \sum [(f_o - f_e)^2 \div f_e]$

where f_o = Observed frequency on some experiment

f_e = Expected frequency on some hypothesis

Sum 1: 60 call centre executives were rated into three groups: Efficient, Satisfactory, Poor in terms of their job performance by their boss as shown below.

Efficient	Satisfactory	Poor
30	25	5

Does the distribution of ratings differ significantly from that to be expected if job performance is normally distributed in the population of call centre executives?

Here the observed frequencies are 30, 25 and 5.

According to the equal probability hypothesis the expected frequencies will be 20 (60÷3) in each category. Thus, the null hypothesis may be stated as follows:

There is no significant difference between the observed and expected frequencies.

Computation of f_e: (30+25+5)/3 = 20

Computation of χ^2 :

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$[(f_o-f_e)^2 /f_e]$
30	20	10	100	5
25	20	5	25	1.25
5	20	-15	225	11.25
				χ^2 = 17.5

(After determining the expected frequencies the number of rows become 3 and the number of columns become 2)

df = (r-1)(c-1) , where r denotes number of rows and c denotes number of columns.

df = (3-1)(2-1) = 2

Tabulated value of χ^2 at 0.05 level = 5.99

Tabulated value of χ^2 at 0.01 level = 9.21

The calculated value of χ^2 (17.5) is greater than the tabulated value of χ square at 0.01 level of significance. Thus, the null hypothesis is rejected. The distribution of ratings differs significantly from that to be expected if job performance is normally distributed in the population of call centre executives.

Sum 2: The responses of two groups of students on an item of an attitude scale questionnaire were recorded as follows:

	Strongly disagree	Disagree	Undecided	Agree	Strongly agree	Total (f_r)
Science students	8 (9.1)	10 (10.9)	5 (5.2)	12 (13)	15 (11.7)	50
Humanities students	13 (11.9)	15 (14.1)	7 (6.8)	18 (17)	12 (15.3)	65
Total (f_c)	21	25	12	30	27	115 (n)

This table follows 5 x 2 contingency table and hence expected frequencies (f_e) are computes as following

$f_e = (f_r \times f_c)/n$

$(21\times50)/115 = 9.1$

$(25\times50)/115 = 10.9$

$(12\times50)/115 = 5.2$

$(30\times50)/115 = 13$

$(27\times50)/115 = 11.7$

$(21\times65)/115= 11.9$

$(25\times65)/115 = 14.1$

$(12\times65)/115 = 6.8$

$(30\times65)/115 = 17$

$(27\times65)/115 = 15.3$

Computation of χ^2:

$\chi^2 = \sum[(f_o-f_e)^2 /f_e]$

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$[(f_o-f_e)^2/f_e]$
8	9.1	-1.1	1.21	0.133
10	10.9	-0.9	0.81	0.074
5	5.2	-0.2	0.04	0.008
12	13	-1	1	0.077
15	11.7	3.3	10.89	0.931
13	11.9	1.1	1.21	0.102
15	14.1	0.9	0.81	0.057
7	6.8	0.2	0.04	0.006
18	17	1	1	0.059
12	15.3	-3.3	10.89	0.712
115				χ^2 =2.159

df=(r-1)(c-1) = (2-1)(5-1) = 4

Tabulated value of χ^2 at 0.05 level = 9.488

Tabulated value of χ^2 at 0.01 level = 13.277

The calculated value of χ^2 is 2.159 which is smaller than the critical value of χ^2 at 0.05 level of significance; therefore the obtained value is insignificant. The null hypothesis is accepted. The given data do not provide an indication of significant difference in attitude held by students of science and humanities.

Sum 3: The table below shows the number of males and females who chose each of the three possible answers to an item on an interest inventory. Does this item differentiate between males and females? Test the null hypothesis.

	Yes	No	Undecided	Total (f_r)
Males	36 (31.6)	15 (21)	20 (18.4)	71
Females	84 (88.4)	65 (59)	50 (51.6)	199

Total (f_c)	120	80	70	270 (N)

Computation of expected frequencies:

$f_e = (f_r \times f_c)/n$ [f_e=expected frequency]

(120×71)/270 = 31.6

(80×71)/270 = 21

(70×71)/270 = 18.4

(120×199)/270 = 88.4

(80×199/270 = 59

(70×199)/270 = 51.6

Computation of χ^2 :

$\chi^2 = \sum[(f_o-f_e)^2/f_e]$

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$[(f_o-f_e)^2/f_e]$
36	31.6	4.4	19.36	0.61
15	21	-6	36	1.71
20	18.4	1.6	2.56	0.14
84	88.4	-4.4	19.36	0.22
65	59	6	26	0.61
50	51.6	-1.6	2.56	0.05
				χ^2 =3.34

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

Tabulated value of χ^2 at 0.05 level = 5.99

Tabulated value of χ^2 at 0.01 level = 9.21

The calculated value of χ^2 is 3.34 which is smaller than the tabulated value of χ^2 at 0.05 level of significance. Thus, the null hypothesis is accepted. The given data do not provide an indication of significant difference in opinion held by males and females.

Sum 4: In a study, working women and non-working women were asked to express their opinion in two categories- Agree and Disagree, on the topic- “Do you prefer shopping online”? Obtained responses are given in the table below. Do the working and non-working women differ in terms of their opinion?

	Agree	Disagree	Total (f_r)
Working women	55 (A)	40 (B)	95
Non-working women	25 (C)	65(D)	90
Total (f_c)	80	105	185 (N)

χ^2 = [N(AD-BC)^2] / [(A+B)(C+D)(A+C)(B+D)]

= 185[(55×65)-(40×25)]^2 / (95×90×80×105)

= [185×(3575-1000)^2] / 71820000

= (185×6630625) / 71820000

= 17.08

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

Tabulated value of χ^2 at 0.05 level = 3.84

Tabulated value of χ^2 at 0.01 level = 6.64

The calculated value of χ^2 is 17.08 which is greater than the critical value of χ^2 at 0.01 level of significance. Thus, the null hypothesis is rejected. The working and non-working women differ significantly in terms of their opinion.

Sum 5: In a study, 12-year old boys of upper class and middle class were asked to express their opinion in three categories-yes, no, and undecided to the question- “Do you prefer playing vieo games to outdoor games”? Obtained responses are shown in the table below. Is socioeconomic status independent of the opinion expressed by the boys?

	Yes	No	Undecided	Total (f_r)
Upper class	32 (26.5)	12 (21.2)	16 (12.2)	60
Middle class	18 (23.5)	28 (18.8)	7 (10.8)	53
Total (f_c)	50	40	23	113 (N)

Computation of expected frequencies:

$f_e = (f_r \times f_c)/n$ [f_e=expected frequency]

(50×60)/113 = 26.5

(40×60)/113 = 21.2

(23×60)/113 = 12.2

(50×53)/113 = 23.5

(40×53)/113 = 18.8

(23×53)/113 = 10.8

Computation of χ^2:

$\chi^2 = \sum[(f_o-f_e)^2/f_e]$

f_o	f_e	f_o-f_e	$(f_o-f_e)^2$	$[(f_o-f_e)^2/f_e]$
32	26.5	5.5	30.25	1.14
12	21.2	-9.2	84.64	3.99
16	12.2	3.8	14.44	1.18
18	23.5	-5.5	30.25	1.29
28	18.8	9.2	84.64	4.50
7	10.8	-3.8	14.44	1.34
				χ^2 =13.44

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

Tabulated value of χ^2 at 0.05 level = 5.99

Tabulated value of χ^2 at 0.01 level = 9.21

The calculated value of χ^2 is 13.44 which is greater than the critical value of χ square at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that there is a statistically significant difference between upper and lower class boys in terms of their opinion.

Sum 6: The opinion of 26 married females and 26 unmarried females were taken as on an attitude scale as shown below. Do the data indicate a significant difference in opinion in terms of marital status of the females?

	Yes	No	Total (f_r)
Married	18 (A)	8 (B)	26
Unmarried	6 (C)	20 (D)	26
Total (f_c)	24	28	52 (N)

χ^2 = [N(AD-BC)^2] / [(A+B)(C+D)(A+C)(B+D)]

= 52[(18×20)-(8×6)]^2 / (26×26×24×28)

= (52×97344)/454272

= 11.14

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

Tabulated value of χ^2 at 0.05 level = 3.84

Tabulated value of χ^2 at 0.01 level = 6.64

The calculated value of χ^2 is 11.14 which is greater than the critical value of χ^2 at 0.01 level of significance. Thus, the null hypothesis is rejected. This shows that the married and unmarried females differ statistically significantly in terms of their opinion.

Yates' Correction: Correction for small frequencies in a 2×2 table

If in a small sample of cases, the least frequency in any cell of a 2×2 contingency table (df = 1) is less than 5, then computing chi square in the usual way mentioned above might give an overestimate of the true value, which in turn may result in rejection of a hypothesis which in reality should not be rejected. This difficulty may be overcome by applying Yates' correction, also known as the correction for continuity.

The rule is as follows: Subtract 0.5 from the absolute value of the difference between observed and expected frequencies, which in turn decreases the value of chi square.

The formula for computing χ^2 using Yates' correction wherever applicable is as follows:

Usual formula for computing χ^2	Formula for computing χ^2 when Yates correction is applied
$\chi^2 = \sum [(f_o - f_e)^2 \div f_e]$	$\chi^2 = \sum [(\lvert f_o - f_e\rvert - 0.5)^2 \div f_e]$
$\chi^2 = [N(AD\text{-}BC)^2] / [(A+B)(C+D)(A+C)(B+D)]$	$\chi^2 = [N(\lvert AD\text{-}BC\rvert - N/2)^2] / [(A+B)(C+D)(A+C)(B+D)]$

Sum 7: The table below shows the number of normal and neurotic who chose each of the two possible answers to an item on a neurosis questionnaire as shown on the table below. Does this item differentiate between the two groups? Test the null hypothesis.

	Yes	No	Total (f_r)
Normal	10 (A)	4 (B)	14
Neurotic	35 (C)	21 (D)	56
Total (f_c)	45	25	70 (N)

$\chi^2 = N(\lvert AD\text{-}BC\rvert\text{-}N/2)^2 \div [(A+B)(C+D)(A+C)(B+D)]$ [Here you have to provide the reason why it is divided by N/2] It is better you change the value of B and write more than 5 and again compute it.

$$= [70(\lvert 10\times21 - 4\times35\rvert - 70/2)^2] / (14\times56\times45\times25)$$

$$= [70(70\text{-}35)^2] / 882000$$

$$= 85750/882000 = 0.097$$

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

Tabulated value of χ^2 at 0.05 level = 3.84

Tabulated value of χ^2 at 0.01 level = 6.64

The calculated value of χ^2 is 0.097 which is smaller than the tabulated value of χ^2 at 0.05 level of significance. Thus, the null hypothesis is accepted. The item does not significantly differentiate between the two groups.

Sum 8: In a study, male and female students were asked to express their preference for three different methods of teaching as shown below. Is preference for a certain type of teaching method independent of the sex of students?

	Lecture Method	Discussion method	Demonstration method	Total (f_c)
Males	3 (7.8)	9 (13.6)	25 (15.6)	37
Females	17 (12.2)	26 (21.4)	15 (24.4)	58
Total (f_r)	20	35	40	95 (N)

Remarks: If any expected frequency is less than 10 or 5 in r x c contingency table then we use Yates correction. The following formula is used.

$$\chi^2_{Yates} = \sum^{k} \frac{(|f_0 - f_e| - 0.5)^2}{f_e}$$

Computation of expected frequencies:

$f_e = (f_r \times f_c)/n$

(20×37)/95 = 7.8

(35×37)/95 = 13.6

(40×37)/95 = 15.6

(20×58)/95 = 12.2

(35×58)/95 = 21.4

(40×58)/95 = 24.4

Computation of χ^2:

$\chi^2 = \sum[(|f_o-f_e|-0.5)^2 / f_e]$

f_o	f_e	f_o-f_e	$(f_o-f_e)-0.5$	$[(f_o-f_e)-0.5]^2$	$[\|f_o-f_e\|-0.5]^2 / f_e$
3	7.8	4.8	4.3	18.49	2.37
9	13.6	4.6	4.1	16.81	1.24
25	15.6	9.4	8.9	79.21	5.08
17	12.2	4.8	4.3	18.49	1.52
26	21.4	4.6	4.1	16.81	0.79
15	24.4	9.4	8.9	79.21	3.25
					$\chi^2 = 14.25$

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

Tabulated value of χ^2 at 0,05 level = 5.99

Tabulated value of χ^2 at 0.01 level = 9.21

The calculated value of χ^2 is 14.25 which is greater than the tabulated value of χ^2 at 0.01 level of significance. Thus, the null hypothesis is rejected. There is a statistically significant difference in the preference of teaching method between males and females.

Application of Chi Square

- Chi square is used as a test of significance when data are expressed in terms of frequencies.
- Normality of distribution is not assumed for computing chi square. Hence, it is a non-parametric or distribution-free statistic.
- It is used as a test of goodness of fit, i.e., to determine the extent to which the experimental results differ from theoretical assumptions. For instance, if a coin is tossed 10 times, theoretically it is expected that 5 out of 10 times the head will occur and the remaining 5 times tail will occur. However, this may not always happen in reality. A null hypothesis may be formulated stating that there is no significant difference between expected and observed frequencies and chi square may be used to test the hypothesis.
- The sum of expected frequencies should be equal to that of the observed frequencies.
- The individual observations should be independent of each other.
- It is formulated to compute data composed of qualitative categories.
- It measures the difference between expectations and actual observed data.
- Yates correction may be used when frequency in one or more cells is small in a 2×2 table.

Exercise

1. What so you understand by chi square test?
2. What is equal probability hypothesis?
3. What do you understand by Yates' correction? Under what circumstance it is applied? State the formula.
4. In a study of food preference, 60 students of a college were asked to select the cuisine they liked best, out of three cuisines – Indian, Chinese and Italian. The obtained data is given below:

Indian	Chinese	Italian
28	18	14

Based on the above data, test the hypothesis that the observed choices do not differ from a random selection.

5. 50 guardians were asked to express their opinions on a five-point scale to the statement – "Children should not be allowed to watch T.V. for more than 30 minutes a day". The obtained data is as follows:

Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
14	16	6	10	4

Do these responses differ significantly from the distribution from the distribution to be expected by chance? [Change first and second example as each row and column must be more than or equal to two]

6. The responses of parents to the question "Should mobile phones be banned in school" are as follows:

	Yes	No	Total
Father	35	25	60
Mother	80	20	100
Total	115	45	160

Test the significance of the responses at 0.05 level of significance.

7. The opinion of three groups of students on a statement were recorded as follows:

	Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree	Total

Science	14	16	5	11	4	50
Humanities	18	24	10	20	13	85
Commerce	20	22	6	12	10	70
Total	52	62	21	43	27	205

Do the data indicate that the opinions expressed are independent of the stream chosen by students?

8. A group of office employees were asked to respond to the statement – “8 hours work schedule per day should be implemented” in terms of two response categories - Agree and Disagree. The obtained responses were as follows:

	Agree	Disagree	Total
Males	14	6	20
Females	16	4	20
Total	30	10	40

Is the gender and opinion expressed independent of each other at the 1 per cent level of significance?

Table: Critical values of χ^2

Levels of significance

df	0.05	0.01
1	3.841	6.635
2	5.991	9.210
3	7.815	11.345
4	9.488	13.277
5	11.070	15.086
6	12.592	16.812
7	14.067	18.475
8	15.507	20.090
9	16.919	21.666
10	18.307	23.209
11	19.675	24.725
12	21.026	26.217
13	22.362	27.688
14	23.685	29.141
15	24.996	30.578
16	26.296	32.000
17	27.587	33.409

18	28.869	34.805
19	30.144	36.191
20	31.410	37.566
21	32.671	38.932
22	33.924	40.289
23	35.172	41.638
24	36.415	42.980
25	37.652	44.314
26	38.885	45.642
27	40.113	46.963
28	41.337	48.278
29	42.557	49.588
30	43.773	50.892

Chapter 8

NORMAL PROBABILITY CURVE

The term normal refers to average. It is intriguing to note that virtually every attribute human beings possess is average. For instance, most human beings are average in terms of height, beauty, honesty, intelligence, etc. However, some people belong to the extremes, either on the positive or negative sides in terms of the attributes. But a general trend is that "farther away a characteristic is from the mean value, the lesser the frequency of occurrence of that particular characteristic". For example, we rarely encounter adult individuals who are more than 6.5 feet tall or those below 4 feet. Likewise, in a classroom, most of the students are average in terms of intelligence and only a few deviate noticeably from the average and they may be either gifted or slow learners. If we plot such a distribution on a graph paper, we get a inverted U-shaped curve or a bell shaped curve which is called the Normal Probability Curve, where 'normal' refers to average and 'probability' refers to the chances of occurrence. Laplace and Gauss (1777-1855) derived this curve independently. They worked on experimental errors in Physics and Astronomy and found the errors to be distributed normally. In the later part of the 19th century Sir Francis Galton started anthropometric measurements of mental traits to study individual differences and arrived at the conclusion that most of the physical and mental traits conformed to the normal curve.

Another fascinating phenomenon in humans is that it is difficult to know who is the most intelligent person in the world and who is the least intelligent. As we all know, it is not possible to get access to every member of a particular population under study so it becomes increasingly difficult to make such comments. Likewise, the Normal Probability Curve never touches the baseline on either side owing to the fact that there might be someone who is even higher than the highest or lower than the lowest in terms of intelligence, beauty, height, etc. This particular characteristic of the Normal Probability Curve is called an 'asymptote', i.e., the curve approaches but never touches the baseline. The curve has its maximum height at the mean of the distribution, its value being 0.3989 in a unit normal curve.

The three measures of central tendency - mean, median and mode lie at the same point for a perfectly normal distribution, they coincide in it and are equal. To find out the deviations from the point of mean, standard deviation of the distribution (σ) is used as a unit of measurement. Since the curve never touches the baseline, it extends theoretically from minus infinity to plus infinity. However, for practical purposes the area covered by the curve is considered to be -3σ to +3σ. The total area under the curve extending from -3σ to +3σ is arbitrarily taken as 10,000 for ease of computation of the fractional parts of the total area found for the mean and the height of the curve at various σ points. It is to be noted that 3413 cases out of 10,000 or 34.13% of the area of the curve lies between -1σ and mean. Similarly, another 34.13% cases lie between the mean and 1σ thus making the total percentage of cases to be 68.26% (34.13+34.13) from -1σ to +1σ distance. Going further, 13.59% cases lie between -2σ and -1σ on one side and 1σ to

2σ on the other thus making the total area percentage of cases to be 95.44% (68.26+27.18) from -2σ to 2σ distance. From -3σ to -2σ there lies 2.15% of the cases on one side and the same from 2σ to 3σ thus making the total percentage of cases to be 99.74% (95.44+4.3). Thus there remains only 100-99.74, i.e., 0.26% cases which lie beyond 3σ on either side (i.e., 0.13% on each side).

Mathematical Formula for Height of a Normal Distribution : The height (ordinate) of a normal curve is defined as

Y =

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

where, X = any value along the X-axis (Any value of the variable X)

Y = the height (ordinate) of a normal curve

μ = mean of the distribution

σ = standard deviation of the distribution

π = 3.1416 = the ratio of the circumference of a circle to its diameter

e = 2.7183 = the base of the Napierian system of logarithms or natural logarithms.

In distributions where the scores deviate markedly from the average, then this divergence from normality may be categorized into:

1. Skewness
2. Kurtosis

In skewness, the curve lacks symmetry. Based on this, the curve may be of two types – positively skewed in which the curve inclines more to the right and negatively skewed in which the curve inclines more to the left. So, the values of the mean, median and mode are different. Skewness may be computed by using the following formula:

Sk = {3(Mean-Median)/Standard Deviation} (a measure of skewness in a frequency distribution)

Sk = {(P90+P10)/2} – P50 (a measure of skewness in terms of percentiles)

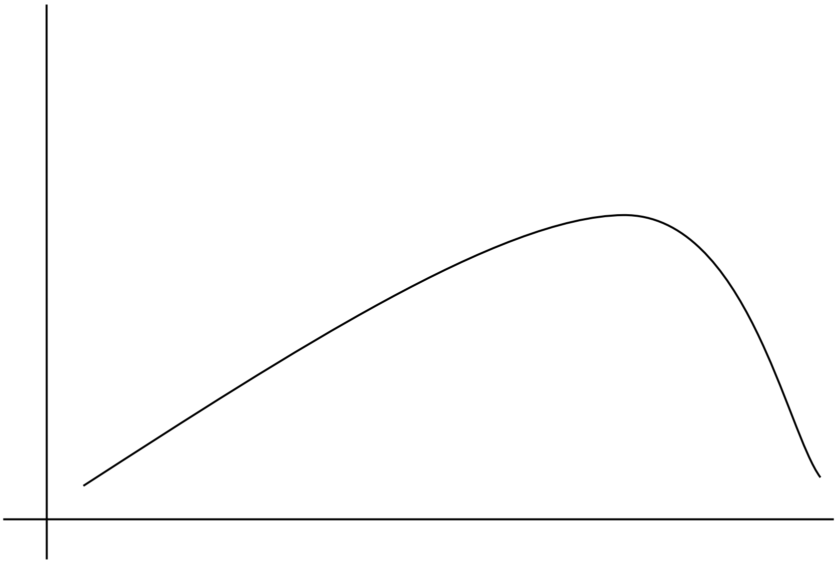


Fig 1: Negative skewness

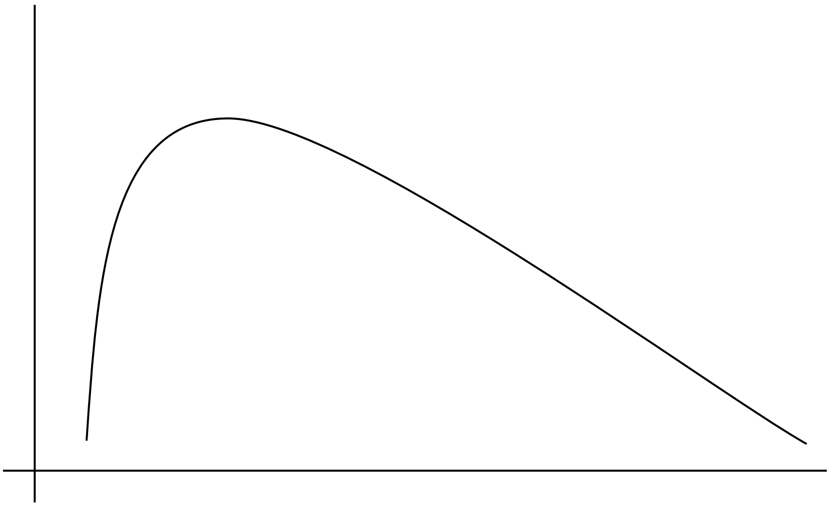


Fig 2: Positive skewness

In Kurtosis, the curve is either more peaked or less peaked than the normal. Based on this, the curve may be of three types – Platykurtic when the curve is flatter than the normal, Leptokurtic when the curve is more peaked than the normal and when a distribution resembles or nearly resembles a normal curve then it is called mesokurtic. Kurtosis may be computed by using the following formula:

Kurotsis = {Quartile Deviation/(90th Percentile-10th Percentile)}

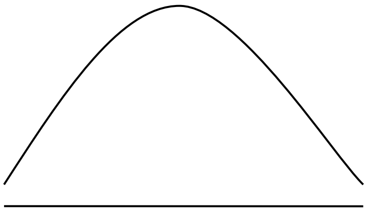


Fig 3: Mesokurtic (Normal curve)

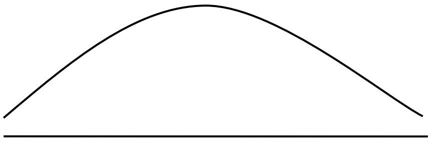


Fig 4: Platykurtic

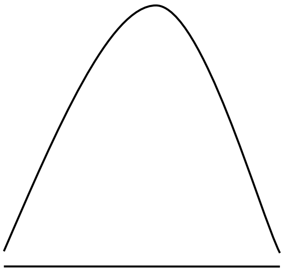


Fig 5: Leptokurtic

Sum 1*:* The given mean for a distribution is 64.5 and S.D. is 15.7. Change the score of 75 into a z or sigma (σ) score.

z = (X-M)/S.D.

= (75-64.5)/15.7 = 0.67σ

Sum 2*:* The given mean for a distribution is 56 and S.D. is 9.6. Convert a z score (σ score) of value 1.25 into a raw score.

z = (X-M)/S.D.

1.25 = (X-56)/9.6

X-56 = 9.6×1.25

X = 12+56 = 68

Sum 3*:* The marks obtained by a student in Physics and Chemistry are 55 and 63 respectively. If the mean and S.D. for the scores in Physics are 48 and 12.5 and that of Chemistry are 60 and 10.3 respectively, then find out in which subject the student did better.

Solution: From the given data, direct comparison on the basis of raw scores cannot be made owing to the fact psychological variables rely on interval scales of measurement and that the measures of central tendency and variability are different for the two subjects. Thus the raw scores are converted into a standard score (σ score) with the help of the NPC table.

Raw score in Physics=55, Mean=48, S.D.=12.5

z = (X-M)/S.D.

= (55-48)/12.5 = 0.56σ

Raw score in Chemistry = 63, Mean=60, S.D.=10.3

z = (X-M)/S.D.

= (63-60)/10.3 = 0.29σ

Thus, it may be concluded that the student did better in Physics than in Chemistry because the z score in Physics is more than z score in Chemistry.

Sum 4*:* In a sample of 5000 cases, the mean and S.D of test scores are 18.2 and 3.9 respectively. Assuming normality of distribution find out how many individuals scored between 15 and 20.

Mean (M) = 18.2   S.D = 3.9

Raw score 1 (X_1) = 15

Raw score 2 (X_2) = 20

for X_1 , z = (X_1-M)/S.D

= (15-18.2)/3.9

= -0.82σ

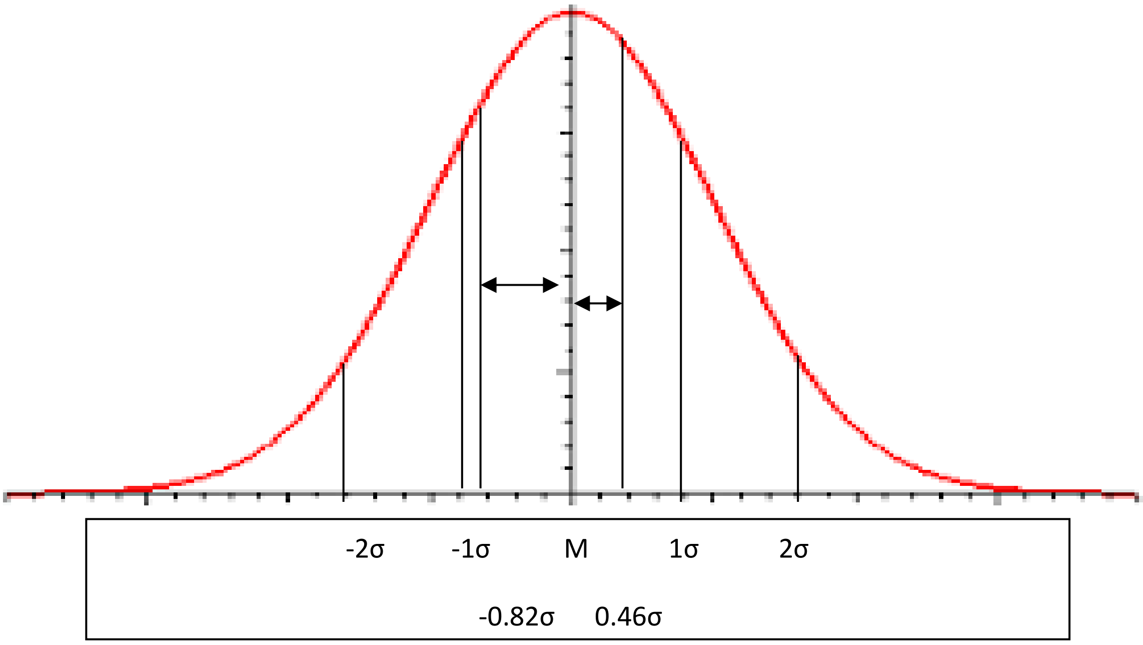
From the normal curve table it may be seen that 2939 cases (out of 10,000), i.e., 29.39% cases lie between -0.82σ and the mean.

Similarly for $z = (X_2\text{-}M)/S.D$

$= (20\text{-}18.2)/3.9$

$= 0.46\sigma$

That is, 0.46σ corresponds to 1772 cases which means that 17.72% cases lie between mean and 0.46σ.



From NPC Table,

0.82σ corresponds to 2939 cases and 0.46σ corresponds to 1772 cases.

Therefore, Total number of cases lying between these two points = 2939 + 1772

= 4711 individuals

That is, 4711/10,000 × 100%, i.e., 47.11%

Therefore, number of individuals who scored between 15 and 20 out of 5000 individuals = (47.11/100) × 5000

= 2355

Hence 2355 individuals scored between 15 and 20.

Sum 5*:* If a distribution is normal with M = 100 and S.D = 15, find out the two points between which the middle 40% of the cases lie.

The middle 40% of the cases are distributed in such a way that 20% (or 2000 out of 10,000) of the cases lie to the left and 20% to the right of the mean.

From the normal curve (Table A, Garrett), the corresponding σ distance for 2000 fractional parts of the total area under the normal curve is to be found out. The nearest number to 2000 in the table is 1985 for which the sigma value is 0.52σ. It means that 20% of the cases lie on the right side of the curve between M and 0.52σ and 20% of the cases lie on the left side of the figure between M and -0.52σ.

The standard z scores are to be converted into raw scores. Let the two raw scores be X_1 and X_2. Now z is computed as

$z = (X_1-M)/S.D$ and $z = (X_2-M)/S.D$

$0.52 = (X_1-100)/15$ $-0.52 = (X_2-100)/15$

$X_1-100 = 7.8$ $X_2-100 = -7.8$

$X_1 = 100 + 7.8$ $X_2 = 100 – 7.8$

$X_1 = 107.8$ $X_2 = 92.2$

≈ 108 ≈ 92

Thus, the middle 40% of the cases lie between the scores 92 and 108.

Sum 6*:* In a normal distribution where N= 1000, Mean = 60 and S.D = 12, find the percentile rank of the individual scoring 70.

In order to find out the percentile rank, we have to determine the percentage of cases lying below the score point 70.

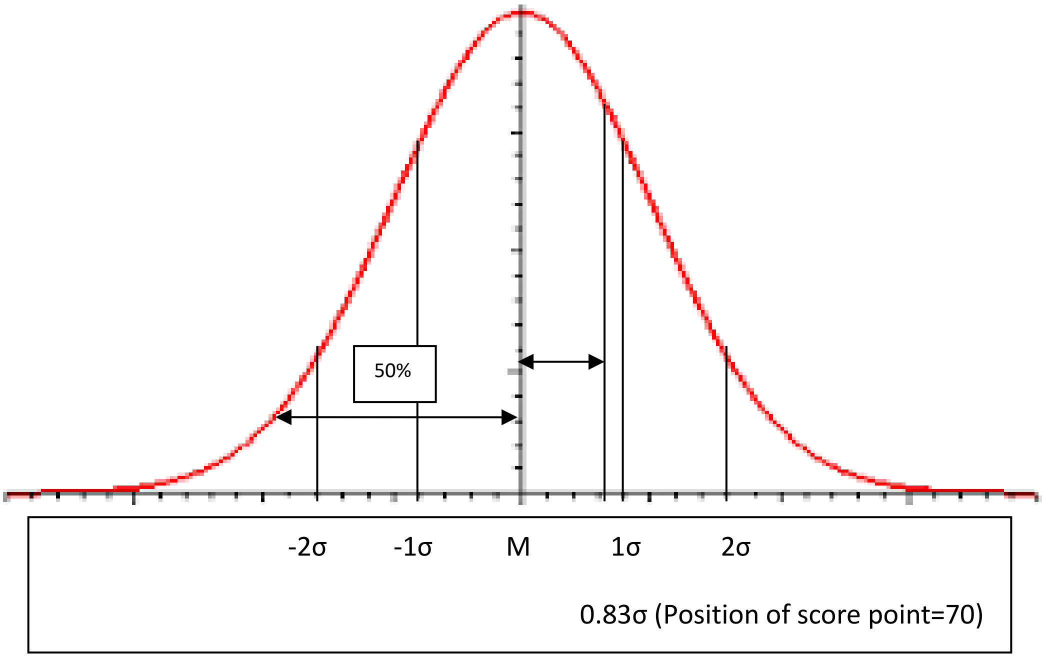
$z = (X-M)/S.D$

$= (70-60)/12$

$= 0.83\sigma$

Hence 0.83σ corresponds to 2967 cases

Thus it can be said that 29.67% cases lie between M and 0.83σ distance. But 50% cases lie up to the mean on the left side of the curve. Therefore, it may be concluded that there are 50+29.67 = 79.67% of the individuals whose scores lie below the score point 70. Thus the percentile rank of the individual scoring 70 is 80. We have also shown this solution using below Normal curve.



Sum 7*:* In a sample of 1000 cases, the mean and S. D. of the distribution is 60 and 8 respectively. Assuming normality of the distribution, find out how many individuals score are above 65 score point.

z = (X-M)/S.D

= (65-60)/8

= 0.625σ

Hence 0.625σ is halfway between 0.62σ and 0.63σ. Therefore, the area between mean and 0.625σ is interpolated with the help of the normal curve table as follows:

2324+{(2357-2324)/2} = 2324+16.5 = 2340.5 or 2341

Thus it can be said that 23.41% cases lie between the mean and 0.625σ distance.

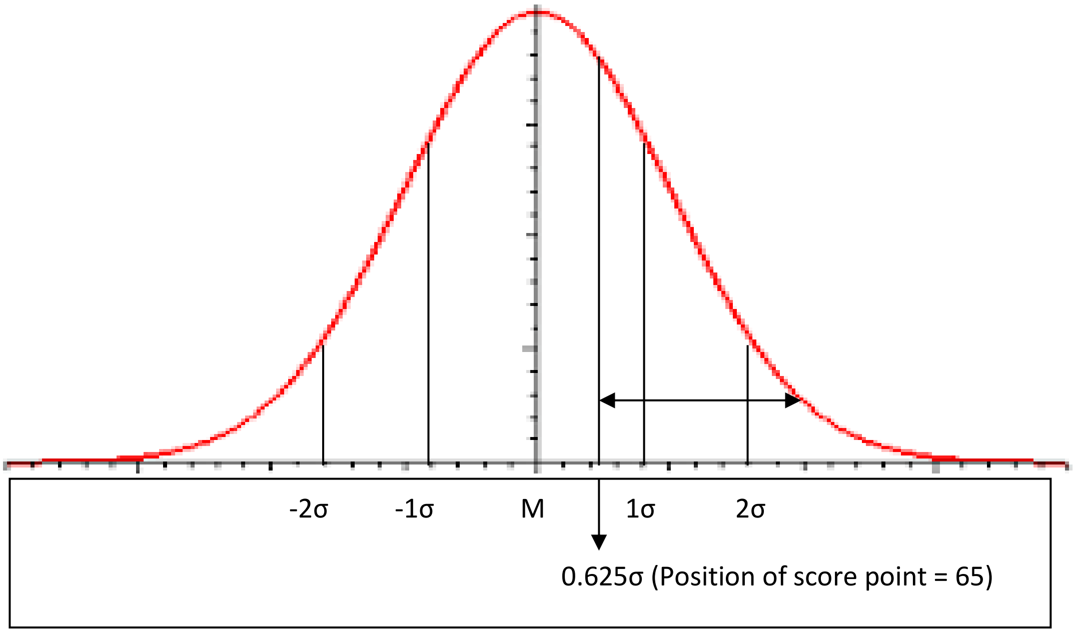
It is also known that a total of 50% cases lie on each side of the mean. Therefore,

50 - 23.41 = 26.59%

That is, (26.59/100) × 1000

= 265.9 or 266

Therefore, 266 individuals out of 1000 score are above the 65 score point. We have also shown this solution using below Normal curve.



Sum 8*:* In a sample of 500 cases, the mean and S.D. of the distribution is 80 and 13 respectively. Assuming normality of the distribution, find out how many individuals score are below the score point 70.

z = (X-M)/S.D

= (70-80)/13

= -0.77

So 0.77σ corresponds to 2794 cases.

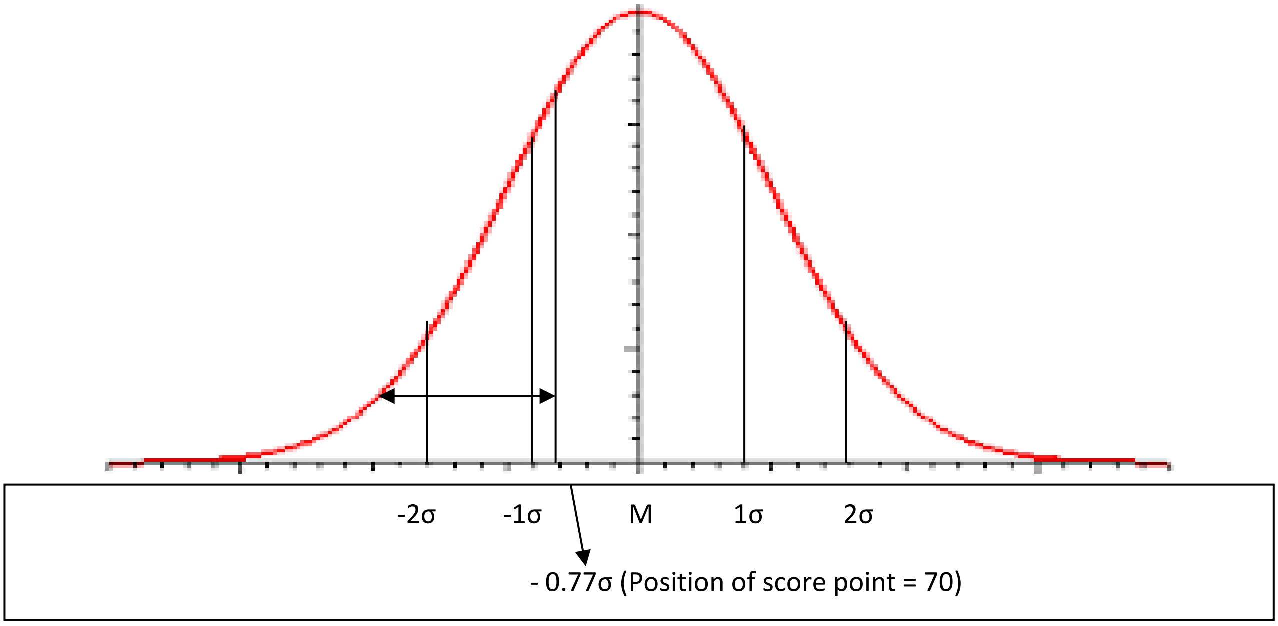
From the normal curve table (Garrett Table A), it is found that 2794 out of 10,000 cases or 27.94% cases lie between -0.77σ and the mean. It is also known that a total of 50% cases lie on either side of the mean. Therefore,

50 - 27.94 = 22.06%

That is (22.06/100) × 500

= 110.3 or 110

Hence 110 individuals out of 500 score are below the score point 70. We have also shown this solution using below Normal curve.



Applications of Normal Probability Curve

The normal probability curve can be used

- To convert raw scores into standard scores (z scores)
- To determine the standard error of measurement
- To compare various distributions. For instance, whether a set of observations is normally distributed or is it skewed.
- To find out percentiles and percentile ranks in a given distribution.
- For grouping individuals or scores into certain specified categories according to their traits. For example, individuals can be categorized as Low Average, Average, High based on IQ scores.
- To determine relative difficulty of test items

Exercise

1. What do you understand by the normal probability curve?
2. What are the characteristics of a normal curve?
3. Write down height (ordinate) of a normal curve.
4. What do you mean by skewness and kurtosis? Give examples.
5. What percentage of a normal distribution is included between the
   i. Mean and 1.14σ
   ii. -0.82σ and Mean
   iii. -1.56σ and 2.07σ
6. In a normal distribution, determine P_{30}, P_{44} and P_{60} in σ units.
7. The following data shows the Mean and Standard Deviation of a group of students and achievement scores of a student:

	History	Political Science
Mean	40	60
Standard Deviation	7	8
Achievement Scores of a student	49	64

   Find out whether the student did better in History or Political Science.

8. In a normal distribution with a mean of 60 and Standard Deviation of 15, find out
   i. What per cent of the cases lie between the scores 50 and 65
   ii. What per cent of the group is expected to have scores greater than 66
9. In a normal distribution with N=100, Mean=30 and S.D.=5, find out
   i. What percent of cases lie between the scores 24-28
   ii. What limits include the middle 50%?
10. Assuming normality, a test has a mean score of 100 and standard deviation of 16. Compute
   i. The score that cuts off the top 15%
   ii. Score that cuts off the lower 30%
   iii. Percentage of cases above 105
   iv. Score limits of the middle 50%

Table: Areas of the Normal Curve (fractional parts of the total area taken as 10,000) and Critical Values of z (i.e. x/σ)

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0.1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0.2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0.3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0.4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0.5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2357	2389	2422	2454	2486	2517	2549
0.7	2580	2611	2642	2673	2704	2734	2764	2794	2823	2852

0.8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0.9	3159	3186	3212	3238	3264	3290	3315	3340	3365	3389
1.0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1.1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1.2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1.3	4032	4049	4066	4082	4099	4115	4131	4147	4162	4177
1.4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1.5	4332	4345	4357	4370	4383	4394	4406	4418	4429	4441
1.6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1.7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1.8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1.9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2.0	4772	4778	4783	4788	4793	4798	4803	4808	4812	4817
2.1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2.2	4861	4864	4868	4871	4875	4878	4881	4884	4887	4890
2.3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2.4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2.5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2.6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2.7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2.8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2.9	4981	4982	4982	4983	4984	4984	4985	4985	4986	4986
3.0	4986.5	4986.9	4987.4	4987.8	4988.2	4988.6	4988.9	4989.3	4989.7	4990.0
3.1	4990.3	4990.6	4991.0	4991.3	4991.6	4991.8	4992.1	4992.4	4992.6	4992.9
3.2	4993.129									
3.3	4995.166									
3.4	4996.631									
3.5	4997.674									
3.6	4998.409									
3.7	4998.922									
3.8	4999.277									
3.9	4999.519									
4.0	4999.683									
4.5	4999.966									
5.0	4999.997133									

ANSWERS

Chapter 3

5) Mean = 41.67

6) i) Median = 16

ii) Median = 5

7) Mode = 14

8) i) Mean = 40.33, Median = 40.82, Mode = 41.80

ii) Mean = 73.06, Median = 72.94, Mode = 72.71

9) Mode = 42.23

Chapter 4

3) i) $Q_1 = 65.61$, $Q_3 = 89.5$, $Q = 11.95$

ii) $Q_1 = 37.56$, $Q_3 = 53.25$, $Q = 7.85$

4) A.D. = 3.00, S.D. = 3.59

5) A.D. = 5.4, S.D. = 6.32

6) i) S.D. = 4.36

ii) S.D. = 2.087

Chapter 5

3. i) r = 0.68 ii) r = 0.196

4. i) $\rho = -0.55$ ii) $\rho = 0.864$

5. r bis = 0.055

6. $r_p$bis = -0.083

7. i) 0.45 ii) 0.42

8. i) 0.50 ii) 0.22

9. $\chi^2 = 59.93$ (Significant at 0.01 level of significance)

Chapter 6

2. 12.89 (Significant at 0.01 level of significance)

3. 0.993 (Not significant)

4. 10.48 (Significant at 0.01 level of significance)

5. 7.47 (Significant at 0.01 level of significance)

6. 4.036 (Significant at 0.01 level of significance)

7. 10.10 (Significant at 0.01 level of significance)

Chapter 7

4. 5.2 (Not significant)

5. 10.4 (Significant at 0.05 level of significance)

6. 8.71 (Significant at 0.01 level of significance)

7. 3.73 (Not significant)

8. 0.13 (Not significant)

Chapter 8

5. i) 37.29%

ii) 29.39%

iii) 92.14%

6. -0.52σ, $-0.15\ \sigma$, $0.25\ \sigma$

7. The student performed better in History.

8. i) 37.79%

ii) 34.46%

9. i) 22.95%

ii) 27 and 33

10. i) 117

ii) 92

iii) 37.83%

iv) 89 and 111

REFERENCES

Aron, A., Coups, E. J., Aron, E. N. (2013) *Statistics for Psychology* (6th Ed.). United States of America, Pearson Education , Inc.

Dancey, C. P., & Reidy, J. (2007). *Statistics without maths for psychology*. Pearson Education.

Das, D., Das, A. (2005) *Statistics in Biology and Psychology* (4th Ed.). Kolkata, Academic Publishers.

Garrett, H. E. (2009) *Statistics in Psychology and Education*. New Delhi, Paragon International Publishers.

Gaur, A. S., Gaur, S. S. (2009), *Statistical Methods for Practice and Research, A guide to data analysis using SPSS* (2nd Ed). Response, Business Books from SAGE

Gravetter, F. J., Wallnau, L. B. (2017). *Statistics for the Behavioural Sciences* (10th Ed). Cengage Learning.

Hanneman, R. A., Kposowa, A. J., Riddle (2013), M. D., *Basic Statistics for Social Research*. Jossey-Bass, A Wiley imprint.

Heiman, G. W. (2011) *Basic Statistics for the Behavioral Scienecs* (6th Ed.). Wadsworth, Cengage Learning.

King, B. M., Rosopa, P. J., Minium, E. W. (2011) *Statistical Reasoning in the Behavioural Sciences* (6th Ed.). John Wiley and Sons, Inc.

Mangal, S. K. (2012). *Statistics in Psychology and Education* (2nd Ed.). New Delhi, PHI Learning Private Limited.

Nolan, S. A., Heinzen, T. E. (2012). *Statistics for the Behavioral Sciences* (2nd Ed.). Worth Publishers

Ramamurti, P.V. (2014) *An Introduction to Psychological Measurements*. Delhi, PHI Learning Private Limited.

Singh, A. K. (2010) *Tests, Measurements and Research Methods in Behavioural Sciences*. New Delhi, Bharati Bhavan (Publishers & Distributors)

Stockemer, D. (2018). *Quantitative Methods for the Social Sciences: A Practical Introduction with Examples in SPSS and Stata*. Springer.

https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/8-chi-squared-tests

ISBN 978-81-949922-8-8

9 788194 992288 >

Printed and Published By
GLOBAL LEARNERS ACADEMY OF DEVELOPMENT ,
26A.Kasthuri Nagar, Korattur, Chennai, Pin - 600 080, India,
www.gladnetwork.in,   Email - editor@gladnetwork.in