



B.Sc. STATISTICS - I YEAR
DJS1B : DESCRIPTIVE STATISTICS
SYLLABUS

Unit - I

Origin, scope, limitations and misuses of Statistics – Collection - Classification- Tabulation of data. Frequency Distributions – Nominal, ordinal, Interval and ratio. Diagrammatic presentation of data: one dimensional and two-dimensional diagrams – graphic representation: line diagram, frequency polygon, frequency curve, histogram and Ogive curves.

Unit - II

Measures of central tendency: mean, median, mode, geometric mean and harmonic mean. Partition values: Quartiles, Deciles and Percentiles. Measures of Dispersion: Mean deviation, Quartile deviation and Standard deviation – Coefficient of variation.

Unit - III

Moments - measures of Skewness - Pearson's and Bowley's Coefficients of skewness, Coefficient of Skewness based on moments – co-efficient of Kurtosis.

Unit - IV

Curve fitting: principle of least squares, fitting of the curves of the form $y = a+bx$, $y = a+bx+cx^2$ and Exponential and Growth curves.

Unit - V

Linear correlation - scatter diagram, Pearson's coefficient of correlation, computation of co-efficient of correlation from a bivariate frequency distribution, Rank correlation, Coefficient of concurrent deviation- Regression equations - properties of regression coefficients.

REFERENCE BOOKS::

1. Anderson, T.W. and Sclove, S.L. (1978) Introduction to Statistical Analysis of data, Houghton Mifflin, Boston.
2. Bhat, B.R., Srivenkataramna, T. and Madhava Rao, K.S. (1996) Statistics A Beginner's Text, Vol. I, New Age International, New Delhi.
3. Croxton, F.E. and Cowden, D.J. (1969) Applied General Statistics, Prentice Hall, New Delhi.
4. Goon, A.M., M.K. Gupta and B. Das Gupta (2002) Fundamentals of Statistics- Vol. I., World Press Ltd, Kolkata.
5. Gupta, S.C. and V.K. Kapoor (2002) Fundamentals of Mathematical Statistics, Sultan Chand & Sons, New Delhi.
6. Spiegel, M.R. and Stephens, L. (2010) Statistics, Schaum's Outline Series, Mc Graw Hill, New York.



CONTENTS

No.	Title	Page No.
	Unit – I	
1.1	Introduction	4
1.2	Origin	7
1.3	Functions of statistics	9
1.4	Scope of statistics	11
1.5	Limitations and Misuses of Statistics	14
1.6	Collection of data	16
1.7	Classification	23
1.8	Tabulation	27
1.9	Frequency Distributions	29
1.10	Diagrammatic and graphical representation	36
	Questions	46
	Unit – II	
2.1	Measures of Central Tendency	47
	Mean	49
	Median	54
	Mode	63
	Geometric mean	71
	Harmonic mean	72
2.2	Partition values	75
	Quartiles	75
	Deciles	80
	Percentiles	82
2.3	Measures of Dispersion	85
	Range	86
	Mean deviation	88
	Quartile deviation	96
	Standard deviation	101
	Coefficient of variation	106
	Questions	110



	Unit – III	
3.1	Moments	114
3.2	Measures of Skewness	119
3.3	Pearson’s and Bowley’s Coefficients of skewness	119
3.4	Measures (or) Coefficient of Skewness based on moments	128
3.5	Kurtosis (or) Co-efficient of Kurtosis	128
	Questions	130
	Unit - IV	
4.1	Curve fitting	133
4.2	Principle of least squares	133
4.3	Fitting of the curves of the form $y = a+bx$,	135
4.4	Fitting of the curves of the form $y = a+bx+cx^2$	141
4.5	Exponential and Growth curves	143
	Questions	144
	Unit - V	
5.1	Linear correlation	145
5.2	Scatter diagram	147
5.3	Pearson’s coefficient of correlation	153
5.4	Computation of co-efficient of correlation from a bivariate frequency distribution	157
5.5	Rank correlation	161
5.6	Coefficient of concurrent deviation	165
5.7	Regression equations	167
5.8	Regression Coefficients and properties of regression coefficients	171
	Questions	176



UNIT – I

1.1 INTRODUCTION

Statistics has been defined differently by different statisticians from time to time. These definitions emphasize precisely the meaning, scope and limitations of the subject. The reasons for such a variety definitions may be stated as follows:

- The field of utility of statistics has been increasing steadily.
- The word statistics has been used to give different meaning in singular (the science of statistical methods) and plural (numerical set of data) sense.

Definitions:

Webster:

“Statistics are the classified facts representing the conditions of the people in a State... specially those facts which can be stated in number or in tables of numbers or in any tabular or classified arrangement”.

A.L. Bowley:

- i. The science of counting
- ii. The science of averages
- iii. The science of measurements of social phenomena, regarded as a whole in all its manifestations.
- iv. A subject not confined to any one science.

Yule and Kendall:

“By Statistics we mean quantitative data affected to a marked extent by multiplicity of causes”

A.M. Tuttle:

“Statistics are measurements, enumerations or estimates of natural phenomenon, usually systematically arranged, analysed and presented as to exhibit important inter-relationships among them”.

**Prof. Horace Secrist:**

“Statistics may be defined as the aggregate of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other”.

Croxton and Cowden:

Statistics may be defined as the science of collection, presentation analysis and interpretation of numerical data from the logical analysis. It is clear that the definition of statistics by Croxton and Cowden is the most scientific and realistic one. According to this definition there are four stages:

Collection of Data:

It is the first step and this is the foundation upon which the entire data set. Careful planning is essential before collecting the data. There are different methods of collection of data such as census, sampling, primary, secondary, etc., and the investigator should make use of correct method.

Presentation of data:

The mass data collected should be presented in a suitable, concise form for further analysis. The collected data may be presented in the form of tabular or diagrammatic or graphic form.

Analysis of data:

The data presented should be carefully analysed for making inference from the presented data such as measures of central tendencies, dispersion, correlation, regression etc.,

Interpretation of data:

The final step is drawing conclusion from the data collected. A valid conclusion must be drawn on the basis of analysis. A high degree of skill and experience is necessary for the interpretation.



Every day, we come across the different types of quantitative information in newspapers, magazines, over radio and television. For example, we may hear or read that population of India had increased at the rate of 2.5% per year (per annum) during the period 1981-1991, number of admission in National Open School had gone up by say 20% during 1996-97 as compared to 1995-96 etc. We would like to know that what these figures mean. These quantitative information or expression called statistical data or statistics.

Statistics is concerned with scientific methods for collecting, organising, summarising, presenting and analysing data as well as deriving valid conclusions and making reasonable decisions on the basis of this analysis. Statistics is concerned with the systematic collection of numerical data and its interpretation. The word 'statistic' is used to refer to

1. Numerical facts, such as the number of people living in particular area.
2. The study of ways of collecting, analysing and interpreting the facts.

1.2 ORIGIN

The science of statistics developed gradually and its field of application widened day by day. The subject of Statistics, as it seems, is not a new discipline but it is as old as the human society itself. Its origin can be traced to the old days when it was regarded as the 'Science of State-craft' and was the by-product of the administrative activity of the State. The word 'Statistics' seems to have been derived from the Latin word 'status' or the Italian word 'statista' or the German word 'statistik' each of which means a 'political state'.

In ancient times, the government used to collect the information regarding the population and 'property or wealth' of the country-the former enabling the government to have an idea of the man-power of the country, and the latter providing it a basis for introducing new taxes and levies.

In India, an efficient system of collecting official and administrative statistics existed even more than 2,000 years ago, in particular, during the reign of Chandra Gupta Maurya (324-300 B.C). From Kautilya's



Arthshastra it is known that even before 300 B.C. a very good system of collecting 'Vital Statistics' and registration of births and deaths was in vogue. During Akbar's reign (1556-1605 A. D.), Raja, Todarmal, the then land and revenue minister, maintained good records of land and agricultural statistics.

In Alina-e-Akbari written by Abul Fazl (in 1596-97), one of the nine gems of Akbar, we find detailed accounts of the administrative and statistical surveys conducted during Akbar's reign. In Germany, the systematic collection of official statistics originated towards the end of the 18th century when, in order to have an idea of the relative strength of different German States, information regarding population and output-industrial and agricultural was collected.

In England, statistics were the outcome of Napoleonic wars. The wars necessitated the systematic collection of numerical data to enable the government to assess the revenues and expenditure with greater precision and then to levy new taxes in order to meet the cost of war. Seventeenth century saw the origin of the 'Vital Statistics'. Captain John Grant of London (1660-1674), known as the 'father' of Vital Statistics, was the first man to study the statistics of births and deaths.

The theoretical development of the so-called, modern statistics came during the mid-seventeenth century with the introduction of '*Theory of Probability*' and '*Theory of Games and Chance*'. The chief contributors being mathematicians and gamblers of France, Germany and England. Statistics is an old science, originated during the time of Mahabharat. For the last few centuries, it has remained a part of mathematics like Pascal (1623-1662), James Bernoulli (1654-1705), De Moivre (1667-1754), Laplace (1749-1827), Gauss (1777-1855), Lagrange, Bayes, Markoff, Euler etc. These mathematicians were mainly interested in the development of the theory of probability as applied to the theory of games and other chance phenomena. Till the early nineteenth century, statistics was mainly concerned population and area of land under cultivation, etc., of a state or kingdom.



A Ronald. A. Fisher (1890-1962) who applied statistics to a variety of diversified fields such as genetics, biometry, psychology and education, agriculture, etc and which is rightly termed as the Father of Statistics. His contributors to the subject of Statistics are described in the following words:

‘R.A. Fisher is the real giant in the development of the theory of Statistics’

The varied and outstanding contributions of R.A. fisher put the subject of Statistics on a very firm footing and earned for it the status of fully fledged science.

Indian statisticians have also made notable contributions to the development of Statistics in various diversified fields. The valuable contributions of P.C. Mahalanobis and P.V. Sukhatme (Sample Surveys); R.C. Bose, Panse, J.N. Srivatsva (Design of experiments in Agriculture); S.N. Roy (Multivariate Analysis); C.R. Rao (Statistical Inference); Parthasarathy (Theory of Probability), to mention only a few, have earned for India a high position in the world map of Statistics.

1.3 FUNCTIONS OF STATISTICS

Statistics is viewed not as a mere device for collecting numerical data but as a means of developing sound techniques for their handling and analysis and drawing valid inferences from them.

We now discuss briefly the functions of statistics. Let us consider the following important functions.

It simplifies facts in a definite Form:

Any conclusions stated numerically are definite and hence more convincing than conclusions stated qualitatively. Statistics presents facts in a precise and definite form and thus helps for a proper comprehension of what is stated.

Condensation:

The generally speaking by the word ‘to condense’, we mean to reduce or to lessen. Condensation is mainly applied at embracing the



understanding of a huge mass of data by providing only few observations. If in a particular class in Chennai School, only marks in an examination are given, no purpose will be served. Instead if we are given the average mark in that particular examination, definitely it serves the better purpose. Similarly the range of marks is also another measure of the data. Thus, Statistical measures help to reduce the complexity of the data and consequently to understand any huge mass of data.

It facilitates Comparison:

The various statistical methods facilitate comparison and enable useful conclusions to be drawn. The classification and tabulation are the two methods that are used to condense the data. They help us to compare data collected from different sources. Grand totals, measures of central tendency measures of dispersion, graphs and diagrams, coefficient of correlation etc provide ample scope for comparison. If we have one group of data, we can compare within itself. If the rice production (in Tonnes) in Tanjore district is known, then we can compare one region with another region within the district. Or if the rice production (in Tonnes) of two different districts within Tamilnadu is known, then also a comparative study can be made. As statistics is an aggregate of facts and figures, comparison is always possible and in fact comparison helps us to understand the data in a better way.

It helps in the formation of policies:

Scientific analysis of statistical data constitutes the starting point in all policy making. Decisions relating to import and export of various commodities, production of particular products etc., are all based on statistics.

It helps in Forecasting:

Plans and policies of organizations are invariably formulated well in advance of the time of their implementation. The word forecasting, mean to predict or to estimate beforehand. Given the data of the last fifteen years connected to rainfall of a particular district in Tamilnadu, it is possible to



predict or forecast the rainfall for the near future. In business also forecasting plays a dominant role in connection with production, sales, profits etc. Analysis of time series and regression analysis plays an important role in forecasting.

Estimation:

One of the main objectives of statistics is drawn inference about a population from the analysis for the sample drawn from that population. The four major branches of statistical inference are

- Estimation theory
- Tests of Hypothesis
- Non Parametric tests
- Sequential analysis

In estimation theory, we estimate the unknown value of the population parameter based on the sample observations.

1.4 SCOPE OF STATISTICS

There are many scopes of statistics. Statistics is not a mere device for collecting numerical data, but as a means of developing sound techniques for their handling, analysing and drawing valid inferences from them. Statistics is applied in every sphere of human activity – social as well as physical – like Biology, Commerce, Education, Planning, Business Management, Information Technology, etc. It is almost impossible to find a single department of human activity where statistics cannot be applied. We now discuss briefly the applications of statistics in other disciplines.

Statistics and Industry:

Statistics is widely used in many industries. In industries, control charts are widely used to maintain a certain quality level. In production engineering, to find whether the product is conforming to specifications or not, statistical tools, namely inspection plans, control charts, etc., are of



extreme importance. In inspection plans we have to resort to some kind of sampling – a very important aspect of Statistics.

Statistics and Commerce:

Statistics are lifeblood of successful commerce. Any businessman cannot afford to either by under stocking or having overstock of his goods. In the beginning he estimates the demand for his goods and then takes steps to adjust with his output or purchases. Thus statistics is indispensable in business and commerce. As so many multinational companies have invaded into our Indian economy, the size and volume of business is increasing. On one side the stiff competition is increasing whereas on the other side the tastes are changing and new fashions are emerging. In this connection, market survey plays an important role to exhibit the present conditions and to forecast the likely changes in future.

Statistics and Agriculture:

Analysis of variance (ANOVA) is one of the statistical tools developed by Professor R.A. Fisher, plays a prominent role in agriculture experiments. In tests of significance based on small samples, it can be shown that statistics is adequate to test the significant difference between two sample means. In analysis of variance, we are concerned with the testing of equality of several population means.

For an example, five fertilizers are applied to five plots each of wheat and the yield of wheat on each of the plots are given. In such a situation, we are interested in finding out whether the effect of these fertilisers on the yield is significantly different or not. In other words, whether the samples are drawn from the same normal population or not. The answer to this problem is provided by the technique of ANOVA and it is used to test the homogeneity of several population means.

Statistics and Economics:

Statistics data and techniques of statistical analysis have proved immensely useful in solving a variety of economic problems.



Statistical methods are useful in measuring numerical changes in complex groups and interpreting collective phenomenon. Nowadays the uses of statistics are abundantly made in any economic study. Both in economic theory and practice, statistical methods play an important role.

Alfred Marshall said, “Statistics are the straw only which I like every other economists have to make the bricks”. It may also be noted that statistical data and techniques of statistical tools are immensely useful in solving many economic problems such as wages, prices, production, distribution of income and wealth and so on. Statistical tools like Index numbers, time series Analysis, Estimation theory, Testing Statistical Hypothesis are extensively used in economics.

Statistics and Education:

Statistics is widely used in education. Research has become a common feature in all branches of activities. Statistics is necessary for the formulation of policies to start new course, consideration of facilities available for new courses etc. There are many people engaged in research work to test the past knowledge and evolve new knowledge. These are possible only through statistics.

Statistics and Planning:

Statistics is indispensable in planning. In the modern world, which can be termed as the “world of planning”, almost all the organisations in the government are seeking the help of planning for efficient working, for the formulation of policy decisions and execution of the same. In order to achieve the above goals, the statistical data relating to production, consumption, demand, supply, prices, investments, income expenditure etc and various advanced statistical techniques for processing, analysing and interpreting such complex data are of importance. In India statistics play an important role in planning, commissioning both at the central and state government levels.



Statistics and Medicine:

Statistical tools are widely used in Medical sciences. In order to test the efficiency of a new drug or medicine, t-test is used to compare the efficiency of two drugs or two medicines; t-test for the two samples is used. More and more applications of statistics are at present used in clinical investigation.

1.5 LIMITATIONS AND MISUSES OF STATISTICS

Although statistics is indispensable to almost all sciences: social, physical and natural and very widely used in most of spheres of human activity. It suffers from the following limitations. Statistics with all its wide application in every sphere of human activity has its own limitations. Some of them are given below.

Statistics deals only with aggregate of facts and not with individuals:

Statistics does not give any specific importance to the individual items; in fact it deals with an aggregate of objects. Individual items, when they are taken individually do not constitute any statistical data and do not serve any purpose for any statistical enquiry.

Statistics does not study of qualitative phenomenon:

Since statistics is basically a science and deals with a set of numerical data, it is applicable to the study of only these subjects of enquiry, which can be expressed in terms of quantitative measurements. As a matter of fact, qualitative phenomenon like honesty, poverty, beauty, intelligence etc, cannot be expressed numerically and any statistical analysis cannot be directly applied on these qualitative phenomena. Nevertheless, statistical techniques may be applied indirectly by first reducing the qualitative expressions to accurate quantitative terms. For example, the intelligence of a group of students can be studied on the basis of their marks in a particular examination.



Statistics laws are true only on an average:

It is well known that mathematical and physical sciences are exact. But statistical laws are not exact and statistical laws are only approximations. Statistical conclusions are not universally true. They are true only on an average.

Statistics table may be misused:

Statistics must be used only by experts; otherwise, statistical methods are the most dangerous tools on the hands of the inexperienced. The use of statistical tools by the inexperienced and untrained persons might lead to wrong conclusions. Statistics can be easily misused by quoting wrong figures of data.

Statistics is only, one of the methods of studying a problem:

Statistical method do not provide complete solution of the problems because problems are to be studied taking the background of the countries culture, philosophy or religion into consideration. Thus the statistical study should be supplemented by other evidences.

Statistics is only inappropriate information:

Unskilled, idle and inexperienced person often collect data. As a result, erroneous, puzzling and partial information is collected. As a result, very often improper decision is taken.

Statistics is purposive Misuses:

The most total limitation of statistics is that its purposive misuse. Very often erroneous information may be collected. But sometimes some institutions use statistics for self interest and puzzling other organizations.

1.6 COLLECTION OF DATA

Everybody collects, interprets and uses information, much of it in numerical or statistical forms in day-to-day life. It is a common practice that people receive large quantities of information everyday through conversations, televisions, computers, the radios, newspapers, posters, notices and instructions. It is just because there is so much information



available that people need to be able to absorb, select and reject it. In everyday life, in business and industry, certain statistical information is necessary and it is independent to know where to find it how to collect it. As consequences, everybody has to compare prices and quality before making any decision about what goods to buy. As employees of any firm, people want to compare their salaries and working conditions, promotion opportunities and so on. In time the firms on their part want to control costs and expand their profits.

One of the main functions of statistics is to provide information which will help on making decisions. Statistics provides the type of information by providing a description of the present, a profile of the past and an estimate of the future. The following are some of the objectives of collecting statistical information.

- To consider the status involved in carrying out a survey.
- To analyse the process involved in observation and interpreting.
- To describe the methods of collecting primary statistical information.
- To define and describe sampling.
- To analyse the basis of sampling.
- To describe a variety of sampling methods.

Statistical investigation is a comprehensive and requires systematic collection of data about some group of people or objects, describing and organizing the data, analyzing the data with the help of different statistical method, summarizing the analysis and using these results for making judgements, decisions and predictions. The validity and accuracy of final judgement is most crucial and depends heavily on how well the data was collected in the first place. The quality of data will greatly affect the conditions and hence at most importance must be given to this process and every possible precaution should be taken to ensure accuracy while collecting the data.



Nature of data:

It may be noted that different types of data can be collected for different purposes. The data can be collected in connection with time or geographical location or in connection with time and location. The following are the three types of data:

- Time series data.
- Spatial data.
- Spacio-temporal data.

Time series data:

It is a collection of a set of numerical values, collected over a period of time. The data might have been collected either at regular intervals of time or irregular intervals of time.

Example

The following is the data for the three types of expenditures in rupees for a family for the four years 2011,2012,2013,2014.

Year	Food	Education	Others	Total
2011	2000	3000	3000	8000
2012	2500	3500	3500	9500
2013	3000	2500	4000	9500
2014	3000	4000	5000	12000

Spatial Data:

If the data collected is connected with that of a place, then it is termed as spatial data. For example, the data may be

- Number of runs scored by a batsman in different test matches in a test series at different places



- District wise rainfall in Tamil Nadu
- Prices of silver in four metropolitan cities
- State wise population in Tamil Nadu

Example

The population of the southern states of India in 2011.

State	Population
Andhra Pradesh	84665533
Karnataka	61130704
Kerala	33387677
Pondicherry	1244464
Tamil Nadu	72138958

Spacio Temporal Data:

If the data collected is connected to the time as well as place then it is known as spacio temporal data.

Example

State	Population	
	1981	1991
Andhra Pradesh	5,34,03,619	6,63,04,854
Karnataka	3,70,43,451	4,48,17,398
Kerala	2,54,03,217	2,90,11,237
Pondicherry	6,04,136	7,89,416
Tamil Nadu	4,82,97,456	5,56,38,318



Categories of data:

Any statistical data can be classified under two categories depending upon the sources utilized. These categories are,

- a) Primary data
- b) Secondary data

a) Primary data

Primary data is the one, which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by survey conducted by individuals or research institution or any organisation.

The primary data can be collected by the following five methods.

1. Direct personal interviews.
2. Indirect Oral interviews.
3. Mailed questionnaire method.
4. Information from correspondents.
5. Schedules sent through enumerators.

b) Secondary Data

Secondary data are those data which have been already collected and analysed by some earlier agency for its own use; and later the same data are used by a different agency. According to W.A. Neiswanger, a primary source is a publication in which the data are published by the same authority which gathered and analysed them. A secondary source is a publication, reporting the data which have been gathered by other authorities and for which others are responsible'.

Sources of Secondary data

In most of the studies the investigator finds it impracticable to collect first-hand information on all related issues and as such he makes use of the data collected by others. There is a vast amount of published information



from which statistical studies may be made and fresh statistics are constantly in a state of production. The sources of secondary data can broadly be classified under two heads:

- a) Published sources
- b) Unpublished sources

a) Published Sources:

Generally, published sources are international, national, govt., semi-Govt, private corporate bodies, trade associations, expert committee and commission reports and research reports. They collect the statistical data in different fields like national income, population, prices, employment, wages, export, import etc. These reports are published on regular basis i.e., annually, quarterly, monthly, fortnightly, weekly, daily and so on. These published sources of the secondary data are given below:

i) Govt. Publications

The Central Statistical Organization (CSO) and various state govt. collect compile and publish data on regular basis. Some of the important such publications are:

- Indian Trade Journals
- Reports on Currency and Finance
- Indian Customs and Central Excise Tariff
- Statistical Abstract of India
- Reserve Bank of India Bulletin
- Labour Gazette
- Agricultural Statistics of India
- Bulletin of Agricultural Prices
- Indian Foreign Statistics
- Economic Survey and so on.



ii) International Bodies

All foreign Governments and international agencies publish regular reports of international significance. These reports are regularly published by the agencies like;

- United Nations Organization
- World Health Organization
- International Labour Organization
- Food and Agriculture Organization
- International Bank for Reconstruction and Development
- World Meteorological Organization.

iii) Semi Govt. Publications

Semi govt, organizations municipalities, District Boards and others also publish reports in respect of birth, death and education, sanitation and many other related fields.

iv) Reports of Committee and Commissions

Central Govt, or State Govt, sometimes appoints committees and commissions on matters of great importance. Reports of such committees are of great significance as they provide invaluable data. These reports are like, Shah Commission Report, Sarkaria Commission Report and Finance Commission Reports etc.

v) Private Publications

Some commercial and research institutes publish reports regularly. They are like Institutes of Economic Growth, Stock Exchanges, National Council of Education Research and Training (NCERT), National Council of Applied Economic Research (NCAER) etc.

vi) Newspapers and Magazines

Various newspapers as well as magazines also do collect data in respect of many social and economic aspects. Some of them are as:



- Economic Times
- Financial Express
- Hindustan Times
- Indian Express
- Business Standard
- Economic and Political Weekly
- Main-stream
- Kurukshetra
- Yojna etc.

vii) Research Scholars:

Individual research scholars collect data to complete their research work which further is published with their research papers.

b) Unpublished Source

There are certain records maintained properly by the govt, agencies, private offices and firms. These data are not published.

Limitations of Secondary Data

One should not use the secondary data without care and precautions. As such, secondary data suffers from pitfalls and limitations as stated below:

- No proper procedure is adopted to collect the data.
- Sometimes, secondary data is influenced by the prejudice of the investigator.
- Secondary data sometimes lacks standard of accuracy.
- Secondary data may not cover the full period of investigation.



1.7 CLASSIFICATION

Classification defined as: “the process of arranging things in groups or classes according to their resemblances and affinities and gives expression to the unity of attributes that may subsist amongst a diversity of individuals”.

The Collected data, also known as raw data or ungrouped data are always in an unorganised form and need to be organised and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis. It is, therefore, essential for an investigator to condense a mass of data into more and more comprehensible and assailable form. The process of grouping into different classes or sub classes according to some characteristics is known as classification, tabulation is concerned with the systematic arrangement and presentation of classified data. Thus classification is the first step in tabulation. For Example, letters in the post office are classified according to their destinations viz., New Delhi, Mumbai, Bangalore, Chennai etc.,

Objects of Classification:

The following are main objectives of classifying the data:

- It eliminates unnecessary details.
- It facilitates comparison and highlights the significant aspect of data.
- It enables one to get a mental picture of the information and helps in drawing inferences.
- It helps in the statistical treatment of the information collected.

Types of classification:

Statistical data are classified in respect of their characteristics. Broadly there are four basic types of classification namely

- Qualitative classification
- Quantitative classification
- Chronological classification



- Geographical classification

Qualitative classification

Qualitative classification is done according to attributes or non-measurable characteristics; like social status, sex, nationality, occupation, etc.

For example, the population of the whole country can be classified into four categories as married, unmarried, widowed and divorced.

When only one attribute, e.g., sex, is used for classification, it is called simple classification.

When more than one attributes, e.g., deafness, sex and religion, are used for classification, it is called manifold classification.

Quantitative classification

If the data are classified on the basis of phenomenon which is capable of quantitative measurements like age, height, weight, prices, production, income expenditure, sales, profits, etc., it is termed as quantitative variable.

For example the daily incomes of different retail shops in a town may be classified as under.

Daily earnings in rupees of 100 retail shop in a town

Daily earnings	No. of retail shops
Upto 100	9
101 - 200	25
201 - 300	33
301 - 400	28
401 - 500	2
Above 500	8



In the above classification, the daily earnings of the shops are termed as variable and the number of shops in each class or group as the frequency. This classification is called grouped frequency distribution. Hence this classification is often called 'classification by variables'.

Variable:

A variable in statistics means any measurable characteristic or quantity which can assume a range of numerical values within certain limits, e.g., income, height, age, weight, wage, price, etc. A variable can be classified as either a) Discrete, b) Continuous.

a) Discrete variable

A variable which can take up only exact values and not any fractional values, is called a 'discrete' variable. Number of workmen in a factory, members of a family, students in a class, number of births in a certain year, number of telephone calls in a month, etc., are examples of discrete-variable.

b) Continuous variable

A variable which can take up any numerical value (integral/fractional) within a certain range is called a 'continuous' variable. Height, weight, rainfall, time, temperature, etc., are examples of continuous variables. Age of students in a school is a continuous variable as it can be measured to the nearest fraction of time, i.e., years, months, days, etc.

Chronological classification

When the data are classified on the basis of time then it is known as chronological classification. Such series are also known as time series because one of the variables in them is time. If the population of India during the last eight censuses is classified it will result in a time series or chronological classification.

The following table would give an idea of chronological classification:

Production of Washing Machine by Company 'X'
--



1966	2600
1967	3400
1968	4800
1969	5100
1970	6900
1971	7300
1972	8600
1973	9800

Geographical classification

When the data are classified by geographical regions or location, like states, provinces, cities, countries etc..., this type of classification is based on geographical or location differences between various items in the data like states, cities, regions, zones etc. For eg. The Rainfall output per Millimetre for different states of India in some given period may be presented as follows:

Rainfall output of different countries in 2015 (Millimetre (mm))

States	Tamil Nadu	Andhra Pradesh	Kerala	Karnataka	Pondicherry
Avg. Output (mm) (Approximate)	70.13	65.33	85.21	75.66	6.23

1.8 TABULATION

Tabulation may be defined as the systematic presentation of numerical data in rows or/and columns according to certain characteristics. It is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information. A table is a systematic arrangement of classified data in columns and rows. Thus, a statistical table makes it possible for the investigator to present a huge mass of data in a detailed and



orderly form. It facilitates comparison and often reveals certain patterns in data which are otherwise not obvious. Classification and 'Tabulation', as a matter of fact, are not two distinct processes. Actually they go together. Before tabulation data are classified and then displayed under different columns and rows of a table.

Objectives of Tabulation:

The main objectives of tabulation are following

- To carry out investigation;
- To do comparison;
- To locate omissions and errors in the data;
- To use space economically;
- To study the trend;
- To simplify data;
- To use it as future reference.

Advantages of Tabulation:

The advantages of Tabulation are following

- It simplifies complex data and the data presented are easily understood.
- It facilitates comparison of related facts.
- It facilitates computation of various statistical measures like averages, dispersion, correlation etc.
- It presents facts in minimum possible space and unnecessary repetitions and explanations are avoided. Moreover, the needed information can be easily located.
- Tabulated data are good for references and they make it easier to present the information in the form of graphs and diagrams.



Table:

The making of a compact table itself an art. This should contain all the information needed within the smallest possible space. What the purpose of tabulation is and how the tabulated information is to be used are the main points to be kept in mind while preparing for a statistical table. An ideal table should consist of the following main parts are: (i) Table number; (ii) Title of the table; (iii) Captions or column headings; (iv) Stubs or row designation; (v) Body of the table; (vi) Footnotes; and (vii) Sources of data.

Types of Tables:

The tables can be classified according to their purpose, stage of enquiry, nature of data or number of characteristics used. On the basis of the number of characteristics, tables classified as follows: (i) Simple or one-way table; (ii) Two way table; and (iii) Manifold table.

A good statistical table is not merely a careless grouping of columns and rows but should be such that it summarizes the total information in an easily accessible form in minimum possible space. Thus while preparing a table, one must have a clear idea of the information to be presented, the facts to be compared and the points to be stressed.

1.9 FREQUENCY DISTRIBUTION:

A frequency distribution is an arrangement where a number of observations with similar or closely related values are put in separate groups, each group being in order of magnitudes in the arrangement based on magnitudes. It is a series when a number of observations with similar or closely related values are put in separate bunches or groups, each group being in order of magnitude in a series. It is simply a table in which the data are grouped into classes and the numbers of cases which fall in each class are recorded. It shows the frequency of occurrence of different values of a single Phenomenon.

The frequency distribution is constructed for three main reasons are: (i) To facilitate the analysis of data.



(ii) To estimate frequencies of the unknown population distribution from the distribution of sample data.

(iii) To facilitate the computation of various statistical measures.

Example

60	70	55	50	80	65	40	30	80	90
35	45	75	65	70	80	82	55	65	80
90	55	38	65	75	85	60	65	45	75

The above figures are nothing but raw or ungrouped data and they are recorded as they occur without any pre consideration. This representation of data does not furnish any useful information and is rather confusing to mind. A better way to express the figures in an ascending or descending order of magnitude and is commonly known as array. But this does not reduce the bulk of the data. The above data when formed into an array is in the following form:

30	35	38	40	45	45	50	55	55	55
60	60	65	65	65	65	65	65	70	70
75	75	75	80	80	80	80	85	90	90

The array helps us to see at once the maximum and minimum values. It also gives a rough idea of the distribution of the items over the range. When we have a large number of items, the formation of an array is very difficult, tedious and cumbersome. The Condensation should be directed for better understanding and may be done in two ways, depending on the nature of the data.



Discrete frequency distribution

Discrete frequency distribution shows the number of times each value and not to a range of values, of the variable occurs in the data set. Discrete frequency distribution is called ungrouped frequency distribution. In this form of distribution, the frequency refers to discrete value. Here the data are presented in a way that exact measurement of units is clearly indicated. There are definite differences between the variables of different groups of items. Each class is distinct and separate from the other class. Non-continuity from one class to another class exists. Data as such facts like, the number of rooms in a house, the number of companies registered in a country, the number of children in a family, etc.

Example

In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

3	1	3	2	1	5	6	6
2	0	0	3	4	2	1	2
1	3	1	5	3	3	2	4
2	2	3	0	2	1	4	5
4	2	3	4	1	2	5	4

Represent the data in the form of a discrete frequency distribution.

Solution:

Frequency distribution of the number of children

Number of Children	Tally Marks	Frequency
0		3
1		7
2		10
3		8



4		6
5		4
6		2
	Total	40

Grouped frequency distribution

Whenever the range of values of the variable is large for example 0 to 100 or 15 to 200 and if the data is represented by discrete frequency distribution, the data will still remain unwieldy and need further processing for condensation and statistical analysis. Grouped frequency distribution is called continuous frequency distribution. In this form of distribution refers to groups of values. This becomes necessary in the case of some variables which can take any fractional value and in which case an exact measurement is not possible. Hence a discrete variable can be presented in the form of a continuous frequency distribution.

Weekly wages (Rs.)	Number of Employees
1500-2000	4
2000-2500	12
2500-3000	22
3000-3500	33
3500-4000	16
4000-4500	8
4500-5000	5
Total	100

To understand the contraction of the Grouped frequency distribution, the following technical terms need definition and its calculations.

- Class interval



The various groups into which the values of the variable are classified are known as class intervals or simply classes. For example, the symbol 25-35 represents a group or class which includes all the values from 25 to 35.

- Class limits

- ❖ The two values (maximum and minimum) specifying the class intervals are called the class limits. The lowest value is called the lower limit and the highest value the upper limit of the class.
- ❖ Length or width of the class is defined as the difference between the upper and the lower limits of the class. That is Class mark or Midpoint of a class = $(\text{Lower Limit} + \text{Upper Limit}) / 2$

Example

Consider a class denoted as 25 – 50

- The class 25-50 includes all the values in between 25 to 50
- The lower limit of the class is 25 and the upper limit 50
- The length of the class is given by:

$$\text{Length} = \text{Upper limit} - \text{Lower limit} = 50 - 25 = 25.$$

- The class mark or mid value of the class is given by

$$\begin{aligned} \text{Mid value} &= (\text{Lower Limit} + \text{Upper Limit}) / 2 \\ &= (25 + 50) / 2 = 37.5 \end{aligned}$$

Nominal:

Let's start with the easiest one to understand. Nominal scales are used for labelling variables, without any quantitative value. "Nominal" scales could simply be called "labels." Here are some examples, below. Notice that all of these scales are mutually exclusive (no overlap) and none of them has any numerical significance. A good way to remember all of this is that



“nominal” sounds a lot like “name” and nominal scales are kind of like “names” or labels.

What is your gender?	What is your hair colour?	What is your live?
M – Male	1- Brown	A – North of the equator
F - female	2- Black	B - South of the equator
	3- Blonde	C - Neither, In the international space station
	4- Gray	
	5- Other	

Ordinal:

Ordinal scales, it is the order of the values is what’s important and significant, but the differences between each one is not really known. Take a look at the example below. In each case, we know that a = 4 is better than a = 3 or 2, but we don’t know—and cannot quantify – how *much* better it is. For example, is the difference between “OK” and “Unhappy” the same as the difference between “Very Happy” and “Happy?” We can’t say. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc. “Ordinal” is easy to remember because is sounds like “order” and that’s the key to remember with “ordinal scales”—it is the *order* that matters, but that’s all you really get from these.

How do you feel today?	How satisfied are you with our service
1 – Very Unhappy	1- Very unsatisfied
2 – Unhappy	2- Somewhat Unsatisfied
3 – OK	3- Neutral
4 – Happy	4- Somewhat Satisfied
5 – Very Happy	5- Very Satisfied

Interval:

Interval scales are numeric scales in which we know not only the order, but also the exact differences between the values. The classic



example of an interval scale is Celsius temperature because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees. Time is another good example of an interval scale in which the increments are known, consistent, and measurable. Interval scales are nice because the realm of statistical analysis on these data sets opens up.

For example, central tendency can be measured by mode, median, or mean; standard deviation can also be calculated. Like the others, you can remember the key points of an “interval scale” pretty easily. “Interval” itself means “space in between,” which is the important thing to remember—interval scales not only tell us about order, but also about the value between each item. Here’s the problem with interval scales: they don’t have a “true zero.” For example, there is no such thing as “no temperature.” Without a true zero, it is impossible to compute ratios. With interval data, we can add and subtract, but cannot multiply or divide. Confused? Ok, consider this: 10 degrees + 10 degrees = 20 degrees. No problem there. 20 degrees is not twice as hot as 10 degrees, however, because there is no such thing as “no temperature” when it comes to the Celsius scale. I hope that makes sense. Bottom line, interval scales are great, but we cannot calculate ratios, which brings us to our last measurement scale...



Ratio:

Ratio scales are the ultimate nirvana when it comes to measurement scales because they tell us about the order, they tell us the exact value between units, and they also have an absolute zero—which allows for a wide range of both descriptive and inferential statistics to be applied. At the risk of repeating myself, everything above about interval data applies to ratio



scales + ratio scales have a clear definition of zero. Good examples of ratio variables include height and weight. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be meaningfully added, subtracted, multiplied, divided (ratios). Central tendency can be measured by mode, median, or mean; measures of dispersion, such as standard deviation and coefficient of variation can also be calculated from ratio scales.



This Device Provides Two Examples of Ratio Scales (height and weight)

1.10 DIAGRAMMATIC AND GRAPHICAL REPRESENTATION

One of the most convincing and appealing ways in which statistical results may be presented is through diagrams and graphs. Representation of statistical data by means of pictures, graphs and geometrical figures is called diagrammatic and graphical representations. The difference between the two is that in the case of diagrammatic representation the quantities are represented by diagrams and pictures and in case of graphical representation they are represented by points which are plotted on a graph paper.

Diagrammatic Representation:

Diagrammatic representation is used when the data relating to different times and places are given and they are independent of one another. One of the most convincing and appealing ways in which statistical results may be presented is through diagrams and graphs. Just one diagram is enough to represent a given data more effectively than thousand words. Moreover even a layman who has nothing to do with numbers can also



understands diagrams. Evidence of this can be found in newspapers, magazines, journals, advertisement, etc. An attempt is made in this chapter to illustrate some of the major types of diagrams and graphs frequently used in presenting statistical data. Diagram is a visual form for presentation of statistical data, highlighting their basic facts and relationship. If we draw diagrams on the basis of the data collected they will easily be understood and appreciated by all. It is readily intelligible and save a considerable amount of time and energy.

Significance of Diagrams and Graphs:

Diagrams and graphs are extremely useful because of the following reasons.

- They are attractive and impressive.
- They make data simple and intelligible.
- They make comparison possible
- They save time and labour.
- They have universal utility.
- They give more information.
- They have a great memorizing effect.

General rules for constructing diagrams:

The construction of diagrams is an art, which can be acquired through practice. However, observance of some general guidelines can help in making them more attractive and effective. The diagrammatic presentation of statistical facts will be advantageous provided the following rules are observed in drawing diagrams.

- A diagram should be neatly drawn and attractive.
- The measurements of geometrical figures used in diagram should be accurate and proportional.
- The size of the diagrams should match the size of the paper.



- Every diagram must have a suitable but short heading.
- The scale should be mentioned in the diagram.
- Diagrams should be neatly as well as accurately drawn with the help of drawing instruments.
- Index must be given for identification so that the reader can easily make out the meaning of the diagram.
- Footnote must be given at the bottom of the diagram.
- Economy in cost and energy should be exercised in drawing diagram.

Types of diagrams:

In practice, a very large variety of diagrams are in use and new ones are constantly being added. For the sake of convenience and simplicity, they may be divided under the following heads:

- One-dimensional diagrams
- Two-dimensional diagrams
- Three-dimensional diagrams
- Pictograms and Cartograms

One-dimensional diagrams

In such diagrams, only one-dimensional measurement, i.e height is used and the width is not considered. These diagrams are in the form of bar or line charts and can be classified as

- Line Diagram
- Simple Diagram
- Multiple Bar Diagram
- Sub-divided Bar Diagram
- Percentage Bar Diagram



Line Diagram:

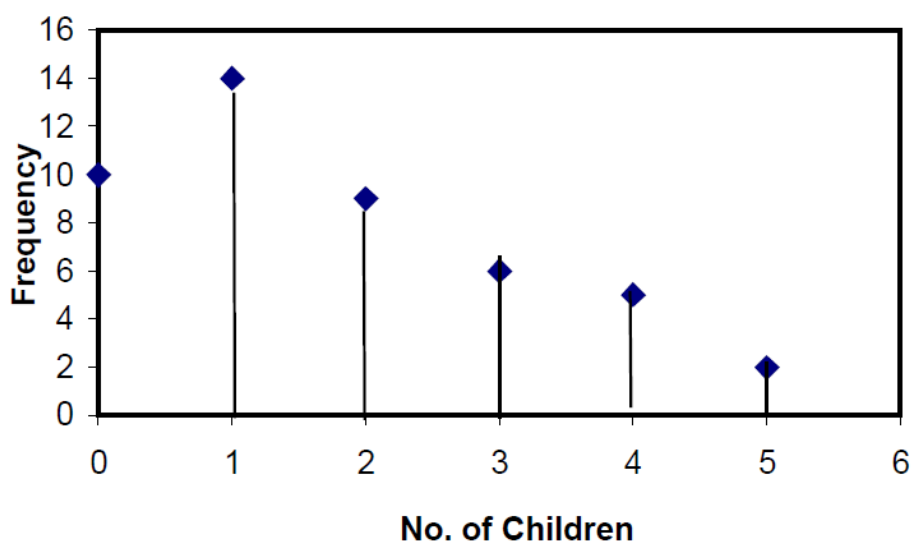
Line diagram is used in case where there are many items to be shown and there is not much of difference in their values. Such diagram is prepared by drawing a vertical line for each item according to the scale. The distance between lines is kept uniform. Line diagram makes comparison easy, but it is less attractive.

Example

Show the following data by a line chart.

No. of Children	0	1	2	3	4	5
Frequency	10	14	9	6	4	2

Line Diagram



Two-dimensional Diagrams

In one-dimensional diagrams, only length is taken into account. But in two-dimensional diagrams the area represents the data and so the length and breadth have both to be taken into account. Such diagrams are also called area diagrams or surface diagrams. The important types of area diagrams are:

- Rectangles



- Squares
- Pie-diagrams

Three-dimensional diagrams

Three-dimensional diagrams, also known as volume diagram, consist of cubes, cylinders, spheres, etc. In such diagrams three things, namely length, width and height have to be taken into account. Of all the figures, making of cubes is easy. Side of a cube is drawn in proportion to the cube root of the magnitude of data.

Pictograms and Cartograms

Pictograms are not abstract presentation such as lines or bars but really depict the kind of data we are dealing with. Pictures are attractive and easy to comprehend and as such this method is particularly useful in presenting statistics to the layman. When Pictograms are used, data are represented through a pictorial symbol that is carefully selected. Cartograms or statistical maps are used to give quantitative information as a geographical basis. They are used to represent spatial distributions. The quantities on the map can be shown in many ways such as through shades or colours or dots or placing pictogram in each geographical unit.

GRAPHICAL REPRESENTATION:

The graphical representation is used when we have to represent the data of a frequency distribution and a time series. A graph represents mathematical relationship between the two variables whereas a diagram does not. A graph is a visual form of presentation of statistical data. A graph is more attractive than a table of figure. Even a common man can understand the message of data from the graph. Comparisons can be made between two or more phenomena very easily with the help of a graph. Finally graphs are more obvious, precise and accurate than diagrams and are quite helpful to the statistician for the study of slopes, rates of changes and estimation, whenever possible. However here we shall discuss only some important types of graphs which are more popular and they are



- Histogram
- Frequency Polygon
- Frequency Curve
- Ogive
- Lorenz Curve

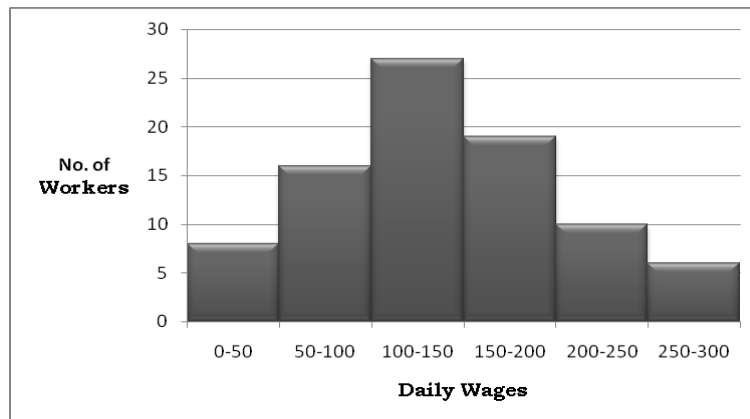
Histogram

A histogram consists of bars or rectangles which are erected over the class intervals, without giving gaps between bars and such that the areas of the bars are proportional to the frequencies of the class intervals. It is a bar chart or graph showing the frequency of occurrence of each value of the variable being analysed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the 'X-axis' and the frequencies on the 'Y-axis'. The height of each rectangle represents the frequency of the class interval. Each rectangle is formed with the other so as to give a continuous picture. Such a graph is also called staircase or block diagram. However, we cannot construct a histogram for distribution with open-end classes. It is also quite misleading if the distribution has unequal intervals and suitable adjustments in frequencies are not made.

Example

Draw a histogram for the following data.

Daily wages	Number of Workers
0-50	8
50-100	16
100-150	27
150-200	19
200-250	10
250-300	6



Frequency Polygon

Frequency Polygon is another device of graphic presentation of a frequency distribution. In case of discrete frequency distribution frequency polygon is obtained on plotting the frequencies on the vertical axis (y-axis) against the corresponding values of the variable on the horizontal axis (x-axis) and joining the points so obtained by straight lines. In case of grouped or continuous frequency distribution the construction of frequency polygon is consist in plotting the frequencies of different classes (along y-axis). The points so obtained are joined by straight lines to obtain the frequency polygon.

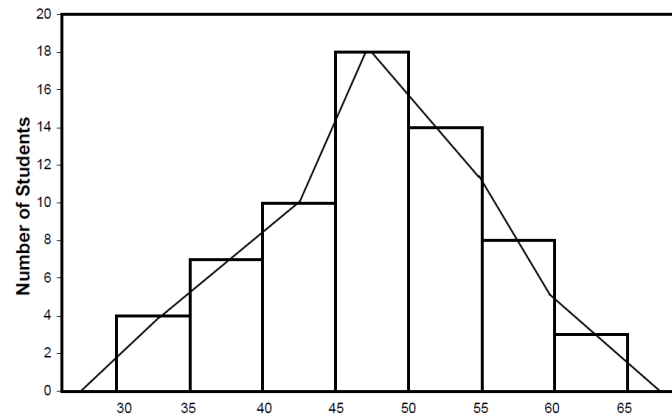
Example

Draw a Frequency polygon for the following data.

Weight (in Kg.)	30-35	35-40	40-45	45-50	50-55	55-60	60-65
No. of Students	4	7	10	18	14	8	3



FREQUENCY POLYGON



Frequency Curve

A frequency curve is a smooth free hand curve drawn through the vertices of frequency polygon. The object of smoothing of the frequency polygon is to eliminate as far as possible, the random or erratic changes that might be present in the data. The area enclosed shape is smooth one and not with sharp edges.

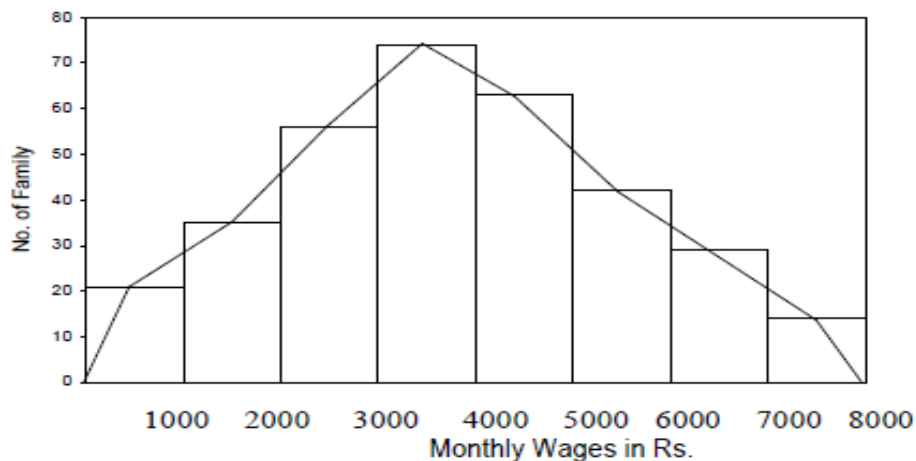
Example

Draw a frequency curve for the following data.

Monthly Wages (in Rs)	No. of family
0-1000	21
1000-2000	35
2000-3000	56
3000-4000	74
4000-5000	63
5000-6000	40
6000-7000	29
7000-8000	14



FREQUENCY CURVE



Ogives or cumulative frequency curves:

For a set of observations, we know how to construct a frequency distribution. In some cases we may require the number of observations less than a given value or more than a given value. This is obtained by accumulating (adding) the frequencies upto (or above) the give value. This accumulated frequency is called cumulative frequency. These cumulative frequencies are then listed in a table is called cumulative frequency table. The curve table is obtained by plotting cumulative frequencies is called a cumulative frequency curve or an ogive. There are two methods of constructing ogive namely: a) The 'less than ogive' method, b) The 'more than ogive' method

In less than ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve. In more than ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

Example

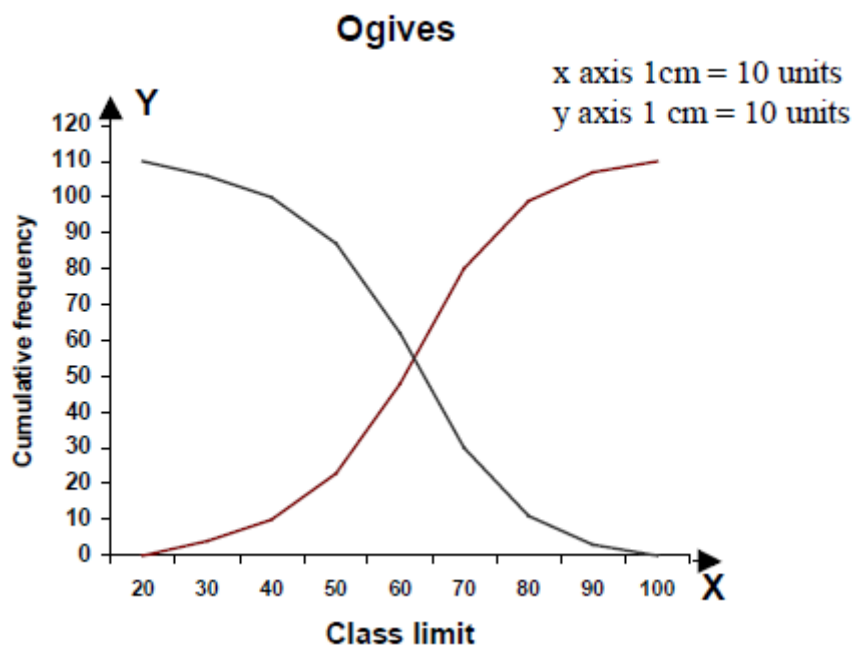
Draw the O gives for the following data.

C.I	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
F	4	6	13	25	32	19	8	3



Solution

Class limit	Less than ogive	More than ogive
20	0	110
30	4	106
40	10	100
50	23	87
60	48	62
70	80	30
80	99	11
90	107	3
100	110	0





QUESTIONS

1. Explain the origin of statistics.
2. Write the meaning and definitions of statistics.
3. Write the definitions of statistics as given by Croxton and Crowden?
Explain the four stages in statistics.
4. Explain the categories of data.
5. Explain the classification and objects of classification.
6. Describe the advantages of tabulation.
7. Explain the process of preparing a table
8. Explain the scales of measurement?
9. Write the general rules of constructing diagrams.
10. Explain the functions of statistics.
11. Describe the scope of statistics.
12. What are the limitations of statistics?
13. Explain the collection of data.
14. Explain the classification and Tabulation.
15. Describe the frequency distribution.
16. Write the types of diagrams and graphs



UNIT – II

2.1 MEASURES OF CENTRAL TENDENCY

The collected data such are not suitable to draw conclusions about the mass from which it has been taken. Some inferences about the population can be drawn from the frequency distribution of the observed values. This process of condensation of data reduces the bulk of data and the frequency distribution is categorised by certain constraints known as parameters.

R. A. Fisher has rightly said, “The inherent inability of the human mind to grasp entirely a large body of numerical data compels us to seek relatively few constants that will adequately describe the data.”

It is not possible to grasp any idea about the characteristic when we look at all the observations. So it is better to get one number for one group. That number must be a good representative one for all the observations to give a clear picture of that characteristic. Such representative number can be a central value for all these observations. This central value is called a measure of central tendency or an average or a measure of locations. There are five averages. Among them mean, median and mode are called simple averages and the other two averages geometric mean and harmonic mean are called special averages.

The meaning of average is nicely given in the following definitions. “A measure of central tendency is a typical value around which other figures congregate.” “An average stands for the whole group of which it forms a part yet represents the whole.” “One of the most widely used set of summary figures is known as measures of location.” There are three popular measures of central tendency namely,

- Mean
- Median
- Mode



Each of these will be discussed in detail here. Besides these, some other measures of location are also dealt with, such as quartiles, deciles and percentiles

Characteristics for a good or Measures of Central tendency:

There are various measures of central tendency. The difficulty lies in choosing the measure as no hard and fast rules have been made to select any one. A measure of central tendency is good or satisfactory if it possesses the following characteristics,

- It should be based on all the observations.
- It should not be affected by the extreme values.
- It should be close to the maximum number of observed values as possible.
- It should be based on all items in the data.
- Its definition shall be in the form of a mathematical formula
- It should be defined rigidly which means that it should have a definite value. The experimenter or investigator should have no discretion.
- It should not be subjected to complicated and tedious calculations.
- It should be capable of further algebraic treatment.
- It should be stable with regard to sampling.

MEAN

Mean of a variable is defined as the sum of the observed values of asset divided by the number of observations in the set is called a mean or an averaged. If the variable x assumes n values $x_1 + x_2 + \dots + x_n$ then the mean, \bar{x} , is given by

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$



$$\bar{x} = \frac{\sum x}{n}$$

This formula is for the ungrouped or raw data.

Under this method an assumed or an arbitrary average (indicated by A) is used as the basis of calculation of deviations from individual values.

The formula is

$$\bar{x} = A + \frac{\sum d}{n}$$

A = the assumed mean or any value in x.

d = the deviation of each value from the assumed mean

Example

The heights of five runners are 160 cm, 137 cm, 149 cm, 153 cm and 161 cm respectively. Find the mean height per runner.

$$\bar{x} = \frac{\sum x}{n}$$

Mean height = Sum of the heights of the runners/number of runners

$$= (160 + 137 + 149 + 153 + 161)/5 \text{ cm}$$

$$= 760/5 \text{ cm}$$

$$= 152 \text{ cm.}$$

Hence, the mean height is 152 cm.

Example

A Student's marks in 5 subjects are 15, 25, 35, 45, 55, 65. Find his average mark.



Solution:

X	d = x-A
15	-30
25	-20
35	-10
45	0
55	10
65	20
Total	-30

$$A = 45$$

$$\begin{aligned}\bar{x} &= A + \frac{\sum d}{n} \\ &= 45 + \frac{-30}{6} \\ &= 45 + (-5) \\ \bar{x} &= 40\end{aligned}$$

Grouped Data:

The mean for grouped data is obtained from the following formula:

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

Where x = the mid-point of individual class

f = the frequency of individual class

N = some of the frequencies of total frequencies



Another method:

$$\bar{x} = A + \frac{\sum fd}{N} \times C \quad \text{Where } d = \frac{x-A}{c}$$

A = any value in x, N = Total frequency, C= width of the class interval

Example

Calculate the arithmetic mean, given the following Income of No.of persons.

Income	10-20	20-30	30-40	40-50	50-60	60-70
No. of persons	4	7	16	2	15	8

Solution:

Income	No. of persons (f)	Mid Value (x)	fx
10-20	4	15	60
20-30	7	25	175
30-40	16	35	560
40-50	2	45	900
50-60	15	55	825
60-70	8	65	520
	70		3040

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

$$\bar{x} = \frac{3040}{70} = 43.3$$

$$\bar{x} = 43.3$$

Example

Calculate arithmetic mean. Following the Number of persons according to different income groups.



Income	10-20	20-30	30-40	40-50	50-60	60-70
No. of persons	4	7	16	2	15	8

Solution:

Income	Number of Persons (f)	Mid value (x)	$d = \frac{x - A}{c}$	fd
10-20	4	15	-3	-12
20-30	7	25	-2	-14
30-40	16	35	-1	-16
40-50	2	45	0	0
50-60	15	55	1	15
60-70	8	65	2	16
	70			-11

$$A = 45, C.I = 10$$

$$\begin{aligned}\text{Mean } \bar{x} &= A + \frac{\sum fd}{N} \times C \\ &= 45 + \frac{-11}{70} \times 10 \\ &= 45 + \frac{-11}{7} \\ &= 45 - 1.57\end{aligned}$$

$$\text{Mean } \bar{x} = 43.43$$

Merits and demerits of Arithmetic mean:

Merits

1. It is rigidly defined.
2. It is easily to understood and easily to calculated



3. It is every item is taken calculated.
4. It can further be subjected to algebraic treatment unlike other measures i.e. mode and median.
5. If the number of items is sufficiently large, it is more accurate and more reliable.
6. It is a calculated value and is not based on its position in the series.
7. It is possible to calculate even if some of the details of the data are lacking.
8. It provides a good basis for comparison.

Demerits

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be computed when class intervals have open ends.
6. It may give impossible and fallacious conclusions.

MEDIAN

Median of a distribution is the value of the variable which divides it into two equal parts. It is the value which exceeds and is exceeded by the same number of observations, i.e., it is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a positional average.

Ungrouped or Raw data:

Arrange the given values in the increasing or decreasing order. If the numbers of values are odd, median is the middle value. If the numbers of values are even, median is the mean of middle two values. By formula



$$\text{Median} = \left[\frac{n+1}{2} \right]^{\text{th}} \text{ item.}$$

Example

When odd number of values are given. Find median for the following data 9, 12, 7, 16, 13.

Solution:

Arranging the data in the increasing order 7, 9, 12, 13, and 16. The middle value is the 3th item i.e., 12 is the median Using formula

$$\begin{aligned} \text{Median} &= \left[\frac{n+1}{2} \right]^{\text{th}} \text{ item} \\ &= \left[\frac{5+1}{2} \right]^{\text{th}} \text{ item} \\ &= \left[\frac{6}{2} \right]^{\text{th}} \text{ item} \\ &= 3^{\text{th}} \text{ item} \end{aligned}$$

Median is 12

Example

Find median for the following data 5, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the data in the increasing order 2, 5, 8, 10, 12, 18, 22, 30. Here median is the mean of the middle two items i.e., mean of (10, 12)

$$= \left[\frac{10+12}{2} \right] = 11$$

Median = 11

Using the formula

$$\text{Median} = \left[\frac{n+1}{2} \right]^{\text{th}} \text{ item}$$



$$\begin{aligned} &= \left[\frac{8+1}{2} \right]^{\text{th}} \text{ item} \\ &= 4.5^{\text{th}} \text{ item} \\ &= 4^{\text{th}} \text{ item} + \left(\frac{1}{2} \right) (5^{\text{th}} \text{ item} - 4^{\text{th}} \text{ item}) \\ &= 10 + \left(\frac{1}{2} \right) (12 - 10) \\ &= 10 + 1 \\ &= 11 \end{aligned}$$

Median = 11

Example

The following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and Accountancy.

S. No	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Accountancy)	57	45	24	31	25	84	43	80	32	72

Indicate in which subject is the level of knowledge higher?

Solution:

For such question, median is the most suitable measure of central tendency. The mark in the two subjects are first arranged in increasing order as follows:

S. No	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Accountancy)	57	45	24	31	25	84	43	80	32	72



$$\begin{aligned}\text{Median} &= \left[\frac{n+1}{2} \right]^{\text{th}} \text{ item} = \left[\frac{10+1}{2} \right]^{\text{th}} \text{ item} \\ &= 5.5^{\text{th}} \text{ item} \\ &= \left[\frac{\text{Value of 5th item} + \text{Value of 6th item}}{2} \right]\end{aligned}$$

$$\text{Median (Statistics)} = \left[\frac{46+47}{2} \right] = 46.5$$

$$\text{Median (Accountancy)} = \left[\frac{43+45}{2} \right] = 44$$

Grouped Data:

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution, cumulative frequencies have to be calculated to know the total number of items.

Cumulative frequency:

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the previous classes, i.e., adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

Discrete Series:

Step 1: Find cumulative frequencies.

Step 2: Find $\left[\frac{N+1}{2} \right]$

Step 3: See in the cumulative frequencies the value just greater than $\left[\frac{N+1}{2} \right]$

Step 4: Then the corresponding value of x is median.

Example

Find the median for the following frequency distribution.

Number of Member (X)	1	2	3	4	5	6	7	8	9
Frequency (f)	8	10	11	16	20	25	15	9	6



Solution:

X	f	c.f
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120
	N = 120	

Median = size of $\left[\frac{N+1}{2}\right]^{\text{th}}$ item

= 60.5th item

The cumulative frequency just greater than 60.5 is 65, and the value of x corresponding to 65 is 5. Hence the median size is 5.

Continuous Series:

The steps given below are followed for the calculation of median in continuous series.

Step 1: Find cumulative frequencies.

Step 2: Find $\left[\frac{N}{2}\right]$

Step 3: See in the cumulative frequency the value first greater than $\left[\frac{N}{2}\right]$, Then the corresponding class interval is called the Median class. Then apply the formula



$$\text{Median} = l + \left[\frac{\frac{N}{2} - m}{f} \right] \times c$$

Where l = Lower limit of the median class

m = cumulative frequency preceding the median

c = width of the median class

f = frequency in the median class.

N = Total frequency.

Note:

If the class intervals are given in inclusive type convert them into exclusive type and call it as true class interval and consider lower limit in this.

Example

Find the median, Based on the grouped data below.

Time to travel to work	Frequency
1 – 10	8
11 – 20	14
21 – 30	12
31 – 40	9
41 – 50	7

Solution:

Time to travel to work	Frequency	Class Interval	Cumulative frequency (c.f)
1 – 10	8	1.5 – 10.5	8
11 – 20	14	10.5 – 20.5	22



21 – 30	12	20.5 – 30.5	34
31 – 40	9	30.5 – 40.5	43
41 – 50	7	40.5 – 50.5	50

$$\text{Median} = l + \left[\frac{\frac{N-m}{2}}{f} \right] \times c$$

$$\frac{N}{2} = \frac{50}{2} = 25$$

..... class median is the 3rd class

Here $l=20.5$, $N = 50$, $f=12$, $c = 10$, $m = 22$

$$\begin{aligned} \text{Median} &= 20.5 + \left[\frac{25-22}{12} \right] \times 10 \\ &= 20.5 + 2.5 \\ &= 23 \end{aligned}$$

Thus, 25 persons take less than 23 minutes to travel to work and another 25 persons take more than 23 minutes to travel to work.

Example

Calculate median from the following data.

Value	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39
Frequency	5	8	10	12	7	6	3	2

Solution:

Value	f	Class interval	c.f
0-4	5	0.5-4.5	5
5-9	8	4.5-9.5	13
10-14	10	9.5-14.5	23
15-19	12	14.5-19.5	35
20-24	7	19.5-24.5	42



25-29	6	24.5-29.5	48
30-34	3	29.5-34.5	51
35-39	2	34.5-39.5	53
	53		

$$\frac{N}{2} = \frac{53}{2} = 26.5$$

$$\begin{aligned}\text{Median} &= l + \left[\frac{\frac{N}{2} - m}{f} \right] \times c \\ &= 14.5 + \left[\frac{26.5 - 23}{12} \right] \times 5\end{aligned}$$

Median is 15.96

Example

Find the median. Following are the daily Production in a Tea industry.

Production	Number of workers
Less than 10	5
Less than 20	3
Less than 30	4
Less than 40	3
Less than 50	3
Less than 60	4
Less than 70	7
Less than 80	9
Less than 90	7
Less than 100	8



Solution:

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 10, hence the width of the class interval equal to 10.

Class interval	f	c.f
0-10	5	5
10-20	3	8
20-30	4	12
30- 40	3	15
40-50	3	18
50-60	4	22
60-70	7	29
70-80	9	38
80-90	7	45
90-100	8	53

$$\frac{N}{2} = \frac{53}{2} = 26.5$$

$$\begin{aligned}\text{Median} &= l + \left[\frac{\frac{N}{2} - m}{f} \right] \times c \\ &= 60 + \left[\frac{26.5 - 22}{7} \right] \times 10\end{aligned}$$

$$\text{Median} = 66.74$$

So, about half the workers produced on tea less than 66.4, and the other half the workers produced on tea more than 66.4.



Example

Find median for the data given below.

Marks	Number of Students
Greater than 10	70
Greater than 20	62
Greater than 30	50
Greater than 40	38
Greater than 50	30
Greater than 60	24
Greater than 70	17
Greater than 80	9
Greater than 90	4

Solution:

Here we are given lower limit and more than cumulative frequencies.

Class interval	f	More than c.f	Less than c.f
10-20	8	70	8
20-30	12	62	20
30-40	12	50	32
40-50	8	38	40
50-60	6	30	46
60-70	7	24	53
70-80	8	17	61
80-90	5	9	66
90-100	4	4	70
	70		



$$\frac{N}{2} = \frac{70}{2} = 35$$

$$\begin{aligned}\text{Median} &= l + \left[\frac{\frac{N}{2} - m}{f} \right] \times c \\ &= 40 + \left[\frac{35 - 32}{8} \right] \times 10 \\ &= 40 + 3.75\end{aligned}$$

$$\text{Median} = 43.75$$

Example

Compute median for the following data.

Mid-Value	5	15	25	35	45	55	65	75
Frequency	7	10	15	17	8	4	6	7

Solution:

Here values in multiples of 10, so width of the class interval is 10.

Mid X	C.I	f	c.f
5	0-10	7	7
15	10-20	10	17
25	20-30	15	32
35	30-40	17	49
45	40-50	8	57
55	50-60	4	61
65	60-70	6	67
75	70-80	7	74

$$\frac{N}{2} = \frac{74}{2} = 37$$



$$\begin{aligned}\text{Median} &= l + \left[\frac{\frac{N}{2} - m}{f} \right] \times c \\ &= 30 + \left[\frac{37 - 32}{17} \right] \times 10 \\ &= 30 + 2.94 \\ \text{Median} &= 32.94\end{aligned}$$

MODE

It is another measure of central tendency. Mode is a value of a particular type of items which occur most frequently. Mode is a variate value which occurs most frequently in a set of values.

In case of discrete distribution, one can find mode by inspection. The variate value having the maximum frequency is the modal value.

Ungrouped or Raw Data:

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example

Find the mode for the following ungrouped data. 6, 5, 5, 7, 4, 8, 3, 4, 4

Answer: Mode is 4

In Some cases the mode may be absent while in some cases there may be more than one mode.

Example

Find the mode of the following observations. 4, 6, 8, 6, 7, 8, 8. In the given data, the observation 8 occurs maximum number of times (3) Mode is 8

Example

Find the mode of the following observations. 2, 4, 6, 8, 10, 12.

Answer: no mode



Grouped data:

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

Continuous distribution:

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the formula.

$$\text{Mode} = M_0 = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

l = Lower limit of the modal class

$$\Delta_1 = f_1 - f_0$$

$$\Delta_2 = f_1 - f_2$$

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

The above formula can also be written as

$$\text{Mode} = M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C$$

Remarks:

1. If $2f_1 - f_0 - f_2$ comes out to be zero, then mode is obtained by the following formula taking absolute differences within vertical lines.

$$2. M_0 = l + \frac{(f_1 - f_0)}{|f_1 - f_0| + |f_1 - f_2|} \times C$$

3. If mode lies in the first class interval, then f_0 is taken as zero.

4. The computation of mode poses no problem in distributions with open-end classes, unless the modal value lies in the open-end class.



Example

A survey conducted on 20 households in a locality by a group of students resulted in the following frequency table for the number of family members in a household:

Family Size	Number of families
1-3	7
3-5	8
5-7	2
7-9	2
9-11	1

Calculate the mode this data.

Solution:

The highest frequency is 8 and corresponding class interval is 3-5, which is the modal class.

$$\text{Mode} = M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C$$

- modal class = 3 – 5, lower limit of modal class = 3, class size (c) = 2
- frequency (f_1) of the modal class = 8,
- frequency (f_0) of class preceding the modal class = 7,
- frequency (f_2) of class succeeding the modal class = 2.

$$= 3 + \frac{8 - 7}{2 \times 8 - 7 - 2} \times 2$$

$$= 3 + \frac{2}{7}$$

$$= 3.286$$

Therefore, the mode of the data above is 3.286.



Example

The marks distribution of 30 students in a mathematics examination. Find the mode for the following data.

Class Interval	Number of Students (f)
10-25	2
25-40	3
40-55	7
55-70	6
70-85	6
85-100	6

Solution:

Since the maximum number of students, (i.e., 7) have got marks in the interval 40 - 55, the modal class is 40 - 55.

- lower limit (l) of the modal class = 40,
- class size (c) = 15,
- frequency (f_1) of modal class = 7,
- frequency (f_0) of the class preceding the modal class = 3,
- frequency (f_2) of the class succeeding the modal class = 6.

$$\text{Mode} = M_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times C$$

$$M_0 = 40 + \frac{7 - 3}{2 \times 7 - 6 - 3} \times 15$$

The mode marks is 52.

Definition of Mode:

In a grouped frequency distribution, it is not possible to determine the mode by looking at the frequencies. Here, we can only locate a class with the maximum frequency, called the **modal class**.



Determination of Modal class:

For a frequency distribution modal class corresponds to the maximum frequency. But in any one (or more) of the following cases

- If the maximum frequency is repeated
- If the maximum frequency occurs in the beginning or at the end of the distribution
- If there are irregularities in the distribution, the modal class is determined by the method of grouping.

Steps for Calculation:

We prepare a grouping table with 6 columns

- In column I, we write down the given frequencies.
- Column II is obtained by combining the frequencies two by two.
- Leave the 1st frequency and combine the remaining frequencies two by two and write in column III
- Column IV is obtained by combining the frequencies three by three.
- Leave the 1st frequency and combine the remaining frequencies three by three and write in column V
- Leave the 1st and 2nd frequencies and combine the remaining frequencies three by three and write in column VI.

Mark the highest frequency in each column. Then form an analysis table to find the modal class. After finding the modal class use the formula to calculate the modal value.

Example

Calculate mode for the following frequency distribution.

Class interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	9	12	15	16	17	15	10	13



Solution:

Grouping Table:

CI	f	2	3	4	5	6
0-5	9	21				
5-10	12		27	36		
10-15	15	31			43	
15-20	16		33			48
20-25	17	32		48		
25-30	15		25		42	
30-35	10	3				38
35-40	3					

Analysis table:

Columns	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
1					1			
2					1	1		
3				1	1			
4				1	1	1		
5		1	1	1				
6			1	1	1			
Total		1	2	4	5	2		

The maximum occurred corresponding to 20-25, and hence it is the modal class.

$$\text{Mode} = M_0 = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

Here $l = 20$, $\Delta_1 = f_1 - f_0 = 17 - 16 = 1$, $\Delta_2 = f_1 - f_2 = 17 - 15 = 1$



$$M_0 = 20 + \frac{1}{1+2} \times 5$$

$$= 20 + 1.67$$

Mode is 21.67

Example

Obtained the marks, Find the mode of the irregular distribution.

Marks	3	4	6	7	9	10	13	15	18	20
Frequency	4	8	15	20	32	16	14	35	10	6

Solution:

Marks	Frequency	3	4	5	6	7
1	2					
3	4					
		12		27		
4	8		23			
6	15					
		35			43	
7	20					
			52			
9	32					67
		48		68		
10	16					
			30		62	
13	14					
		49		59		65
15	35					



			45		51	
18	10					
		16				
20	6					

Columns	3	4	6	7	9	10	13	15	18	20
2								1		
3								1		
4				1	1					
5				1	1	1	1			
6					1		1			
7			1	1	1	1				
Total			1	3	4	2	2	2		

In the above table $X = 9$, the maximum sum of 1 is 4, hence the modal value is 9.

Merits and Demerits of Mode:

1. It is not affected by extreme values of a set of observations.
2. It can be calculated for distributions with open end classes.



3. The main drawback of mode is that often it does not exist.
4. Often its value is not unique.
5. It does not fulfil most of the requirements of a good measure of central tendency.

GEOMETRIC MEAN:

The geometric mean of a series containing n observations is the nth root of the product of the values. If x_1, x_2, \dots, x_n are observations then

$$\begin{aligned} \text{Geometric mean} &= \sqrt[n]{x_1 \cdot x_2 \dots x_n} \\ &= (x_1 \cdot x_2 \dots x_n)^{\frac{1}{n}} \\ &= \frac{1}{n} (\log x_1 + \log x_2 + \dots + \log x_n) \\ &= \frac{\sum \log x_i}{n} \end{aligned}$$

$$\text{Geometric mean} = \text{Antilog } \frac{\sum \log x_i}{n}$$

For grouped data

$$\text{Geometric mean} = \text{Antilog } \left[\frac{\sum f \log x_i}{N} \right]$$

Example

Calculate the geometric mean of the following series of monthly income of a bench of families 180, 250, 490, 1400, 1050

x	logx
180	2.2553
250	2.3979
490	2.6902
1400	3.1461
1050	3.0212
	13.5107



$$\begin{aligned}\text{Geometric mean} &= \text{Antilog } \frac{\sum \log x_i}{n} \\ &= \text{Antilog } \frac{13.5107}{5} \\ &= \text{Antilog } 20.7021\end{aligned}$$

Geometric mean = 503.6

Merits of Geometric mean:

1. It is rigidly defined
2. It is based on all items
3. It is very suitable for averaging ratios, rates and percentages
4. It is capable of further mathematical treatment.
5. Unlike AM, it is not affected much by the presence of extreme values.

Demerits of Geometric mean:

1. It cannot be used when the values are negative or if any of the observations is zero
2. It is difficult to calculate particularly when the items are very large or when there is a frequency distribution.
3. It brings out the property of the ratio of the change and not the absolute difference of change as the case in arithmetic mean.
4. The GM may not be the actual value of the series.

HARMONIC MEAN

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If x_1, x_2, \dots, x_n are n observations then

$$\text{Harmonic Mean} = \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i}\right)}$$

For a frequency distribution



$$\text{Harmonic mean} = \frac{N}{\sum_{i=1}^n f\left(\frac{1}{x_i}\right)}$$

Example

From the given data calculate Harmonic mean 5, 10, 17, 24, 30.

Solution:

x	$\frac{1}{x}$
5	0.2000
10	0.1000
17	0.0588
24	0.0417
30	0.0333
	0.4338

$$\begin{aligned}\text{Harmonic Mean} &= \frac{n}{\sum_{i=1}^n \left(\frac{1}{x_i}\right)} \\ &= \frac{5}{0.4338}\end{aligned}$$

Harmonic mean = 11.526

Example

The marks secured by some students of a class are given below. Calculate the harmonic mean.

x	20	21	22	23	24	25
f	4	2	7	1	3	1



Solution:

Marks x	No of students f	$\frac{1}{x}$	$f\left(\frac{1}{x}\right)$
20	4	0.0500	0.2000
21	2	0.0476	0.0952
22	7	0.0454	0.3178
23	1	0.0435	0.435
24	3	0.0417	0.1251
25	1	0.0400	0.0400
	18		0.8216

$$\begin{aligned}\text{Harmonic mean} &= \frac{N}{\sum_{i=1}^n f\left(\frac{1}{x_i}\right)} \\ &= \left(\frac{18}{0.1968}\right)\end{aligned}$$

Harmonic mean = 21.91

Merits of Harmonic mean:

1. It is rigidly defined.
2. It is defined on all observations.
3. It is amenable to further algebraic treatment.
4. It is the most suitable average when it is desired to give greater weight to smaller observations and less weight to the larger ones.

Demerits of Harmonic mean:

1. It is not easily understood.
2. It is difficult to compute.



3. It is only a summary figure and may not be the actual item in the series
4. It gives greater importance to small items and is therefore, useful only when small items have to be given greater weightage.

2.2 PARTITION VALUES

When data is arranged in ascending order, it can be divided into various parts by different values such as quartiles, deciles and percentiles. These values are collectively called quantiles and are the extension of median formula which divides data into two equal parts. Since the basic purpose of these partition values is to divide data into different parts therefore a relationship exists between them.

QUARTILES

The quartiles are the three values that divided the ordered data set in four, perceptually equal, parts. There are three quartiles, usually represented by Q1, Q2, and Q3. The first quartile, Q1, has the lowest value that is greater than a one fourth of the data; that is, the variable's value that is greater than 25% of the observations and is smaller than the 75% of the observations.

The second quartile, Q2, (that coincides, it is identical or similar to the median, $Q2 = M d$), is the lowest value that is greater than half of the data, that is 50% of the observations have a greater value than the median and 50% have a lower value.

The third quartile, Q3, has the lowest value that is greater than three fourths of the data, that is, the value of the variable that has a value greater 75% of the observations and of a lower value than 25% of the observations.

Raw or ungrouped data:

First arrange the given data in the increasing order and use the formula for Q1 and Q3 then quartile deviation, Q.D is given by

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



Where $Q_1 = \left[\frac{n+1}{4} \right]^{\text{th}}$ item and $Q_3 = 3 \left[\frac{n+1}{4} \right]^{\text{th}}$ item

Example

Compute quartiles for the data given below 25, 18, 30, 8, 15, 5, 10, 35, 40, and 45.

Solution:

Now arranging the ascending order 5, 8, 10, 15, 18, 25, 30, 35, 40, 45

$$Q_1 = \left[\frac{n+1}{4} \right]^{\text{th}} \text{ item}$$

$$= \left[\frac{10+1}{4} \right]^{\text{th}} \text{ item}$$

$$= (2.75)^{\text{th}} \text{ item}$$

$$= 2^{\text{nd}} \text{ item} + \frac{3}{4} (3^{\text{rd}} \text{ item} - 2^{\text{nd}} \text{ item})$$

$$= 8 + \frac{3}{4} (10 - 8)$$

$$= 8 + \frac{3}{4} \times 2$$

$$= 8 + 1.5$$

$$Q_1 = 9.5$$

$$Q_3 = 3 \left[\frac{n+1}{4} \right]^{\text{th}} \text{ item}$$

$$= 3 \times (2.75)^{\text{th}} \text{ item}$$

$$= (8.25)^{\text{th}} \text{ item}$$

$$= 8^{\text{th}} \text{ item} + \frac{1}{4} (9^{\text{th}} \text{ item} - 8^{\text{th}} \text{ item})$$

$$= 35 + \frac{1}{4} \times [40 - 35]$$

$$= 35 + 1.25$$

$$Q_3 = 36.25$$



Discrete Series:

Step 1: Find cumulative frequencies.

Step 2: Find $\left[\frac{N+1}{4}\right]$

Step 3: See in the cumulative frequencies, the value just greater than $\left[\frac{N+1}{4}\right]$, and then the corresponding value of x is Q_1

Step 4: Find $3\left[\frac{N+1}{4}\right]$

Step 5: See in the cumulative frequencies, the value just greater than $3\left[\frac{N+1}{4}\right]$, then the corresponding value of x is Q_3 .

Example

Compute quartiles for the data given below.

53	74	82	42	39	20	81	68	58	28
67	54	93	70	30	55	36	37	29	61

Solution:

$$Q_1 = \left[\frac{N+1}{4}\right]^{\text{th}} \text{ item} = \left[\frac{20+1}{4}\right] = \left[\frac{21}{4}\right] = 5.25^{\text{th}} \text{ item}$$

The value of 5th item is 36 and that of the 6th item is 37. Thus the first quartile is a value 0.25th of the way between 36 and 37, which are 36.25.

Therefore, $Q_1 = 36.25$. Similarly,

$$Q_3 = 3\left[\frac{N+1}{4}\right]^{\text{th}} \text{ item} = 3\left[\frac{20+1}{4}\right] = 3\left[\frac{21}{4}\right] = 15.75^{\text{th}} \text{ item}$$

The value of the 15th item is 68 and that of the 16th item is 70. Thus the third quartile is a value 0.75th of the way between 68 and 70. As the difference between 68 and 70 is 2, so the third quartile will be $68 + 2(0.75) = 69.5$. Therefore, $Q_3 = 69.5$.



Continuous series:

Step1: Find cumulative frequencies

Step 2: Find $\left(\frac{N}{4}\right)$

Step3: See in the cumulative frequencies, the value just greater than $\left(\frac{N}{4}\right)$, and then the corresponding class interval is called first quartile class.

Step 4: Find $3 \left[\frac{N}{4}\right]$ see in the cumulative frequencies the value just greater than $3 \left[\frac{N}{4}\right]$ then the corresponding class interval is called 3rd quartile class.

Then apply the respective formulae.

$$Q_1 = l_1 + \frac{\frac{N}{4} \times m_1}{f_1} \times c_1 \qquad Q_3 = l_3 + \frac{3 \left[\frac{N}{4}\right] - m_3}{f_3} \times c_3$$

Where l_1 = lower limit of the first quartile class

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = c.f. preceding the first quartile class

l_3 = lower limit of the 3rd quartile class

f_3 = frequency of the 3rd quartile class

c_3 = width of the 3rd quartile class

m_3 = c.f. preceding the 3rd quartile class

Example

The following series relates to the Monthly per capita expenditure classes.

Find the quartiles.

Monthly per capita expenditure classes (Rs.)	Number of families
140-150	17
150-160	29
160-170	42



170-180	72
180-190	84
190-200	107
200-210	49
210-220	34
220-230	31
230-240	16
240-250	12

Solution:

Monthly per capita expenditure classes (Rs.)	Number of families	Cumulative frequency
140-150	17	17
150-160	29	46
160-170	42	88
170-180	72	160
180-190	84	244
190-200	107	351
200-210	49	400
210-220	34	434
220-230	31	465
230-240	16	481
240-250	12	493

$$\left[\frac{N}{4} \right] = \left[\frac{493}{4} \right] = 123.25 \quad \dots \quad 3 \left[\frac{N}{4} \right] = 369.75$$



$$Q_1 = l_1 + \frac{\frac{N}{4} \times m_1}{f_1} \times c_1$$

$$Q_1 = 170 + \frac{123.25 - 88}{72} \times 10 = 170 + 4.90 = 174.90$$

$$Q_1 = 174.90$$

$$Q_3 = l_3 + \frac{3\left[\frac{N}{4}\right] - m_3}{f_3} \times c_3$$

$$= 200 + \frac{369.75 - 351}{49} \times 10 = 200 + 3.83 = 203.83$$

$$Q_3 = 203.83$$

DECILES

The procedure for locating the i^{th} decile class is to calculate $iN/10$ and search that minimum cumulative frequency in which this value is contained. The class corresponding to this cumulative frequency is i^{th} decile class. The unique value of i^{th} decile can be calculated by the formula.

$$\therefore D_i = l + \frac{i\left[\frac{N}{10}\right] - m}{f} \times c$$

Deciles for Raw data or ungrouped data:

Example

Compute D_5 for the data given below 5, 24, 36, 12, 20, 8.

Solution:

Arranging the given values in the increasing order 5, 8, 12, 20, 24, 36

$$D_5 = \left[\frac{5(n+1)}{10}\right]^{\text{th}} \text{ observation}$$

$$= \left[\frac{5(6+1)}{10}\right]^{\text{th}} \text{ observation}$$

$$= (3.5)^{\text{th}} \text{ observation}$$

$$= 3^{\text{rd}} \text{ item} + \frac{1}{2} [4^{\text{th}} \text{ item} - 3^{\text{rd}} \text{ item}]$$

$$= 12 + \frac{1}{2} [20 - 12]$$



$$= 12+4$$

$$D_5 = 16$$

Deciles for Grouped data:

Example

In a work study investigation, the times taken by 20 men in a firm to do a particular job were tabulated as follows.

Time taken (min)	8-10	11-13	14-16	17-19	20-22	23-25
Frequency	2	4	6	4	3	1

Solution:

x	f	Class Boundaries	c.f
8-10	2	7.5-10.5	2
11-13	4	10.5-13.5	6
14-16	6	13.5-16.5	12
17-19	4	16.5-19.5	16
20-22	3	19.5-22.5	19
23-25	1	22.5-25.5	20
	20		

$$\begin{aligned} D_5 \text{ item} &= \left[\frac{5N}{10} \right]^{\text{th}} \text{ item} \\ &= \left[\frac{5 \times 20}{10} \right]^{\text{th}} \text{ item} \\ &= 10^{\text{th}} \text{ item} \end{aligned}$$

Since the 10th item is in the interval (13.5-16.5)

$$\begin{aligned} \therefore D_i &= l + \frac{i \left[\frac{N}{10} \right] - m}{f} \times c \\ D_5 &= 13.5 + \frac{10 - 6}{6} \times 3 \end{aligned}$$



$$D_5 = 13.5 + 2$$

$$D_5 = 15.5$$

$$\begin{aligned} D_7 \text{ item} &= \left[\frac{7N}{10} \right]^{\text{th}} \text{ item} \\ &= \left[\frac{7 \times 20}{10} \right]^{\text{th}} \text{ item} \\ &= 4^{\text{th}} \text{ item} \end{aligned}$$

Since the 10th item is in the interval (10.5 – 13.5)

$$\begin{aligned} \therefore D_7 &= l + \frac{i \left[\frac{N}{10} \right] - m}{f} \times c \\ D_7 &= 10.5 + \frac{4 - 6}{6} \times 3 \\ D_7 &= 10.5 - 1 \\ D_7 &= 9.5 \end{aligned}$$

PERCENTILES

To locate i^{th} percentile class, calculate $iN/100$ ($i = 1, 2, \dots, 99$) and that minimum cumulative frequency which contains this value. The class corresponding to this minimum cumulative frequency is the percentile class. Unique value of i^{th} percentile can be calculated by the formula,

$$P_i = l + \frac{\left[\frac{iN}{100} \right] - m}{f} \times c$$

Relationship:

$P_{25} = Q_1$; $P_{50} = D_5 = Q_2 = \text{Median}$ and $P_{75} = Q_3$

Percentile for Raw Data or Ungrouped Data:

Example

Calculate P_{15} for the data given below: 5, 24, 36, 12, 20, 8

Solution:

Arranging the given values in the increasing order. 5, 8, 12, 20, 24, 36



$$\begin{aligned}P_{15} &= \left[\frac{15(n+1)}{100} \right]^{\text{th}} \text{ item} \\&= \left[\frac{15 \times 7}{100} \right]^{\text{th}} \text{ item} \\&= (1.05)^{\text{th}} \text{ item} \\&= 1^{\text{st}} \text{ item} + 0.05(2^{\text{nd}} \text{ item} - 1^{\text{st}} \text{ item}) \\&= 5 + 0.05(8 - 5) \\&= 5 + 0.15 \\&= 5.15\end{aligned}$$

Percentile for grouped data:

Example

Find P25, P50 for the following frequency distribution.

Class Interval	f
85.5-90.5	6
90.5-95.5	4
95.5-100.5	10
100.5-105.5	6
105.5-110.5	3
110.5-115.5	1

Solution:

Class Interval	f	C.f
85.5-90.5	6	6
90.5-95.5	4	10
95.5-100.5	10	20



100.5-105.5	6	26
105.5-110.5	3	29
110.5-115.5	1	30
	30	

$$P_i = l + \frac{\left[\frac{iN}{100} \right] - m}{f} \times c$$

$$P_{25} = \frac{iN}{100} = \frac{25 \times 30}{100} = 7.5$$

so, P_{25} group is 90.5–95.5 containing the 7.5th observation

$$\begin{aligned} P_{25} &= 90.5 + \frac{\left[\frac{25 \times 30}{100} \right] - 6}{4} \times 5 \\ &= 90.5 + 1.875 \end{aligned}$$

$$P_{25} = 92.375$$

$$P_{50} = \frac{iN}{100} = \frac{50 \times 30}{100} = 15$$

so, P_{50} group is 95.5–100.5 containing the 15th observation

$$\begin{aligned} P_{50} &= 95.5 + \frac{\left[\frac{50 \times 30}{100} \right] - 10}{10} \times 5 \\ &= 95.5 + 2.5 \end{aligned}$$

$$P_{50} = 98$$

2.3 MEASURES OF DISPERSION

The measure of central tendency serves to locate the centre of the distribution, but they do not reveal how the items are spread out on either side of the centre. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal.



There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates high uniformity of the items, while large dispersion indicates less uniformity.

Characteristics of a good measure of dispersion:

An ideal measure of dispersion is expected to possess the following properties

1. It should be rigidly defined
2. It should be based on all the items.
3. It should not be unduly affected by extreme items.
4. It should lend itself for algebraic manipulation.
5. It should be simple to understand and easy to calculate.

Absolute and Relative Measures:

There are two kinds of measures of dispersion, namely

- Absolute measure of dispersion
- Relative measure of dispersion

Absolute measure of dispersion indicates the amount of variation in a set of values in terms of units of observations. For example, when rainfalls on different days are available in mm, any absolute measure of dispersion gives the variation in rainfall in mm. On the other hand relative measures of dispersion are free from the units of measurements of the observations. They are pure numbers. They are used to compare the variation in two or more sets, which are having different units of measurements of observations. The various absolute and relative measures of dispersion are listed below.

Absolute measure

1. Range
2. Quartile deviation
3. Mean deviation

Relative measure

1. Co-efficient of Range
2. Co-efficient of Quartile deviation
3. Co-efficient of Mean deviation



4. Standard deviation

4. Co-efficient of variation

RANGE

The range is difference between two extreme observations of the distribution. If L and S are the greatest and smallest observations respectively in a distribution, then its range is given by

$$\text{Range} = L - S.$$

Where L = Largest value.

S = Smallest value.

In individual observations and discrete series, L and S are easily identified.

In continuous series, the following two methods are followed.

Method 1:

L = Upper boundary of the highest class

S = Lower boundary of the lowest class.

Method 2

L = Mid value of the highest class.

S = Mid value of the lowest class.

Co-efficient of Range:

$$\text{Co-efficient of Range} = \frac{L-S}{L+S}$$

Example

Find the value of range and its co-efficient range for the following data.

89, 73, 84, 91, 87, 77, 94.

Solution:

73, 77, 84, 87, 89, 91, 94.

L=94, S = 73.

Range = L - S = 94 - 73

Range= 21



$$\begin{aligned}\text{Co-efficient of Range} &= \frac{L-S}{L+S} \\ &= \frac{94-73}{94+73} \\ &= \frac{21}{167}\end{aligned}$$

Co- efficient of range = 0.125

Example

Calculate range and its co efficient from the following distribution.

Size	10-15	15-20	20-25	25-30	30-35
Number	2	8	10	14	8

Solution:

L = Upper boundary of the highest class = 35

S = Lower boundary of the lowest class = 10

$$\text{Range} = L-S = 35-10$$

$$\text{Range} = 15$$

$$\text{Co-efficient of range} = \frac{35-10}{35+10} = \frac{25}{45}$$

$$\text{Co-efficient of range} = 0.555$$

Merits and Demerits of Range:

Merits

1. It is simple to understand.
2. It is easy to calculate.
3. In certain types of problems like quality control, weather forecasts, share price analysis, range is most widely used.

Demerits

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.



3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

MEAN DEVIATION

The measures of dispersion discussed so far are not satisfactory in the sense that they lack most of the requirements of a good measure. Mean deviation is a better measure than range and Quartile deviation.

Definition:

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored. According to Clark and Schekade, "Average deviation is the average amount scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviations". We usually compute mean deviation about any one of the three averages mean, median or mode. Sometimes mode may be ill defined and as such mean deviation is computed from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

Coefficient of mean deviation:

Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

$$\text{Coefficient of mean deviation} = \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}}$$

If the result is desired in percentage, the coefficient of mean

$$\text{Deviation} = \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}} \times 100$$



Example

Calculate Mean deviation from mean and median for the following data, also calculate coefficients of Mean Deviation. 111.5, 111.2, 102.3, 112.4, 108.8, 125.3, 116.5, 132.7

Solution

Arrange the data in ascending order 102.3, 104.6, 108.8, 111.2, 111.5, 112.4, 116.5, 125.3, 132.7

$$\text{Mean} = \bar{X} = \frac{\sum X}{n} = \frac{1025.3}{9} = 113.92$$

$$\text{Mean} = 113.92$$

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

$$= \text{Value of } \left(\frac{9+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Value of } 5^{\text{th}} \text{ item}$$

$$\text{Median} = 111.$$

X	$ D = X - \bar{X} $	$ D = X - Md $
102.3	11.62	9.2
104.6	9.32	6.9
108.8	5.12	2.7
111.2	2.72	0.3
111.5	2.42	0
112.4	1.52	0.9
116.5	2.58	5
125.3	11.38	13.8
132.7	18.78	21.2
1025.3	65.46	60



$$\begin{aligned} \text{M.D from mean} &= \frac{\sum |D|}{n} \\ &= \frac{65.46}{9} \end{aligned}$$

$$\text{M.D from mean} = 7.27$$

$$\text{Co-efficient of MD from mean} = \frac{M.D}{\bar{x}} = \frac{7.27}{113.92}$$

$$\text{Co-efficient of MD} = 0.063$$

$$\begin{aligned} \text{MD from median} &= \frac{\sum |D|}{n} \\ &= \frac{60}{9} = 6.66 \end{aligned}$$

$$\text{Co-efficient of MD from median} = \frac{M.D}{\text{median}} = \frac{6.66}{111.5}$$

$$\text{Co-efficient of M.D} = 0.059$$

Mean Deviation – Discrete series:

Steps:

1. Find out an average (mean, median or mode)
2. Find out the deviation of the variable values from the average, ignoring signs and denote them by $|D|$
3. Multiply the deviation of each value by its respective frequency and find out the total $\sum f|D|$
4. Divide $\sum f|D|$ by the total frequencies N Symbolically, MD from median = $\frac{\sum f|D|}{N}$

Example

Compute Mean deviation from mean and median from the following data:

Height (in cm)	158	159	160	161	162	163	164	165	166
No. of persons	15	20	32	35	33	22	20	10	8



Also compute coefficient of mean deviation.

Solution:

Height X	No. of persons	D=x-A A=162	fd	D = X – Md	f D
158	15	-4	-60	3.51	52.65
159	20	-3	-60	2.51	50.20
160	32	-2	-64	1.51	48.32
161	35	-1	-35	0.51	17.85
162	33	0	0	0.49	16.17
163	22	1	22	1.49	32.78
164	20	2	40	2.49	49.80
165	10	3	30	3.49	34.90
166	8	4	32	4.49	35.92
	195		-95		338.59

$$\bar{X} = A + \frac{\sum f|D|}{N}$$

$$\bar{X} = 162 + \frac{-95}{195}$$

$$= 162 - 0.49$$

$$= 161.51$$

$$\text{MD from median} = \frac{\sum f|D|}{N}$$

$$= \frac{338.59}{195}$$

$$= 1.74$$

$$\text{Co-efficient of MD} = \frac{M.D}{\bar{X}}$$



$$= \frac{1.74}{161.51}$$

$$= 0.0108$$

Height X	No. of Persons f	C.f	D = X - Median	f D
158	15	15	3	45
159	20	35	2	40
160	32	67	1	32
161	35	102	0	0
162	33	135	1	33
163	22	157	2	44
164	20	177	3	60
165	10	187	4	40
166	8	195	5	40
	195			334

Median = Size of $\left(\frac{N+1}{2}\right)$ th item

= Size of $\left(\frac{195+1}{2}\right)$ th item

= Size of 98th item

MD from median = $\frac{\sum f|D|}{N}$

$$= \frac{334}{195} = 1.71$$

Co-efficient = $\frac{M.D}{\text{Median}}$

$$= \frac{1.71}{161} = 0.0106$$

Co-efficient = 0.0106



Mean deviation-Continuous series:

The method of calculating mean deviation in a continuous series same as the discrete series. In continuous series we have to find out the mid points of the various classes and take deviation of these points from the average selected. Thus

$$\text{MD from median} = \frac{\sum f|D|}{N}$$

$$D = m - \text{average}$$

$$M = \text{Mid point}$$

Example

Find out the mean deviation from mean and median from the following series.

Age (in years)	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of Persons	20	25	32	40	42	35	10	8

Solution:

X	m	f	$d = \frac{m - A}{c}$ (A=35, C=10)	fd	$ D = m - \bar{X} $	$f D $
0-10	20	20	-3	-60	31.5	630
10-20	25	25	-2	-50	21.5	537.5
20-30	32	32	-1	-32	11.5	368
30-40	40	40	0	0	1.5	60
40-50	42	42	1	42	8.5	357



50-60	35	35	2	70	18.5	647
60-70	10	10	3	30	28.5	285
70-80	8	8	4	32	38.5	308
				32		3193

$$\bar{X} = A + \frac{\sum fd}{N} \times c$$

$$= 35 + \frac{320}{212} \times 10$$

$$= 35 + \frac{320}{212} = 35 + 1.5 = 36.5$$

$$\text{M.D} = \frac{\sum f|D|}{N} = \frac{3193}{212} = 15.06$$

Mean deviation = 15.06

Calculation of median and M.D from median

X	m	f	cf	D = m - d	f D
0-10	20	20	20	32.25	645
10-20	25	25	45	22.25	556.25
20-30	32	32	77	12.25	392
30-40	40	40	117	2.25	90
40-50	42	42	159	7.75	325.50
50-60	35	35	194	17.75	621.25
60-70	10	10	204	27.75	277.50
70-80	8	8	212	37.75	302
				Total	3209.50



$$\frac{N}{2} = \frac{212}{2} = 106$$

$$l = 30, m = 77, f = 40, c = 10$$

$$\begin{aligned}\text{Median} &= l + \frac{\frac{N}{2} - m}{f} \times c \\ &= 30 + \frac{106 - 77}{40} \times 10 \\ &= 30 + \frac{29}{4} \\ &= 30 + 7.25 \\ &= 37.25\end{aligned}$$

$$\begin{aligned}\text{M.D} &= \frac{\sum f|D|}{N} \\ &= \frac{3209.5}{212} \\ &= 15.14\end{aligned}$$

$$\begin{aligned}\text{Coefficient of M.D} &= \frac{\text{M.D}}{\text{Median}} \\ &= \frac{15.14}{37.25} = 0.41\end{aligned}$$

Co efficient mean deviation = 0.41

Merits and Demerits of Mean Deviation:

Merits

1. It is simple to understand and easy to compute.
2. It is based on all the observations of series.
3. It is not very much affected by the fluctuations of sampling.
4. It is less affected by the extreme items.
5. It is flexible, because it can be calculated from any average.



6. It facilitates comparison between different items of a series.
7. It has practical usefulness in the field of business and commerce.

Demerits

1. It is not a very accurate measure of dispersion.
2. It is not suitable for further mathematical calculation.
3. It is rarely used. It is not as popular as standard deviation.
4. Algebraic positive and negative signs are ignored.
5. It is not suitable for sociological study.

QUARTILE DEVIATION (Q.D):

Co-efficient of quartile deviation is an absolute quantity and is useful to compare the variability among the middle 50% observations. The quartiles are the values which divide the whole distribution into four equal parts.

Definition:

The interquartile range, half of the difference between the first and third quartiles interquartile range that is Quartile deviation. Hence, it is called Semi Inter Quartile Range.

In Symbols, $Q.D = \frac{Q_3 - Q_1}{2}$. Among the quartiles Q_1 , Q_2 and Q_3 , the range $Q_3 - Q_1$ is called inter quartile range and $\frac{Q_3 - Q_1}{2}$, Semi inter quartile range.

Co-efficient of Quartile Deviation:

$$\text{Co-efficient of Q.D} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Example

The wheat production (in Kg) of 20 acres is given as: 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, and 1885. Find the quartile deviation and coefficient of quartile deviation.



Solution:

Arranging the observations in ascending order

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$Q_1 \text{ is } \frac{n+1}{4}$$

$$Q_1 = \frac{20+1}{4}$$

$$Q_1 = 5.25^{\text{th}} \text{ value}$$

$$Q_1 = 5^{\text{th}} \text{ value} + 0.25 (6^{\text{th}} \text{ value} - 5^{\text{th}} \text{ value})$$

$$= 1240 + 0.25(1320 - 1240)$$

$$= 1240 + 20$$

$$Q_1 = 1260$$

$$Q_3 \text{ is } \frac{3(n+1)}{4}$$

$$Q_3 = \frac{3(20+1)}{4} = \frac{3(20+1)}{4} = 15.75^{\text{th}} \text{ value}$$

$$Q_3 = 15^{\text{th}} \text{ value} + 0.75 (6^{\text{th}} \text{ value} - 5^{\text{th}} \text{ value})$$

$$Q_3 = 1750 + 0.75 (1755 - 1750)$$

$$Q_3 = 1753.75.$$


$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{1753.75 - 1260}{2}$$

$$= \frac{492.75}{2}$$

$$\text{Quartile deviation} = 246.875$$

$$\text{Co-efficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$



$$= \frac{1753.75 - 1260}{1753.75 + 1260}$$

Co-efficient of Quartile deviation = 0.164

Example

Calculate the quartile deviation and coefficient of quartile deviation from the data given below

Maximum Weight (Short tonnes)	Number of Packets
9.25-9.75	2
9.75-10.25	5
10.25-10.75	12
10.75-11.25	17
11.25-11.75	14
11.75-12.25	6
12.25-12.75	3
12.75-13.25	1

Solution:

Maximum Weight (Short tonnes)	Number of Packets	C.f
9.25-9.75	2	2
9.75-10.25	5	7
10.25-10.75	12	19
10.75-11.25	17	36
11.25-11.75	14	50
11.75-12.25	6	56
12.25-12.75	3	59



12.75-13.25	1	60
-------------	---	----

Position of $Q_1 = \frac{N+1}{4} = \frac{60+1}{4} = 15.25^{\text{th}}$ item the class boundaries is 10.25-10.75

$$Q_1 = 15^{\text{th}} \text{ value} + 0.25(16^{\text{th}} \text{ value} - 15^{\text{th}} \text{ value})$$

$$= 10.25 + 0.25(10.75-10.25)$$

$$= 10.25+0.25(0.50)$$

$$Q_1 = 10.40$$

Q_3 is $\frac{N+1}{4} = 3 \times 15.25 = 45.75^{\text{th}}$ item the class boundaries is 11.25 – 11.75

$$= 45^{\text{th}} \text{ value} + 0.75(46^{\text{th}} \text{ value} - 45^{\text{th}} \text{ value})$$

$$= 11.25^{\text{th}} \text{ value} + 0.75(11.75^{\text{th}} \text{ value} - 11.25^{\text{th}} \text{ value})$$

$$= 11.25+0.375Q_3$$

$$=11.62$$

$$\begin{aligned} \text{Quartile deviation} &= \frac{Q_3 - Q_1}{2} \\ &= \frac{11.62 - 10.40}{2} \end{aligned}$$

$$\text{Quartile deviation} = 0.61$$

$$\begin{aligned} \text{Co-efficient of Quartile deviation} &= \frac{Q_3 - Q_1}{Q_3 + Q_1} \\ &= \frac{11.62 - 10.40}{11.62 + 10.40} \\ &= \frac{1.22}{22.02} \end{aligned}$$

$$\text{Co-efficient of Quartile deviation} = 0.055$$

Example

For the data give below, give the quartile deviation and Coefficient of quartile deviation.

x	351-500	501-650	651-800	801-950	951-1100
f	48	189	88	4	28



Solution:

x	f	True class intervals	Cumulative frequency
351-500	48	350.5-500.5	48
501-650	189	500.5-650.5	237
651-800	88	650.5-800.5	325
801-950	4	800.5-950.5	372
951-1100	28	950.5-110.5	400
Total	N=400		

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times C_1$$

$$\frac{N}{4} = \frac{400}{4} = 100,$$

Q_1 Class is 500.5-650.5

$$l_1 = 500.5, m_1 = 48, f_1 = 189, C_1 = 150$$

$$Q_1 = 500 + \frac{100 - 48}{189} \times 150$$

$$= 500.5 + \frac{52}{189} \times 150$$

$$= 500.5 + 41.27$$

$$Q_1 = 541.77$$

$$Q_3 = l_3 + \frac{3\frac{N}{4} - m_3}{f_3} \times C_3$$

$$3\frac{N}{4} = 3 \times 100 = 300,$$

Q_3 Class is 650.5-800.5

$$l_3 = 650.5, m_3 = 237, f_3 = 88, C_3 = 150$$

$$Q_3 = l_3 + \frac{3\frac{N}{4} - m_3}{f_3} \times C_3$$

$$= 650.5 + \frac{300 - 237}{88} \times 150$$



$$= 650.5 + \frac{63}{88} \times 150$$

$$= 650.5 + 107.39$$

$$Q_3 = 757.89$$

$$\text{Quartile Deviation} = \frac{Q_3 - Q_1}{2}$$

$$= \frac{757.89 - 541.77}{2}$$

$$= \frac{216.12}{2}$$

$$\text{Quartile Deviation} = 108.06$$

$$\text{Co-efficient of Quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

$$= \frac{757.89 - 541.77}{757.89 + 541.77}$$

$$= \frac{216.12}{1299.66}$$

$$\text{Co-efficient of Quartile deviation} = 0.1663$$

Merits and demerits of Quartile Deviation:

Merits

1. It is Simple to understand and easy to calculate
2. It is not affected by extreme values.
3. It can be calculated for data with open end classes also.

Demerits

1. It is not based on all the items. It is based on two positional values Q1 and Q3 and ignores the extreme 50% of the items.
2. It is not amenable to further mathematical treatment.
3. It is affected by sampling fluctuations.



STANDARD DEVIATION

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also called Root-Mean Square Deviation. The reason is that it is the square-root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by the Greek letter σ (sigma).

Calculation of Standard deviation-Individual Series:

There are two methods of calculating Standard deviation in an individual series.

- a) Deviations taken from Actual mean
- b) Deviation taken from Assumed mean

a) Deviation taken from Actual mean:

This method is adopted when the mean is a whole number.

$$\sigma = \sqrt{\left(\frac{\sum x^2}{n}\right)} \text{ or } \sqrt{\left(\frac{\sum (x - \bar{x})^2}{n}\right)}$$

b) Deviations taken from assumed mean:

This method is adopted when the arithmetic mean is fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, we apply short-cut method; deviations are taken from an assumed mean. The formula is:

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



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

Note: We can also use the simplified formula for standard deviation.

$$\sigma = \frac{1}{n} \sqrt{n \sum d^2 - (\sum d)^2}$$

For the frequency distribution

$$\sigma = \frac{c}{N} \sqrt{N \sum fd^2 - (\sum fd)^2}$$

Example

Calculate the standard deviation from the following data 6, 2, 3, 1.

Solution:

Find the square of the distance from each data point to the mean.

x	$X - \bar{X}$	$(X - \bar{X})^2$
6	3	9
2	-1	1
3	0	0
1	2	4
12		14

$$\bar{X} = \frac{12}{4} = 3$$

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n}}$$

$$\sigma = \sqrt{\frac{14}{4}} = \sqrt{3.5}$$

Standard deviation = 1.87



Example

Calculate the standard deviation, for the following data.

Student	1	2	3	4	5	6
Marks	2	4	6	8	10	12

Solution: (Deviations from assumed mean)

No.	Marks (x)	d= x-A x = 6	d ²
1	2	-4	16
2	4	-2	4
3	6	2	4
4	8	0	0
5	10	4	16
6	12	6	36
n=10		∑d= 6	∑d ² =76

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

$$\sigma = \sqrt{\frac{76}{6} - \left(\frac{6}{6}\right)^2}$$

$$= \sqrt{12.66}$$

$$\sigma = 3.5$$

Calculation of standard deviation: Discrete Series:

There are three methods for calculating standard deviation in discrete series:

- Actual mean methods



- Assumed mean method
- Step-deviation method.

Merits and Demerits of Standard Deviation:

Merits

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is capable of further algebraic treatment as it has a lot of algebraic properties.
5. It is less affected by the fluctuations of sampling and hence stable.
6. It is the basis for measuring the coefficient of correlation and sampling.
7. It can be calculated through a good number of methods yielding the same results.
8. It has a good number of algebraic properties for which it is possible to determine the number of many connected factors like combined standard deviation of two or more series.

Demerits

1. It is not easy to understand and it is difficult to calculate.
2. It gives more weight to extreme values because the values are squared up.
3. As it is an absolute measure of variability, it cannot be used for the purpose of comparison.
4. It cannot be used for comparing the dispersion of two or more series given in different units.



COEFFICIENT OF VARIATION

All the measures of dispersion discussed so far have units. If two series differ in their units of measurement, their variability cannot be compared by any measure given so far. Also, the size of measure of dispersion depends upon the size of values. Hence in situations where either the two series have different units of measurements, or their means differ sufficiently in size, the coefficient of variation should be used as a measure of dispersion and also takes into account size of the means of the two series. It is the best measure to compare the variability of two series or sets of observations. A series with less coefficient of variation is considered more consistent.

Definition:

Coefficient of variation of a series of variate values is the ratio of the standard deviation to the mean multiplied by 100.

$$\text{Coefficient of Variation (C.V)} = \frac{\sigma}{\bar{x}} \times 100$$

$\sigma = \text{Standard deviation}$

$\bar{x} = \text{mean}$

Example

Calculate the Coefficient of variation for the following data.

Prices of Rice	No. of Centres
1.75	3
1.72	2
1.73	4
1.76	5
1.71	6
1.80	2
1.87	7
2.34	1



Solution:

Prices of Rice (x)	No. of Centres (f)	fx	$x - \bar{x}$	$f(x - \bar{x})^2$	$f(x - \bar{x})$
1.75	3	5.25	-0.04	-0.12	0.0048
1.72	2	3.44	-0.07	-0.14	0.0098
1.73	4	6.92	-0.06	-0.24	0.0144
1.76	5	8.80	-0.03	-0.15	0.0045
1.71	6	10.26	-0.08	-0.48	0.0384
1.80	2	3.60	0.01	0.02	0.0002
1.87	7	13.09	0.08	0.56	0.0448
2.34	1	2.34	0.055	0.55	0.3025
	30	53.70			0.4194

$$\bar{x} = \frac{\sum fx}{n}$$

$$\bar{x} = \frac{53.70}{30}$$

$$\bar{x} = 1.79$$

$$\sigma = \sqrt{\frac{\sum f(X - \bar{X})^2}{n}}$$

$$\sigma = \sqrt{\frac{0.4194}{30}}$$

$$\sigma = 0.1182$$

$$\text{Coefficient of Variation (C.V)} = \frac{\sigma}{\bar{x}} \times 100$$

$\sigma = \text{Standard deviation}$

$\bar{x} = \text{mean}$



$$\text{Coefficient of Variation} = \frac{0.1182}{1.79} \times 100$$

$$\text{Coefficient of Variation} = 6.603$$

Example

Prices of a particular commodity in five years in two cities are given below:

Price in city A	Price in city B
20	10
22	20
19	18
23	12
16	15

Which city has more stable prices?

Solution:

Actual mean method:

City A			City B		
Prices (X)	Deviations from $\bar{X}=20$ dX	dX ²	Prices (Y)	Deviations from $\bar{Y}=20$ dY	dY ²
20	0	0	10	-5	25
22	2	4	20	5	25
19	-1	1	18	3	9
23	3	9	12	-3	9
16	-4	16	15	0	0
100	0	30	75	0	68



$$\text{City A: } \bar{X} = \frac{\sum X}{n} = \frac{100}{5} = 20$$

$$\sigma = \sqrt{\frac{\sum (X - \bar{X})^2}{n}} = \sqrt{\frac{\sum dX^2}{n}}$$

$$= \sqrt{\frac{30}{5}} = 2.45$$

$$\text{C.V(X)} = \frac{\sigma_x}{\bar{x}} \times 100$$

$$= \frac{2.45}{20} \times 100$$

$$= 12.25 \%$$

$$\text{City B: } \bar{Y} = \frac{\sum Y}{n} = \frac{75}{5} = 15$$

$$\sigma = \sqrt{\frac{\sum (Y - \bar{Y})^2}{n}} = \sqrt{\frac{\sum dY^2}{n}}$$

$$= \sqrt{\frac{68}{5}} = 3.69$$

$$\text{C.V(X)} = \frac{\sigma_y}{\bar{y}} \times 100$$

$$= \frac{3.69}{15} \times 100$$

$$= 24.6 \%$$

City A had more stable prices than City B, because the coefficient of variation is less in City A.



QUESTIONS

1. What are the desirable characteristics of a good measurable central tendency?
2. Define Mean (or) Arithmetic mean.
3. Find the mean 6, 8, 11, 5, 2, 9, 7, and 8.
4. A student's marks in 5 subjects are 75, 68, 80, 92, and 56. Find his average mark.
5. Calculate the mean for the following data.

X	5	8	12	15	20	24
F	3	4	6	5	3	2

6. Given the following frequency distribution. Calculate the arithmetic mean.

Marks	64	63	62	61	60	59
No. of Students	8	18	12	9	7	6

7. Calculate the Arithmetic mean.

Income	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of persons	6	8	10	12	7	4	3

8. Give the merits and demerits of Arithmetic mean.
9. Define Median.



10. Determine median following data?

i) 8, 10, 18, 20, 25, 27, 30, 42, 53.

ii) 25, 20, 15, 45, 18, 17, 10, 38, 12.

11. Find median for the following data. 5, 8, 12, 30, 18, 10, 2, 22.

12. Find median size of the family.

No. of Member (x)	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	3	5	6	10	13	9	5	3	2	2	1

13. Define Mode?

14. Find the mode. 2, 8, 5, 12, 5, 7, 1, 10, 5, 6.

15. Describe the merits and demerits of mode?

16. Describe the merits and demerits of Geometric mean?

17. Describe the merits and demerits of harmonic mean?

18. Compute quartiles for the data given below. 25, 18, 30, 8, 15, 5, 10, 35, 40 and 45.

19. Compute quartiles for the data given below.

X	5	8	12	15	19	24	30
F	4	3	2	4	5	2	4

20. Explain the characteristics of a good measure of dispersion

21. Define range and Co-efficient range.

22. Give the merits and demerits of range.

23. Define mean deviation and Co-efficient mean deviation.

24. Explain merits and demerits of mean deviation.

25. Define Quartile deviation and Quartile deviation.

26. Explain merits and demerits of Quartile deviation.

27. Define standard deviation and Co-efficient standard deviation.



28. Explain merits and demerits of standard deviation.

29. Compute quartile deviation from the following data.

X	58	59	60	61	62	63	64	65	66
F	15	20	32	35	33	22	20	10	8

30. Find the median of the following data.

Wages (In Rs.)	60-70	50-60	40-50	30-40	20-30
Number of Workers	7	21	11	6	5

31. Find the Median the following table represents the marks obtained by a batch of 10 students in certain class tests in statistics and accountancy?

S. No	1	2	3	4	5	6	7	8	9	10
Marks (Statistics)	53	55	52	32	30	60	47	46	35	28
Marks (Accountancy)	57	45	24	31	25	84	43	80	32	72

32. The table below gives the relative frequency distribution of annual pay roll for 100 small retail establishments in a city.

Annual pay roll (1000 rupees)	Establishments
Less than 10	8
10 and Less than 20	12
20 and Less than 30	18
30 and Less than 40	30
40 and Less than 50	20
50 and Less than 60	12
	100



Calculate Median pay.

33. Calculate the median from the data given below.

Wages (in Rs.)	Above 30	Above 40	Above 50	Above 60	Above 70	Above 80	Above 90
No. of Workers	520	470	399	210	105	45	7

34. Calculate median from the following data

Value	0-4	5-9	10- 14	15- 19	20- 24	25- 29	30- 34	35- 39
Frequency	5	8	10	12	7	6	3	2

35. Find median for the data given below.

Marks	Number of Students
Greater than 10	70
Greater than 20	62
Greater than 30	50
Greater than 40	38
Greater than 50	30
Greater than 60	24
Greater than 70	17
Greater than 80	9
Greater than 90	4

36. Find the mode for the following data.

Class Interval	0-5	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	9	12	15	16	17	15	10	13



37. Find the mode following distribution

X	1	2	3	4	5	6	7	8	9	10	11	12
F	3	8	15	23	35	40	32	28	40	45	14	6

39. Calculate the geometric mean of the following series of monthly income of a batch of families 180, 250, 490, 1400, 1050.

40. The marks secured by some students of a class are given below.

Calculate the harmonic mean.

Marks	20	21	22	23	24	25
Number of Students	4	2	7	1	3	1

41. Calculate D3 and D7 for the data given below

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70
f	5	7	12	16	10	8	4

42. Calculate mean, median, mode for the following data given below.

X	10	20	30	40	50	60	70	80	90	100
Cumulative Frequency	140	133	118	100	75	45	25	9	20	0

43. Calculate mean deviation from mean and median for the following data, 100, 150, 200, 250, 360, 490, 500, 600, 671 also calculate coefficients of mean deviation.

44. Calculate the standard deviation of the following data.

Size	6	7	8	9	10	11	12
Frequency	3	6	9	13	8	5	4



UNIT -III

3.1 MOMENTS

Moments can be defined as the arithmetic mean of various powers of deviations taken from the mean of a distribution. These moments are known as central moments. The first four moments about arithmetic mean or central moments are defined below.

	Individual series	Discrete series
First moments about the mean; μ_1	$\frac{\sum(X - \bar{X})}{n} = 0$	$\frac{\sum f(X - \bar{X})}{N} = 0$
Second moments about the mean; μ_2	$\frac{\sum(X - \bar{X})^2}{n} = \sigma^2$	$\frac{\sum f(X - \bar{X})^2}{N}$
Third moments about the mean; μ_3	$\frac{\sum(X - \bar{X})^3}{n}$	$\frac{\sum f(X - \bar{X})^3}{N}$
Fourth moments about the mean; μ_4	$\frac{\sum(X - \bar{X})^4}{n}$	$\frac{\sum f(X - \bar{X})^4}{N}$

μ is a Greek letter, pronounced as 'mu'.

If the mean is a fractional value, then it becomes a difficult task to work out the moments. In such cases, we can calculate moments about a working origin and then change it into moments about the actual mean. The moments about an origin are known as raw moments.



$$\mu_1 = \frac{\sum(X-A)}{N} = \frac{\sum d}{N} \qquad \mu_2 = \frac{\sum(X-A)^2}{N} = \frac{\sum d^2}{N}$$

$$\mu_3 = \frac{\sum(X-A)^3}{N} = \frac{\sum d^3}{N} \qquad \mu_4 = \frac{\sum(X-A)^4}{N} = \frac{\sum d^4}{N}$$

A – Any origin,

$$d=X-A$$

The first four raw moments – Discrete series (step – deviation method)

$$\mu_1' = \frac{\sum f d'}{N} \times C \qquad \mu_2' = \frac{\sum f d'^2}{N} \times C^2$$

$$\mu_3' = \frac{\sum f d'^3}{N} \times C^3 \qquad \mu_4' = \frac{\sum f d'^4}{N} \times C^4$$

Where $d = \frac{X-A}{C}$, A- origin, C- Common Point. The first four raw moments – Continuous Series.

$$\mu_1' = \frac{\sum f d'}{N} \times C \qquad \mu_2' = \frac{\sum f d'^2}{N} \times C^2$$

$$\mu_3' = \frac{\sum f d'^3}{N} \times C^3 \qquad \mu_4' = \frac{\sum f d'^4}{N} \times C^4$$

Where $d = \frac{X-A}{c}$, A- origin, C- Class interval.

Relationship between Raw Moments and Central moments:

The raw moments (or 'moments about zero') μ_i' of a distribution are defined as

$$\mu_i' = \int_{-\infty}^{+\infty} x^i f(x) dx$$

for continuous distributions with Probability distribution function f(x) and

$$\mu_i' = \sum_{k=0}^{\infty} x_k^i p_k$$



for discrete distributions with Probability mass function p_i . The central moments (or 'moments about the mean') μ_i for $i \geq 2$ are defined as:

$$\mu_i = \int_{-\infty}^{+\infty} (x - \mu)^i f(x) dx$$

with analogue definitions for discrete variables. The lower central moments are directly related to the variance, skewness and kurtosis. The second, third and fourth central moments can be expressed in terms of the raw moments as follows:

$$\mu_1 = \mu_1 - \mu_1 = 0$$

$$\mu_2 = \mu_2 - \mu_1^2$$

$$\mu_3 = \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3$$

$$\mu_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$$

Example

Calculate first four moments from the following data.

x	0	1	2	3	4	5	6	7	8
f	1	8	28	56	70	56	28	8	1

Solution:

x	f	$d = X - m$	fd	fd ²	fd ³	fd ⁴
0	1	-4	-4	16	-64	256
1	8	-3	-24	72	-216	648
2	28	-2	-56	112	-224	448
3	56	-1	-56	56	-56	56
4	70	0	0	0	0	0
5	56	1	56	56	56	56
6	28	2	56	112	224	448
7	8	3	24	72	216	648
8	1	4	4	16	64	256



	256	0	0	512	0	2816
--	-----	---	---	-----	---	------

Moments about the midpoint $x = 4$

$$\mu_1 = \frac{\sum fd}{N} = \frac{0}{256} = 0$$

$$\mu_2 = \frac{\sum fd^2}{N} = \frac{512}{256} = 2$$

$$\mu_3 = \frac{\sum fd^3}{N} = \frac{0}{256} = 0$$

$$\mu_4 = \frac{\sum fd^4}{N} = \frac{2816}{256} = 11$$

Moments about the mean

$$\mu_1 = 0$$

$$\mu_2 = \mu_2 - \mu_1^2$$

$$= 2 - (0)^2$$

$$= 2$$

$$\mu_3 = \mu_3 - 3\mu_1\mu_2 + 2(\mu_1)^3$$

$$= 0 - 3(0)(2) + 2(0)^3$$

$$= 0$$

$$\mu_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$$

$$= 11 - 4(0)(0) + 6(2)(0)^2 - 3(0)^4$$

$$= 11$$

Example

From the data given below, first calculate the four moments about an arbitrary origin and then calculate the first four moments about the mean.

x	30-33	33-36	36-39	39-42	42-45	42-45
f	2	4	26	47	15	6



Solution

X	Mid values (m)	f	$d = \frac{m - A}{C}$ A = 37.5	fd	fd ²	fd ³	fd ⁴
30-33	31.5	2	-2	-4	8	-16	32
33-36	34.5	4	-1	-4	4	-4	4
36-39	37.5	26	0	0	0	0	0
39-42	40.5	47	1	47	47	47	47
42-45	43.5	15	2	30	60	120	240
45-48	46.5	6	3	18	54	162	486
		N=100		87	173	309	809

$$\mu_1' = \frac{\sum fd'}{N} \times C = \frac{87}{100} \times C = \frac{261}{100} = 2.61$$

$$\mu_2' = \frac{\sum fd'^2}{N} \times C^2 = \frac{173}{100} \times 9 = \frac{1557}{100} = 15.57$$

$$\mu_3' = \frac{\sum fd'^3}{N} \times C^3 = \frac{309}{100} \times 27 = \frac{8343}{100} = 83.43$$

$$\mu_4' = \frac{\sum fd'^4}{N} \times C^4 = \frac{809}{100} \times 81 = \frac{65529}{100} = 655.29$$

Moments about mean

$$\mu_1 = 0$$

$$\mu_2 = \mu_2' - \mu_1'^2$$

$$= 15.57 - (2.61)^2$$

$$= 15.57 - 6.81$$

$$= 8.76$$

$$\mu_3 = \mu_3' - 3\mu_1'\mu_2' + 2(\mu_1')^3$$

$$= 83.43 - 3(2.61)(15.57) + 2(2.61)^3$$



$$= 83.43 - 121.9 + 35.56 = -2.91$$

$$\mu_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4$$

$$= 665.29 - 4(83.43)(2.61) + 6(15.57)(2.61)^2 - 3(2.61)^4$$

$$= 665.29 - 871.01 + 636.39 - 139.214$$

$$= 291.454$$

3.2 MEASURES OF SKEWNESS

The important measures of skewness consisting of absolute and relative measures. The relative measure is called the coefficient of skewness. The signs of the skewness indicate whether the distribution is positively skewed or negatively.

- Karl – Pearson's coefficient of skewness.
- Bowley's coefficient of skewness.
- Measure of skewness based on moments.

3.3 PEARSON'S AND BOWLEY'S COEFFICIENT OF SKEWNESS

Karl – Pearson's Coefficient of skewness:

According to Karl – Pearson, the absolute measure of skewness = mean – mode. This measure is not suitable for making valid comparison of the skewness in two or more distributions because the unit of measurement may be different in different series. To avoid this difficulty use relative measure of skewness called Karl – Pearson's coefficient of skewness given by:

$$\text{Karl – Pearson's Coefficient Skewness} = \frac{\text{Mean} - \text{Mode}}{S.D.}$$

In case of mode is ill – defined, the coefficient can be determined by the formula:

$$\text{Coefficient of skewness} = \frac{3(\text{Mean} - \text{Median})}{S.D.}$$



Example

Find the coefficient of skewness from the data given below.

Size	30	40	50	60	70	80	90	100
Frequency	7	10	14	35	102	136	43	8

Solution:

$$A = 10, C = 10$$

x	frequency (f)	$dx = \frac{(x - A)}{c}$	fdx	fd ² x
30	7	4	-28	112
40	10	-3	-30	90
50	14	-2	-28	56
60	35	-1	-35	35
70	102	0	0	0
80	136	1	136	136
90	43	2	86	172
100	8	3	24	72
	335		125	673

$$\sigma = \sqrt{\frac{\sum fd^2x}{n} - \left(\frac{\sum fdx}{n}\right)^2} \times c$$

$$\sigma = \sqrt{\frac{673}{355} - \left(\frac{125}{355}\right)^2} \times 10$$

$$\sigma = \sqrt{1.895 - 0.1239} \times 10$$

$$\sigma = \sqrt{1.33} \times 10$$

$$\sigma = 13.3$$



Maximum frequency is 136 hence the value of mode is 80

$$\text{mean} = A + \frac{\sum f dx}{n} \times c$$

$$= 70 + \frac{125}{355} \times 10$$

$$\text{mean} = 73.52$$

$$\text{Karl Pearson coefficient Skewness} = \frac{(\text{Mean} - \text{Mode})}{S.D.}$$

$$= \frac{(73.52 - 80)}{13.3}$$

$$\text{Karl Pearson coefficient Skewness} = 0.487$$

Example

Calculate the Karl Pearson's coefficient of skewness for the following data.

Variable	0-5	5-10	10-15	15-20	20-25	25-30	30-5	35-40
No. of Workers	2	5	7	13	21	16	8	3

Solution

$$\text{Karl Pearson coefficient Skewness} = \frac{(\text{Mean} - \text{Mode})}{S.D.}$$

$$A = 175, C = 5$$

x	f	Mid x	dx	fdx	fd ² x
0-5	2	2.5	-3	-6	18
5-10	5	7.5	-2	-10	20
10-15	7	12.5	-1	-7	7
15-20	13	17.5	0	0	0
20-25	21	22.5	1	21	21
25-30	16	27.5	2	32	64



30-35	8	32.5	3	24	72
35-40	3	37.5	4	12	48
	75			66	250

$$\text{Mode} = L + \frac{f_0 - f_1}{2f_1 - f_0 - f_2} 21 - (L_2 - L_1)$$

$$= 20 + \frac{21-13}{2(21) - 13 - 6} (25 - 20)$$

$$\text{Mode} = 23.07$$

$$\text{mean} = A + \frac{\sum fdx}{n} \times c$$

$$= 17.5 + \frac{66}{75} \times 5$$

$$= 17.5 + 4.4$$

$$\text{Mean} = 21.9$$

$$\sigma = \sqrt{\frac{\sum fd^2x}{n} - \left(\frac{\sum fdx}{n}\right)^2} \times c$$

$$\sigma = \sqrt{\frac{250}{75} - \left(\frac{250}{n75}\right)^2} \times 5$$

$$\sigma = \sqrt{3.33 - 0.7744} \times 5$$

$$\sigma = \sqrt{2.5556} \times 5$$

$$\sigma = 7.993$$

$$\text{Karl Pearson coefficient Skewness} = \frac{(\text{Mean} - \text{Mode})}{S.D.}$$

$$= \frac{21.9 - 23.07}{7.993}$$

$$\text{Karl Pearson coefficient Skewness} = 0.1463$$



Bowley's Coefficient of skewness:

In Karl – Pearson's method of measuring skewness the whole of the series is needed. Prof. Bowley has suggested a formula based on relative position of quartiles. In a symmetrical, the quartiles are equidistant from the value of the mean i.e., Median $-Q_1 = Q_3$ -Median. But in a skewed distribution, the quartiles will not be equidistant from the median. Hence Bowley has suggested the following formula

$$\text{Bowley's Coefficient of skewness (sk)} = \frac{\text{sum of two quartiles} - 2\text{median}}{\text{different of two quartiles}}$$

$$\text{Bowley's Coefficient of skewness (sk)} = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

Example

Find the Bowley's coefficient of skewness for the following series. 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22

Solution:

The given data in order 2, 4, 6, 10, 12, 14, 16, 18, 20, 22.

$$\text{Bowley's coefficient of skewness} = \text{Size of } \left(\frac{11+1}{4}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 3^{\text{rd}} \text{ item} = 6$$

$$Q_1 = \text{Size of } 3\left(\frac{n+1}{4}\right)^{\text{th}} \text{ item}$$

$$Q_1 = \text{Size of } 3\left(\frac{11+1}{4}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 9^{\text{th}} \text{ item} = 18$$

$$\text{Median} = \text{Size of } \left(\frac{n+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } \left(\frac{11+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 6^{\text{th}} \text{ item} = 12$$



$$\text{Bowley's Coefficient of skewness (sk)} = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

$$= \frac{18 + 6 - 2 \times 12}{18 - 6} = 0$$

Since $sk = 0$, the given series is a symmetrical data.

Example

From the data given below, calculate coefficient of skewness.

x	125	130	135	140	145	150	155	160	165	170
f	10	12	8	5	15	13	7	14	6	10

Solution:

x	f	c.f
125	10	10
130	12	22
135	8	30
140	5	35
145	15	50
150	13	63
155	7	70
160	14	84
165	6	90
170	10	100
	100	

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



$$= \text{Size of } \left(\frac{100+1}{2}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } 50.5^{\text{th}} \text{ item} = 150$$

Median = 150

$$Q_1 = \text{Size of } \left(\frac{N+1}{4}\right)^{\text{th}} \text{ item}$$

$$= \text{Size of } \left(\frac{100+1}{4}\right)^{\text{th}} \text{ item}$$

$$= 25.25$$

$$Q_1 = 135$$

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

$$= \text{Size of } 3\left(\frac{100+1}{4}\right)^{\text{th}} \text{ item}$$

$$= 75.75$$

$$Q_3 = 160$$

$$\text{Bowley's Coefficient of skewness (sk)} = \frac{Q_3 + Q_1 - 2\text{Median}}{Q_3 - Q_1}$$

$$= \frac{16+135 - 2(150)}{160 - 135}$$

Bowley's Coefficient of skewness = 0.2

Example

Calculate the value of the Bowley's coefficient of skewness from the following series.

Wages (in Rs)	10-20	20-30	30-40	40-50	50-60	60-70	70-80
No. of persons	1	3	11	21	43	32	9



Solution:

Wages (Rs)	f	C.f
10-20	1	1
20-30	3	4
30-40	11	15
40-50	21	36
50-60	43	79
60-70	32	111
70-80	9	120
	N=120	

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times C_1$$

$$\frac{N}{4} = \frac{120}{4} = 30$$

Q_1 Class = 40-50

$$l_1 = 40, m_1 = 15, f_1 = 21, C_1 = 10$$

$$Q_1 = 40 + \frac{\frac{N}{4} - 15}{21} \times 10$$

$$= 40 + \frac{30-15}{21} \times 10$$

$$= 40 + \frac{150}{21}$$

$$= 40 + 7.14$$

$$= 47.14$$

$$Q_2 = l + \frac{\frac{N}{2} - m}{f} \times c$$



$$\frac{N}{2} = \frac{120}{4} = 60$$

Q_1 Class = 50-60

$$l = 50, m = 36, f = 43, C = 10$$

$$Q_2 = 50 + \frac{60 - 36}{43} \times 10$$

$$\begin{aligned} Q_2 &= 50 + \frac{240}{43} \\ &= 50 + 5.58 \\ &= 55.58 \end{aligned}$$

$$Q_3 = l_3 + \frac{\frac{3N}{4} - m_3}{f_3} \times C_3$$

$$3 \frac{N}{4} = 3 \times \frac{120}{4} = 90$$

Q_1 Class = 60-70

$$l_3 = 40, m_3 = 79, f_3 = 32, C_3 = 10$$

$$Q_3 = 60 + \frac{90 - 79}{32} \times 10$$

$$\begin{aligned} Q_3 &= 60 + \frac{110}{32} \\ &= 63.44 \end{aligned}$$

$$\text{Bowley's Coefficient of skewness (sk)} = \frac{63.44 + 47.14 - 2 \times 55.58}{63.44 - 47.14}$$

$$= \frac{110.58 - 111.16}{16.30}$$

$$= \frac{-0.58}{16.30}$$

Bowley's Coefficient of skewness = -0.0356



3.4 MEASURE OR COEFFICIENT OF SKEWNESS BASED ON MOMENTS

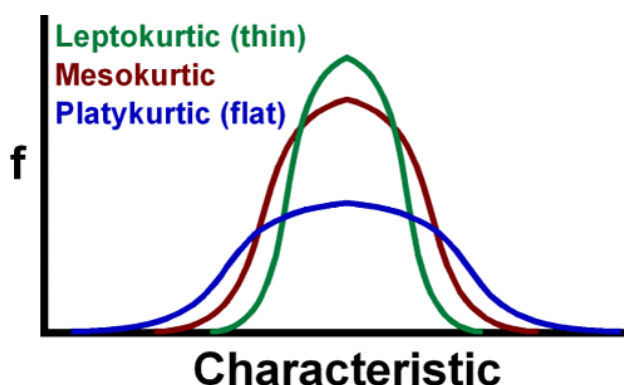
The measure of skewness based on moments is denoted by β_1 and is given by:

$$\beta_1 = \frac{\mu_3}{\mu_2^{3/2}}$$

If μ_3 is negative, then β_1 is negative.

3.5 KURTOSIS OR CO-EFFICIENT OF KURTOSIS

The coefficient of kurtosis, or simply kurtosis, measures the peakedness of a distribution. The criterion for determining the shape of a unimodal frequency curve is its peakedness. Kurtosis is a greek word and it means bulginess. The term kurtosis was introduced by Karl Pearson in 1906. If the frequency curve is highly peaked, a large number of observations have low frequency and are spread in the mid of interval. In both these situations, the curve is said to be a kurtic curve. If a frequency curve is more peaked than normal, it is called a platykurtic curve. If a curve is properly peaked, it is called a mesokurtic curve. The three types of curve are shown in figure.



Measure of Kurtosis

The measure of kurtosis of a frequency distribution based moments is denoted by β_2 and is given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

If $\beta_2 = 3$, the distribution is said to be normal and the curve is mesokurtic.



If $\beta_2 > 3$, the distribution is said to be more peaked and the curve is leptokurtic.

If $\beta_2 < 3$, the distribution is said to be flat topped and the curve is platykurtic.

Example

Calculate β_1 and β_2 for the following data.

x	0	1	2	3	4	5	6	7	8
f	5	10	15	20	25	30	35	40	45

Solution:

$$\mu_1 = 0$$

$$\mu_1 = \frac{\sum f d^2}{N} = \frac{500}{125} = 4$$

$$\mu_3 = \frac{\sum f d^3}{N} = 0$$

$$\mu_4 = \frac{\sum f d^4}{N} = \frac{4700}{125} = 37.6$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{0}{64} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{37.6}{4^2} = 2.35$$

The value of β_2 is less than 3, hence the curve is platykurtic.

Example

The first four central moments of a distribution are 0, 2.5, 0.7, 18.75, Examine the Kurtosis of the distribution.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_1 = \frac{0.7^2}{2.5^3} = 0.031$$

$$\beta_1 = 0.031$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{2.5^2} = 3$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{2.5^2}$$



$$\beta_2 = 3$$

Hence the curve is mesokurtic curve

QUESTIONS

1. Define moments?
2. Define Karl-Pearson's Co-efficient skewness
3. Calculate the Karl-Pearson's Co-efficient of skewness.

Daily wages (in Rs.)	150	200	250	300	350	400	450
No. of People	3	25	19	16	4	5	6

4. Define Bowley's Co-efficient skewness.
5. Find the Bowley's Co-efficient of skewness for the following series 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22.
5. Define kurtosis and measure of kurtosis.
6. Calculate β_1 and β_2 for the following data.

x	0	1	2	3	4	5	6	7	8
f	5	10	15	20	25	30	35	40	45

7. Calculate first four moments from the following data?

x	0	1	2	3	4	5	6	7	8
f	5	10	15	20	25	20	15	10	5

8. Calculate the first four moments about the mean

x	30- 33	33- 36	36- 39	39- 42	42- 45	45- 48
f	2	4	26	47	15	6



9. Calculate Bowley's co-efficient skewness from the following data.

x	5-7	8-10	11- 13	14- 16	17- 19
f	14	24	38	20	4

10. Find the co-efficient skewness from the data given below.

x	3	4	5	6	7	8	9	10
f	7	10	14	35	102	136	43	8

11. Find the Bowley's co-efficient skewness of the following series.

x	4	4.5	5	5.5	6	6.5	7	7.5	8
f	10	18	22	25	40	15	10	8	7

12. Calculate β_1 and β_2 for the following data.

Wages (in Rs.)	170- 180	180- 190	190- 200	200- 210	210- 220	220- 230	230- 240	240- 250
No. of persons	52	68	85	92	100	95	70	28



UNIT – IV

4.1 CURVE FITTING

So far, the discussion about trend is confined mostly to the linear trend. But the linear trend does not always exist. The linear trend means a constant rate of growth or decay, which in many situations is refuted. Thus, one has to switch over to some other kind of trend which would naturally be a curvilinear trend. Such a situation generally arises when the x-values for years (periods) are in arithmetic progression and corresponding y-values are either in geometric progression, or follow some other law except arithmetic progression.

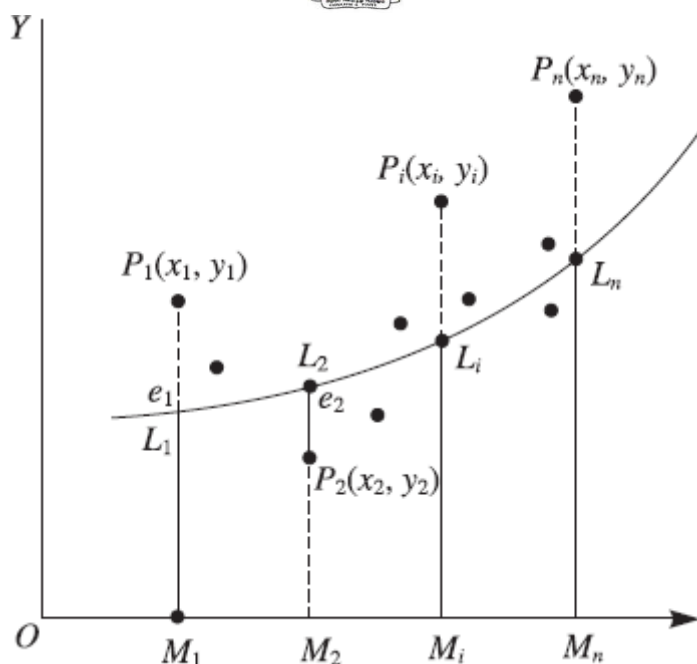
The process of finding the equation of the curve of best fit, which may be most suitable for predicting the unknown values, is known as curve fitting. Therefore, curve fitting means an exact relationship between two variables by algebraic equations. There are following methods for fitting a curve.

- Graphic method
- Method of group averages
- Method of moments
- Principle of least square.

Out of above four methods, we will only discuss and study here principle of least square.

4.2 PRINCIPLE OF LEAST SQUARES:

The graphical method has the drawback in that the straight line drawn may not be unique but principle of least squares provides a unique set of values to the constants and hence suggests a curve of best fit to the given data. The method of least square is probably the most systematic procedure to fit a unique curve through the given data points.



Let the curve $Y = a + bX + cX^2 + \dots + kX^{m-1}$

be fitted to the set of n data points (x_1, y_1) (x_2, y_2) (x_3, y_3) (x_n, y_n) . At $(x = x_i)$ the observed value of the ordinate is $y_i = P_i M_i$ and the corresponding value of the fitting curve is $a + bX + cX^2 + \dots + kX^m = L_i M_i$ which is expected or calculated value. The difference of the observed and expected value is $P_i M_i - L_i M_i = e_i$ this difference is called error at $(x = x_i)$ clearly some of the error $e_1, e_2, e_3, \dots, e_n$ will be positive and other negative. To make all errors positive we square each of the errors $S = e_1^2 + e_2^2 + e_3^2 + \dots + e_n^2$ the curve best fit is that for which e 's are as small as possible S , the sum of the square of the errors is a minimum this is known as the principle of least square.

4.3 LEAST SQUARE METHOD OF FITTING A REGRESSION LINE

The equation for a regression line of Y on X for the population is given as

$$Y = a + bX \quad \dots\dots\dots 1$$

equation 1 is also known as the mathematical model for linear regression. The main difference between the Cartesian equation of a line and a regression line is that a regression line is a probabilistic model which



enables one to develop procedures for making inferences about the parameters a and b model. In this model, the expected value of Y is a linear function of X , But for fixed X , the variable Y differs from its expected value by a random amount. As a special case, the form $Y = a + bX$ is called the deterministic model; the actual observed value of Y is a linear function of c . In this equation 'a' is the intercept which the line cuts on the axis of Y and b is the slope of the line. 'b' is also called the regression coefficient and is defined as, 'b' is the measure of change in the dependent variable (Y) corresponding to a unit change in the independent variable (X). 'b' is often written as b_{yx} to indicate that it is the regression coefficient of Y on X . In case no suffix is attached to 'b', it is considered by itself. B can take any real value within the range $-\infty$ to ∞ .

Suppose the regression line given by equation 1 is to be fitted on the basis of n pairs of sample observations, $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Each pair (x_i, y_i) for $i=1, 2, \dots, n$ will satisfy the regression line equation (1).

$$Y_i = a + b_{x_i} + e_i \dots \dots \dots 1.1$$

or

$$e_i = (Y_i - a - b_{x_i}) \dots \dots \dots 1.2$$

e_i may be positive or negative in case y_i is greater than or than $(a + b_x)$ respectively. Whether the error is positive negative, it does not matter as an error is after all an error and confining to its magnitude only, square both sides of equation 1.2 and take the sum over n pairs of observations. This gives,

$$\sum_i e_i^2 = \sum_i (Y_i - a - b_{x_i})^2 \dots \dots \dots 1.3$$

In Legendre's principle of least squares, the quantity which is minimized is the residuals or error sum of squares. Here the assumption is that each e_i is normally distributed with mean zero and variance $\sigma^2 e$. Thus, the quantity which is to be minimized here is $\sum_i e_i^2$. Let us denote this quantity by Q . Hence,



$$Q = \sum_i (Y_i - a - b_{x_i})^2 \dots\dots\dots 1.4$$

To get the least square estimates of 'a' and 'b', so that Q is minimum, differentiate Q partially with respect to 'a' and 'b', respectively, and equate to zero. Also replace 'a' and 'b' by their estimated values, say 'a' and 'b' respectively. Thus, we get two equations as give below. These equations are called normal equations.

$$\frac{\partial Q}{\partial a} = -2 \sum_i (Y_i - a - b_{x_i}) = 0 \dots\dots\dots 2$$

$$\frac{\partial Q}{\partial b} = -2 \sum_i (Y_i - a - b_{x_i}) x_i = 0 \dots\dots\dots 3$$

On rearranging we get,

$$na + b \sum_i x_i = \sum_i y_i \dots\dots\dots 4$$

$$a \sum_i x_i + b \sum_i x_i^2 = \sum_i y_i x_i \dots\dots\dots 5$$

From (1.4)

$$a + b \frac{1}{n} \sum_i x_i = \frac{1}{n} \sum_i y_i \dots\dots\dots 6$$

$$a + b\bar{x} = \bar{y}$$

$$a = \bar{y} - b\bar{x} \dots\dots\dots 7$$

(or)

Substituting the value of a from equation 1 in equation 5, we get

$$b = \frac{\sum_i y_i x_i - \frac{1}{n} (\sum_i x_i) (\sum_i y_i)}{\sum_i x_i^2 - \frac{1}{n} (\sum_i x_i)^2} \dots\dots\dots 8$$

Expression equation 8 can be easily written as,

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \dots\dots\dots 8.1$$

Suppose $x_i - \bar{x} = u_i$ and $y_i - \bar{y} = v_i$ under the transformation

$$b = \frac{\sum_i u_i v_i}{\sum_i u_i^2} \dots\dots\dots 8.2$$

If we divide the numerator and denominator by n in equation 1 we get



$$b = \frac{\text{cov}(X,Y)}{\text{var}(X)} \dots\dots\dots 8.3$$

$$b = \frac{S_{XY}}{S_X^2} \dots\dots\dots 8.4$$

B is the estimated regression coefficient of Y on X and is also symbolised as b_{YX} .

$$S_{XY} = rS_XS_Y$$

Substituting the value of S_{XY} in terms of rS_X and S_Y in equation 8.4 we obtain

$$b_{YX} = r \frac{S_Y}{S_X} \dots\dots\dots 8.5$$

Where r is the sample correlation coefficient between X and Y

Properties of regression coefficient (1) it can take any value between $-\infty$ to $+\infty$. Its sign is same as that of S_{XY} i.e. $\text{cov}(X, Y)$.

$$b = \frac{\sigma_{XY}}{\sigma_X^2} \dots\dots\dots 9$$

In case of population regression coefficient b of Y on X, which can elaborately be specified as b_{YX} we can express it as,

$$b_{YX} = \rho \frac{\sigma_Y}{\sigma_X}$$

Where ρ is the population correlation coefficient between X and Y.

As a and b are the estimated values of a and b respectively, the equation of the estimated regression line is

$$\hat{Y} = a + bX \dots\dots\dots 10$$

Substituting the value of a from equation 7, the line of best fit is

$$\hat{Y} = (\bar{y} - b\bar{x}) + bx$$

$$(\hat{Y} - \bar{y}) = b(X - \bar{x})$$

Prediction equation:

The regression line is also known as prediction equation. Once the constants a and b are calculated, there remains two unknown variables in



the regression equation viz., Y and X. Moreover, we know y depends on X in the case of regression equation of Y on X. Under the presumption that the trend of change in Y corresponding to X remains the same, the value of Y can be estimated for any value of X. But such a presumption rarely holds good for a very wide range of X values. Hence, the fitted regression line gives a better estimate of Y for a given value of X, which is within the range of X values, taken into consideration at the time of calculation a and b, or not much beyond the values of X involved in the calculations. For example, we know that the crop yield increases with the increase in the quantity of fertilizers applied in the field. But beyond certain fertilizer-dose, the increase in yield is negligible. Hence the estimation of the yield of a crop for a fertilizer does should be restricted only for doses within certain limits

4.3 FITTING OF THE CURVES OF THE FORM $Y=a+bx$

Regression line of X and Y often we come across situations in which two variables (Y and X) are such that not only Y depends on X but X also depends on Y. For example, the heights and weights of people are two variables where heights of people depend on weights and weights depend on heights. In such a case we can find not only the regression line of Y on X but also of X on Y. Suppose the regression line X on Y is,

$$Y=a+bx \dots\dots\dots 1$$

The parameters a and b can be estimated in the same way as a and b in equation 1. Instead of repeating the derivation, it will be worthwhile to write directly the estimated values of a and b, say a and b

$$a = (\bar{x} - b\bar{y}) \dots\dots\dots 2$$

$$b = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (y_i - \bar{y})^2} \dots\dots\dots 3$$

$$= \frac{\sum_i x_i y_i - (\sum_i x_i)(\sum_i y_i)/n}{\sum_i y_i^2 - \sum_i (y_i)^2/n} \dots\dots\dots 4$$

$$b = \frac{S_{XY}}{S_Y^2}$$



We can express b, which is the estimated regression coefficient of X and Y and symbolically denoted as

$$b_{YX} = r \frac{s_Y}{s_X} \dots \dots \dots 5$$

Where r is the sample correlation coefficient between x and y.

The equation of the estimated regression line of X on Y is.

$$(\hat{Y} - \bar{y}) = b(Y - \bar{y})$$

Also, the population regression coefficient of X on Y may be given as follows

$$b_{XY} = \frac{\sigma_{XY}}{\sigma_Y^2}$$

The population regression coefficient b of X on Y, which is often symbolised as b_{XY} , can be expressed as

$$b_{XY} = \rho \frac{\sigma_X}{\sigma_Y}$$

Where ρ is the population correlation coefficient between X and Y.

Example

Find the best fit values of a and b so that $y = a + bx$ fits the data given in the table.

x	0	1	2	3	4
y	1	1.8	3.3	4.5	6.3

Solution:

Let the straight line $y = a + bx$

x	y	xy	X ²
0	1	0	0
1	1.8	1.8	1
2	3.3	6.6	4
3	4.5	13.5	9



4	6.3	25.2	16
10	16.9	47.1	30

Normal equations are,

$$\sum y = na + b \sum x$$
$$\sum xy = a \sum x + na + \sum x^2$$

$$n = 5, \sum x = 10, \sum y = 16.9, \sum xy = 47.1, \sum x^2 = 30$$

Putting the values in Normal equations.

$$16.9 = 5a + 10b$$

$$47.1 = 10a + 30b$$

On solving these two equations we get

$$a = 0.72 \quad b = 1.33$$

so required line $y = 0.72 + 1.33x$

4.4 FITTING OF THE CURVES OF THE FORM $Y=a+bX+cX^2$

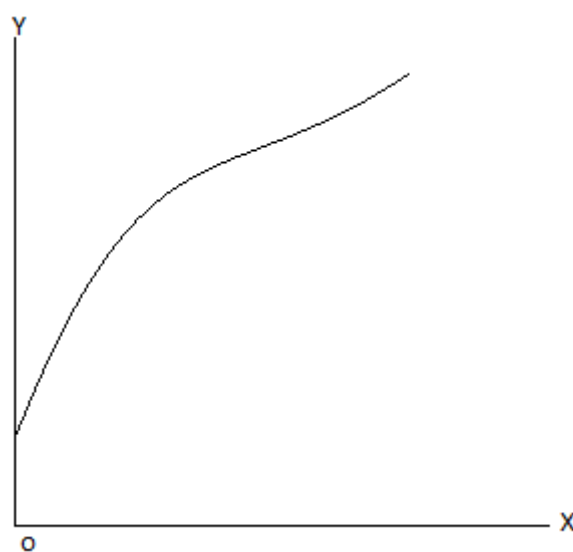


Figure.1 parabola



The shape of the curve is the upper half of the parabola. It is generally used for a relationship between the production of a crop and the quantity of fertilizer applied per unit area.

Let a Parabola $y = a + bx + cx^2$ 1

Which is fitted to a given data $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$

Let y_λ be the theoretical value for x_1 the $e_1 = y_1 - y_\lambda$

$$e_1 = y_1 - (a + bx_1 + cx_1^2)$$

$$e_1^2 = (y_1 - a - bx_1 - cx_1^2)^2$$

Now we have

$$S = \sum_{i=1}^n e_i^2$$

$$S = \sum_{i=1}^n (y_i - a - bx_i - cx_i^2)^2$$

By the principle least squares, the value of S is minimum therefore

$$\frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b} = 0, \text{ and } \frac{\partial S}{\partial c} = 0, \quad \dots \dots 2$$

Solving equation (2) and dropping suffix, we have

$$\sum y = na + b \sum x + c \sum x^2 \quad \dots \dots 3$$

$$\sum xy = a \sum x + b \sum x^2 + c \sum x^3 \quad \dots \dots 4$$

$$\sum x^2 y = a \sum x^2 + b \sum x^3 + c \sum x^4 \quad \dots \dots 5$$

The equation (3), (4) and (5) are known as normal equations.

On solving equations (3), (4) and (5), we get the values a, b and c. Putting the values of a, b and c in equation (1), we get the equation of the parabola of best fit.

4.5 EXPONENTIAL GROWTH CURVE

Mathematical equation of the curve is,

$$Y = a\beta^x$$

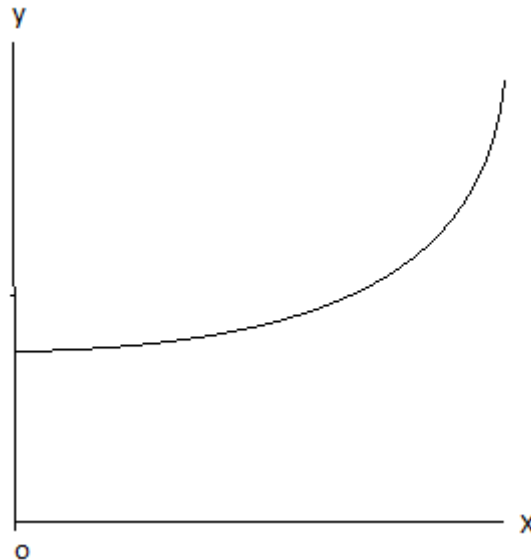


Figure.2. Growth Curve

If we put $\beta = 1 + i$, where i is the rate of interest and X the number of years, then, y gives the amount to which the initial amount a will rise. By taking the logarithm of both sides, the models become linear in log terms. Mathematical equation of exponential growth curve is also given as

$$Y = ae^{ix}$$

QUESTIONS

1. Explain the prediction equation.
2. Explain the exponential growth curve.
3. Explain the principle least squares.
4. Explain least squares method of fitting a regression line.
5. Explain the fitting of the curves of the form $Y = a + bX$.
6. Explain the fitting of the curves of the form $Y = a + bX + cX^2$



UNIT- V

5.1 LINEAR CORRELATION

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work. The study related to the characteristics of only variable such as height, weight, ages, marks, wages, etc., is known as univariate analysis. The statistical Analysis related to the study of the relationship between two variables is known as Bi-Variate Analysis. Sometimes the variables may be inter-related. In health sciences we study the relationship between blood pressure and age, consumption level of some nutrient and weight gain, total income and medical expenditure, etc. The nature and strength of relationship may be examined by correlation and Regression analysis. Thus Correlation refers to the relationship of two variables or more. Correlation is statistical Analysis which measures and analyses the degree or extent to which the two variables fluctuate with reference to each other. The word relationship is important. It indicates that there is some connection between the variables. It measures the closeness of the relationship. Correlation does not indicate cause and effect relationship. Price and supply, income and expenditure are correlated.

Meaning of Correlation:

In a bivariate distribution we may interested to find out if there is any correlation or covariation between the two variables under study. If the change in one variable affects a change in the other variable, the variables are said to be correlated. If the two variables deviate in the same direction, i.e., if the increase in one results in a corresponding increase in the other, correlation is said to be direct or positive.

Example

- The heights or weights of a group of persons
- The income and expenditure is positive and correlation between



- ❖ Price and demand of a commodity
- ❖ The volume and pressure of a perfect gas; is negative

Correlation is said to be perfect if the deviation one variable is followed by a corresponding and proportional deviation in the other.

Definitions:

Ya-Kun-Chou:

Correlation Analysis attempts to determine the degree of relationship between variables.

A.M. Tuttle:

Correlation is an analysis of the covariation between two or more variables. Correlation expresses the inter-dependence of two sets of variables upon each other. One variable may be called as (subject) independent and the other relative variable (dependent). Relative variable is measured in terms of subject.

Uses of correlation:

1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity.
4. Businessmen estimates costs, sales, price etc. using correlation.
4. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
5. Sampling error can be calculated.
6. It is the basis for the concept of regression.

5.2 SCATTER DIAGRAM

Scatter diagram pertaining independent variables, it is easily verifiable that if any line is drawn through the plotted points, not more than two points will be lying on the line most of the other points will be at a considerable distance from this line. Scatter diagram that the two variables

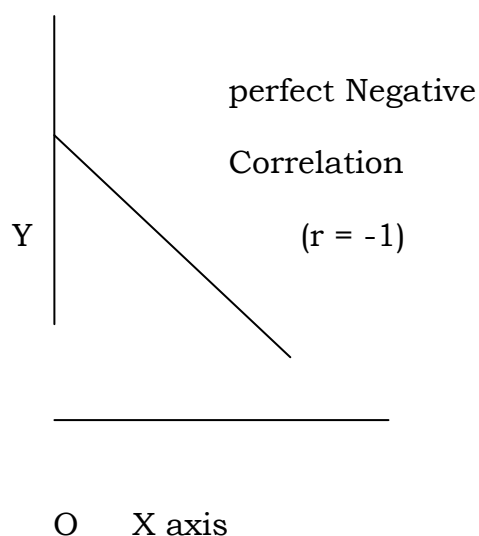
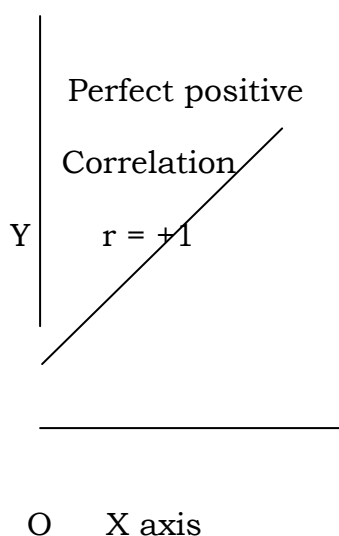


are linearly related, the problem arises on deciding which of the many possible lines the best fitted line is. The least square method is the most widely accepted method of fitting a straight line and is discussed here adequately.

Use of Scatter Diagram:

- When you have paired numerical data
- When trying to identify potential root causes of problems.
- After brain storming causes and effects using a fishbone diagram, to determine objectively whether a particular cause and effect are related.
- When determining whether two effects that appear to be related both occur with the same cause

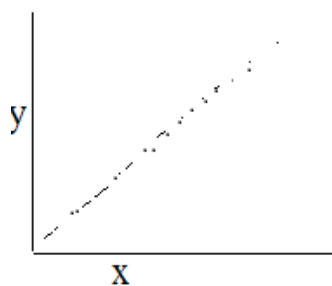
It is the simplest method of studying the relationship between two variables diagrammatically. One variable is represented along the horizontal axis and the second variable along the vertical axis. For each pair of observations of two variables, we put a dot in the plane. There are as many dots in the plane as the number of paired observations of two variables. The direction of dots shows the scatter or concentration of various points. This will show the type of correlation.



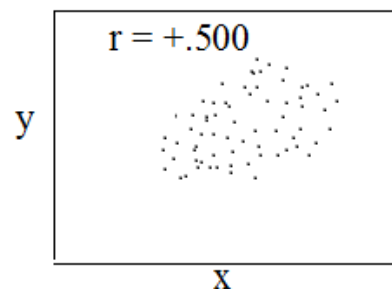


1. If all the plotted dots lie on a straight line falling from upper left hand corner to lower right hand corner, there is a perfect negative correlation between the two variables. In this case the coefficient of correlation takes the value $r = -1$.
2. If all the plotted points form a straight line from lower left hand corner to the upper right hand corner then there is Perfect positive correlation. We denote this as $r = +1$
3. If the plotted points in the plane form a band and they show a rising trend from the lower left hand corner to the upper right hand corner the two variables are highly positively correlated. Highly Positive
Highly Negative

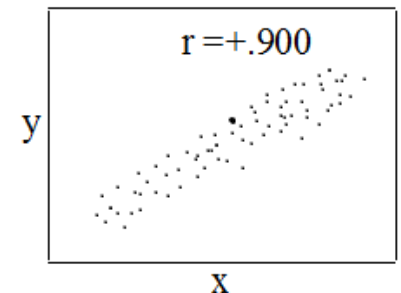
Perfectly + ve



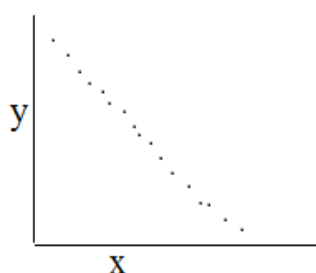
less degree + ve



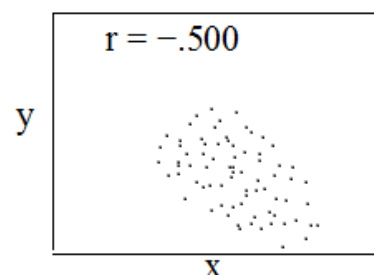
high degree + ve



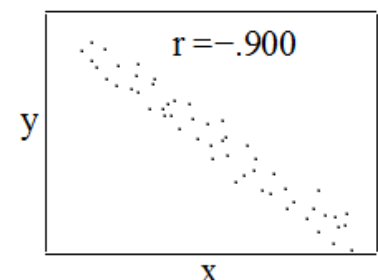
Perfectly - ve



less degree - ve



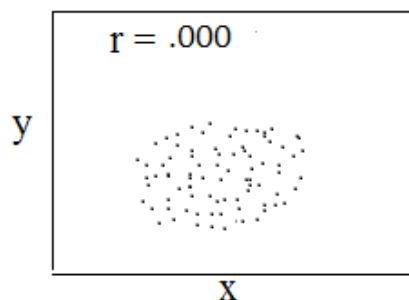
high degree - ve



1. If the points fall in a narrow band from the upper left hand corner to the lower right hand corner, there will be a high degree of negative correlation.



2. If the plotted points in the plane are spread all over the diagram there is no correlation between the two variables.



Merits:

1. It is a simplest and attractive method of finding the nature of correlation between the two variables.
2. It is a non-mathematical method of studying correlation. It is easy to understand.
3. It is not affected by extreme items.
4. It is the first step in finding out the relation between the two variables.
5. We can have a rough idea at a glance whether it is a positive correlation or negative correlation.

Demerits:

By this method we cannot get the exact degree or correlation between the two variables.

Types of Correlation:

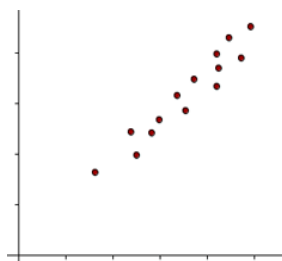
Correlation is classified into various types. The most important ones are

- Positive and negative.
- Linear and non-linear.
- Partial and total.
- Simple and Multiple.

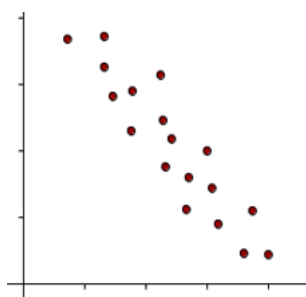


Positive and Negative Correlation

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (i. e) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.



If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.



Linear and Non-linear correlation:

If the ratio of change between the two variables is a constant then there will be linear correlation between them.



Example

Consider the variables with the following values.

X	10	20	30	40	50
Y	20	40	60	80	100

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curve-linear (or) non-linear correlation. The graph will be a curve.

Example

Consider the variables with the following values

X	10	20	30	40	50
Y	10	30	70	90	120

Here there is a non linear relationship between the variables. The ratio between them is not fixed for all points. Also if we plot them on the graph, the points will not be in a straight line. It will be a curve.

Simple and Multiple correlation:

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlations.

Partial and total correlation:

The study of two variables excluding some other variable is called Partial correlation. While mathematically controlling the influence of a third variable by holding is constant.



$$r_{xyz} = \frac{r_{xy} - (r_{xy} - r_{yz})}{\sqrt{(1 - r_{xz}^2 - 1 - r_{yz}^2)}}$$

Example

We study price and demand eliminating supply side. Relationship between one dependent and one independent variable. This is called Total Correlation.

Coefficient of simple determination: $r^2_{yx_1}$ (or) $r^2_{yx_2}$ (or) $r^2_{x_1x_2}$

Computation of correlation

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by 'r'.

Co-variation:

Covariance and correlation are both describe the degree of similarity between two random variables. Suppose that X and Y are real-valued random variables for the experiment with means $E(X)$, $E(Y)$ and variances $\text{Var}(X)$, $\text{Var}(Y)$, respectively. The co-variation between the variables x and y is defined as

$$\text{Cov}(X, Y) = \frac{\sum (x - \bar{x})(y - \bar{y})}{n}$$

Where \bar{x} , \bar{y} are respectively means of x and y and 'n' is the number of pairs of observations.

5.3 KARL PEARSON'S COEFFICIENT OF CORRELATION

Karl Pearson's a British Biometrician (1867 – 1936), degree of linear relationship between the two variables. It is most widely used method in practice and it is known as Pearson's coefficient of correlation. It is denoted by 'r'.

Correlation coefficient between two random variables X and Y usually denoted by $r(X, Y)$ or simply r_{xy} is a numerical measure of linear relationship between them and is defined as.



$$r = \frac{\text{Cov}(x, y)}{\sigma_x \cdot \sigma_y} \text{ where } \sigma_x, \sigma_y \text{ are S.D of } x \text{ and } y \text{ respectively}$$

$$r = \frac{\sum xy}{n \sigma_x \sigma_y}$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \cdot \sum Y^2}}, X = x - \bar{x}, Y = y - \bar{y}$$

When the deviations are taken from the actual mean we can apply any one of these methods. Simple formula is the third one. The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series.

Steps:

1. Find the mean of the two series x and y .
2. Take deviations of the two series from x and y . $X = x - \bar{x}$, $Y = y - \bar{y}$
3. Square the deviations and get the total, of the respective squares of deviations of x and y and

denote by $\sum X^2$, $\sum Y^2$ respectively.

4. Multiply the deviations of x and y and get the total and Divide by n . This is covariance.
5. Substitute the values in the formula.

$$r = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y} = \frac{\sum (x - \bar{x})(y - \bar{y}) / n}{\sqrt{\frac{\sum (x - \bar{x})^2}{n}} \sqrt{\frac{\sum (y - \bar{y})^2}{n}}}$$

The above formula is simplified as follows

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \cdot \sum Y^2}}, X = x - \bar{x}, Y = y - \bar{y}$$



Example

Find Karl Pearson's coefficient of correlation from the following data between height of father (x) and son (y).

x	64	65	66	67	68	69	70
y	66	67	65	68	70	68	72

Solution:

x	y	X = X-x x = 67	X ²	Y = Y-y y = 68	Y ²	XY
64	66	-3	9	-2	4	6
65	67	-2	4	-1	1	2
66	65	-1	1	-3	9	3
67	68	0	0	0	0	0
68	70	1	1	2	4	2
69	68	2	4	0	0	0
70	72	3	9	4	16	12
469	476	0	28	0	34	25

$$\bar{x} = \frac{469}{7} = 67; \bar{y} = \frac{476}{7} = 68$$

$$r = \frac{\sum XY}{\sqrt{\sum X^2 \cdot \sum Y^2}} = \frac{25}{\sqrt{28 \times 34}} = \frac{25}{\sqrt{952}} = \frac{25}{30.85} = 0.81$$

$$r = 0.91$$



Example:

Calculate coefficient of correlation from the following data.

x	1	2	3	4	5	6	7	8	9
y	9	8	10	12	11	13	14	16	15

Solution:

x	y	x ²	y ²	xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
45	108	285	1356	597

$$r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$
$$r = \frac{9 \times 597 - 45 \times 108}{\sqrt{(9 \times 285 - (45)^2) \cdot (9 \times 1356 - (108)^2)}}$$
$$r = \frac{5373 - 4860}{\sqrt{(2565 - 2025) \cdot (12204 - 11664)}}$$
$$r = \frac{513}{\sqrt{540 \times 540}} = \frac{513}{540} = 0.95$$

$$r = 0.95$$



Example

Calculate Pearson's Coefficient of correlation.

x	45	55	56	58	60	65	68	70	75	80	85
y	56	50	48	60	62	64	65	70	74	82	90

Solution:

x	y	u=x-A	v=y-B	u ²	v ²	uv
45	56	-20	-14	400	196	280
55	50	-10	-20	100	400	200
56	48	-9	-22	81	484	198
58	60	-7	-10	49	100	70
60	62	-5	-8	25	64	40
65	64	0	-6	0	36	0
68	65	3	-5	9	25	-15
70	70	5	0	25	0	0
75	74	10	4	100	16	40
80	82	15	12	225	144	180
85	90	20	20	400	400	400
		2	-49	1414	1865	1393

$$r = \frac{n \sum uv - (\sum u)(\sum v)}{\sqrt{[n \sum u^2 - (\sum u)^2][n \sum v^2 - (\sum v)^2]}}$$

$$r = \frac{11 \times 1393 - 2 \times (-49)}{\sqrt{(1414 \times 11 - (2)^2) \times (1865 \times 11 - (-49)^2)}}$$

$$r = \frac{15421}{\sqrt{15550 \times 18114}} = \frac{15421}{16783.11} = 0.92$$

$$r = 0.92$$



5.4 COMPUTATION OF CO-EFFICIENT OF CORRELATION OF GROUPED BI-VARIATE DATA:

When the number of observations is very large, the data is classified into two way frequency distribution or correlation table. The class intervals for 'y' are in the column headings and for 'x' in the stubs. The order can also be reversed. In such a case it is not advisable to deal with every worker's age and days lost individually. So the data are grouped and the correlation is calculated. Here the basic formula and interpretation remain the same. The frequencies for each cell of the table are obtained. The formula for calculation of correlation coefficient 'r' is

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \text{ where } \text{cov}(x, y) = \frac{\sum f(x - \bar{x})(y - \bar{y})}{N}$$

$$= \frac{\sum fxy}{N} - \bar{x}\bar{y}$$

$$\sigma_x^2 = \frac{\sum fx^2}{N} - \bar{x}^2; \sigma_y^2 = \frac{\sum fy^2}{N} - \bar{y}^2$$

N – total frequency

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

Steps:

1. Take the step deviations of the variable x and denote these deviations by u.
2. Take the step deviations of the variable y and denote these deviations by v.
3. Multiply **u**, **v** and the respective frequency of each cell and unit the figure obtained in the right hand bottom corner of each cell.
4. Add the corrected (all) as calculated in step 3 and obtain the total $\sum fuv$.
5. Multiply the frequencies of the variable x by the deviations of x and obtain the total $\sum fu$.
6. Take the squares of the step deviations of the variable x and multiply them by the respective frequencies and obtain the $\sum fu^2$.



Properties of Correlation:

Property 1: Correlation coefficient lies between -1 and $+1$

$$(i.e) -1 \leq r \leq +1$$

$$\text{Let } x' = \frac{x - \bar{x}}{\sigma_x} ; y' = \frac{y - \bar{y}}{\sigma_y}$$

Since $\sum (x'+y')^2$ being sum of squares is always non – negative

$$\sum (x'+y')^2 \geq 0$$

$$\sum x'^2 + \sum y'^2 + 2\sum x'y' \geq 0$$

$$\sum \left(\frac{x - \bar{x}}{\sigma_x} \right)^2 + \sum \left(\frac{y - \bar{y}}{\sigma_y} \right)^2 + 2\sum \left(\frac{x - \bar{x}}{\sigma_x} \right) \left(\frac{y - \bar{y}}{\sigma_y} \right) \geq 0$$

$$\frac{\sum (x - \bar{x})^2}{\sigma_x^2} + \frac{\sum (y - \bar{y})^2}{\sigma_y^2} + \frac{2\sum (x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} \geq 0$$

dividing by 'n' we get

$$\frac{1}{\sigma_x^2} \cdot \frac{1}{n} \sum (x - \bar{x})^2 + \frac{1}{\sigma_y^2} \cdot \frac{1}{n} \sum (y - \bar{y})^2 + \frac{2}{\sigma_x \sigma_y} \cdot \frac{1}{n} \sum (x - \bar{x})(y - \bar{y}) \geq 0$$

$$\frac{1}{\sigma_x^2} \sigma_x^2 + \frac{1}{\sigma_y^2} \sigma_y^2 + \frac{2}{\sigma_x \sigma_y} \cdot \text{cov}(x, y) \geq 0$$

$$1 + 1 + 2r \geq 0$$

$$2 + 2r \geq 0$$

$$2(1 + r) \geq 0$$

$$(1 + r) \geq 0$$

$$-1 \leq r \dots \dots \dots (1)$$

$$\text{Similarly, } \sum (x'-y')^2 \geq 0$$

$$2(1 - r) \geq 0$$

$$1 - r \geq 0$$

$$r \leq +1 \dots \dots \dots (2)$$

$$(1) + (2) \text{ gives } -1 \leq r \leq 1$$

Note: $r = +1$ perfect +ve correlation. $r = -1$ perfect –ve correlation between the variables.

Property 2: 'r' is independent of change of origin and scale.

Property 3: It is a pure number independent of units of measurement.



Property 4: Independent variables are uncorrelated but the converse is not true.

Property 5: Correlation coefficient is the geometric mean of two regression coefficients.

Property 6: The correlation coefficient of x and y is symmetric $r_{xy} = r_{yx}$.

Limitations:

- Correlation coefficient assumes linear relationship regardless of the assumption is correct or not.
- Extreme items of variables are being unduly operated on correlation coefficient.
- Existence of correlation does not necessarily indicate cause effect relation.

Interpretation:

The following rules helps in interpreting the value of 'r'.

- When $r = 1$, there is perfect +ve relationship between the variables.
- When $r = -1$, there is perfect -ve relationship between the variables.
- When $r = 0$, there is no relationship between the variables.
- If the correlation is +1 or -1, it signifies that there is a high degree of correlation, (+ve or -ve) between the two variables. If r is near to zero (i.e.) 0.1, -0.1, (or) 0.2 there is less correlation.

5.5 RANK CORRELATION

It is not always possible to take measurements on units or objects. Many characters are expressed in comparative terms such as beauty, smartness, temperament, honesty, colour, beauty, intelligence, character, morality etc. In such case the units are ranked pertaining to that particular character instead of taking measurements on them. Sometimes, the units are also ranked according to their quantitative measure. In these types of studies, two situations arise.



- The same set of units is ranked according to two characters.
- Two judges give ranks to the same set of units independently.

This method is based on ranks. The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by the psychologist; Charles Edward Spearman in 1906 developed a formula for correlation coefficient, which is known as Rank correlation or Spearman's correlation. It is defined as

$$r = 1 - \frac{6 \sum D^2}{n^3 - n}$$

r = rank correlation coefficient.

The formula is

$$r = 1 - \frac{6 \left[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \dots \right]}{n^3 - n}$$

Where 'm' is the number of items. The value of 'r' also lies between -1 and +1.

Example

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price?

Price of Tea	88	90	95	70	60	75	50
Price of Coffee	120	134	150	115	110	140	100

Solution:

Price of Tea	Rank	Price of Coffee	Rank	D	D ²
88	3	120	4	1	1
90	2	134	3	1	1
95	1	150	1	0	0



70	5	115	5	0	0
60	6	110	6	0	0
75	4	140	2	2	4
50	7	100	7	0	0
50	7	100	7	0	0

$$\begin{aligned}
 r &= 1 - \frac{6 \sum D^2}{n^3 - n} = 1 - \frac{6 \times 6}{7^3 - 7} \\
 &= 1 - \frac{36}{336} = 1 - 0.1071 \\
 &= 0.8929
 \end{aligned}$$

The relation between price of tea and coffee is positive at 0.89. Based on quality the association between price of tea and price of coffee is highly positive.

Example

In an evaluation of answer script the following marks are awarded by the examiners.

1st	88	95	70	960	50	80	75	85
2nd	84	90	88	55	48	85	82	72

Do you agree the evaluation by the two examiners is fair?

Solution:

x	R ₁	y	R ₂	D	D ²
88	2	84	4	2	4
95	1	90	1	0	0
70	6	88	2	4	16
60	7	55	7	0	0
50	8	48	8	0	0



80	4	85	3	1	1
85	3	75	6	3	9
					30

$$\begin{aligned}
 r &= 1 - \frac{6 \sum D^2}{n^3 - n} = 1 - \frac{6 \times 30}{8^3 - 8} \\
 &= 1 - \frac{180}{540} = 1 - 0.357 \\
 &= 0.643
 \end{aligned}$$

$r = 0.643$ shows fair in awarding marks in the sense that uniformity has arisen in evaluating the answer scripts between the two examiners.

Example

The ranks of same 16 students in Statistics and Mathematics are as follows. Two numbers within brackets denote the ranks of the students in mathematics and Statistics.

(1,1) (2,10) (3,3) (4,4) (5,5) (6,7) (7,2) (8,6) (9,8) (10,11) (11,15) (12,9) (13,14) (14,12) (15,16) (16,13).

Calculate the rank correlation coefficient for proficiencies of this group in Statistics and Mathematics.

Solution:

Ranks in Statistics	Ranks in Mathematics	$d = X - Y$	d^2
1	1	0	0
2	10	-8	64
3	3	0	0
4	4	0	0
5	5	0	0
6	7	-1	1
7	2	5	25



8	6	2	4
9	8	1	1
10	11	-1	1
11	15	-4	16
12	9	3	9
13	14	-1	1
14	12	2	4
15	16	-1	1
16	13	3	3
			136

Rank correlation coefficient is given by

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$$r = 1 - \frac{6 \times 136}{16 \times 255} = \frac{4}{5} \quad r = 0.8$$

Example

Rank Correlation for tied observations. Following are the marks obtained by 10 students in a class in two tests.

Students	A	B	C	D	E	F	G	H	I	J
Test 1	70	68	67	55	60	60	75	63	60	72
Test 2	65	65	80	60	68	58	75	63	60	70

Calculate the rank correlation coefficient between the marks of two tests.



Solution:

Student	Test 1	R ₁	Test 2	R ₂	D	D ²
A	70	3	65	5.5	-2.5	6.25
B	68	4	65	5.5	-1.5	2.25
C	67	5	80	1.0	4.0	16.00
D	55	10	60	8.5	1.5	2.25
E	60	8	68	4.0	4.0	16.00
F	60	8	58	10.0	-2.0	4.00
G	75	1	75	2.0	-1.0	1.00
H	63	6	62	7.0	-1.0	1.00
I	60	8	60	8.5	0.5	0.25
J	72	2	70	3.0	-1.0	1.00
						50.00

60 are repeated 3 times in test 1. 60, 65 are repeated twice in test 2.

$$m = 3; m = 2; m = 2$$

$$\begin{aligned}
 r &= 1 - \frac{6 \left[\sum D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) \right]}{n^3 - n} \\
 &= 1 - \frac{6 \left[50 + \frac{1}{12}(3^3 - 3) + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2) \right]}{10^3 - 10} \\
 &= 1 - \frac{6 \times 53}{990} = \frac{672}{990} = 0.68
 \end{aligned}$$

Interpretation: There is uniformity in the performance of students in the two tests.



5.6 CO-EFFICIENT OF CONCURRENT DEVIATIONS

A very simple and casual method of finding correlation when we are not serious about the magnitude of the two variables is the application of concurrent deviations. This method involves in attaching a positive sign for a x-value (except the first) if this value is more than the previous value and assigning a negative value if this value is less than the previous value. This is done for the y-series as well. The deviation in the x-value and the corresponding y-value is known to be concurrent if both the deviations have the same sign. Denoting the number of concurrent deviation by c and total number of deviations as m (which must be one less than the number of pairs of x and y values), the coefficient of concurrent deviation is given by

$$r_c = \pm \sqrt{\pm \frac{(2c - m)}{m}}$$

If $(2c-m) > 0$, then we take the positive sign both inside and outside the radical sign and if $(2c-m) < 0$, we are to consider the negative sign both inside and outside the radical sign. Like Pearson's correlation coefficient and Spearman's rank correlation coefficient, the coefficient of concurrent deviations also lies between -1 and 1 , both inclusive.

Example

Find the coefficient of concurrent deviations from the following data.

Year	1990	1991	1992	1993	1994	1995	1996	1997
Price	25	28	30	23	23	38	39	42
Demand	35	34	35	30	29	28	26	23

Solution:

year	price	Sign of deviation from the previous figure (a)	Demand	Sign of deviation from the previous figure (b)	Product deviation of (ab)
1990	25		35		



1991	28	+	34	-	-
1992	30	+	35	+	+
1993	23	-	30	-	+
1994	35	+	29	-	-
1995	38	+	28	-	-
1996	39	+	26	-	-
1997	42	+	23	-	-

In this case, m = number of pairs of deviations = 7

c = No. of positive signs in the product of deviation column = Number of concurrent deviations = 2.

$$= \pm \sqrt{\pm \frac{(2c - m)}{m}}$$

$$= \pm \sqrt{\pm \frac{(4 - 7)}{7}}$$

$$= \pm \sqrt{\pm \frac{(-3)}{7}}$$

$$= -\sqrt{\frac{3}{7}}$$

$$= -0.65$$

Thus there is a negative correlation between price and demand.

5.7 REGRESSION EQUATION

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known



as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear. Linear regression equation of Y on X is also known as the mathematical model for linear regression equation, the main difference between the Cartesian equation of a line is that a regression line is a probabilistic model which enables to develop procedures for making inferences about the parameters a and b of the model. In this model, the expected value of Y is a linear function of X, but for fixed X, the variable Y differs from its expected value by a random amount.

$$Y = a + bX \dots\dots (1)$$

And X on Y is

$$X = a + bY \dots\dots (2)$$

a, b are constants.

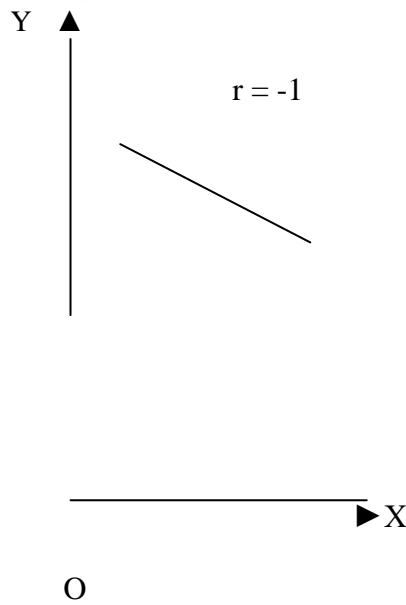
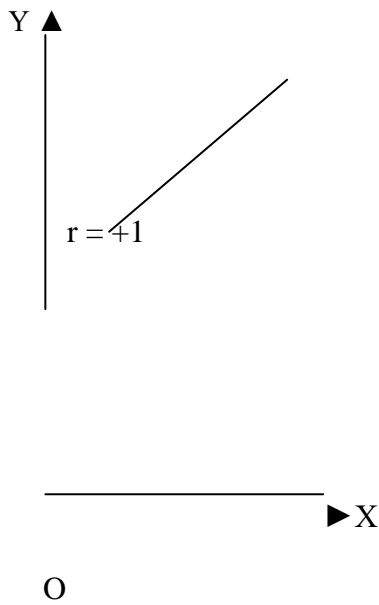
From (1) We can estimate Y for known value of X.

(2) We can estimate X for known value of Y.

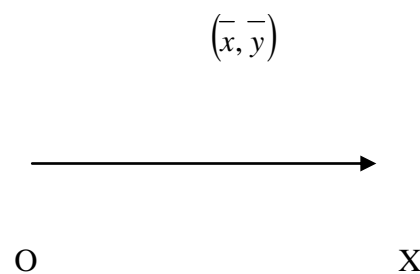
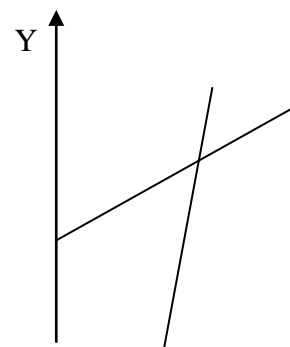
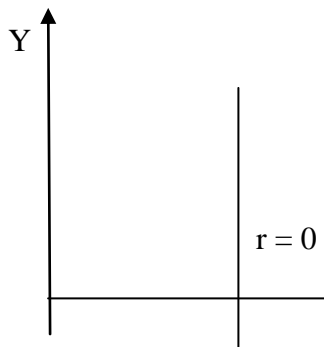
As a special case, the form $Y = a + bX$ is called the deterministic model. In this model, the actual observed value of Y is a linear function of X.

Regression Lines:

For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables. For perfect correlation, positive or negative i.e., $r = + 1$, the two lines coincide i.e., we will find only one straight line. If $r = 0$, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axis.



Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X-axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y-axis will touch the mean value of Y.



Principle of Least Squares:

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of “least



squares”. This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares. A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

- The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero.

$$\text{i.e., } \sum(X - X_c) = 0 \text{ or } \sum(Y - Y_c) = 0$$

Where X_c and Y_c are the values obtained by regression analysis.

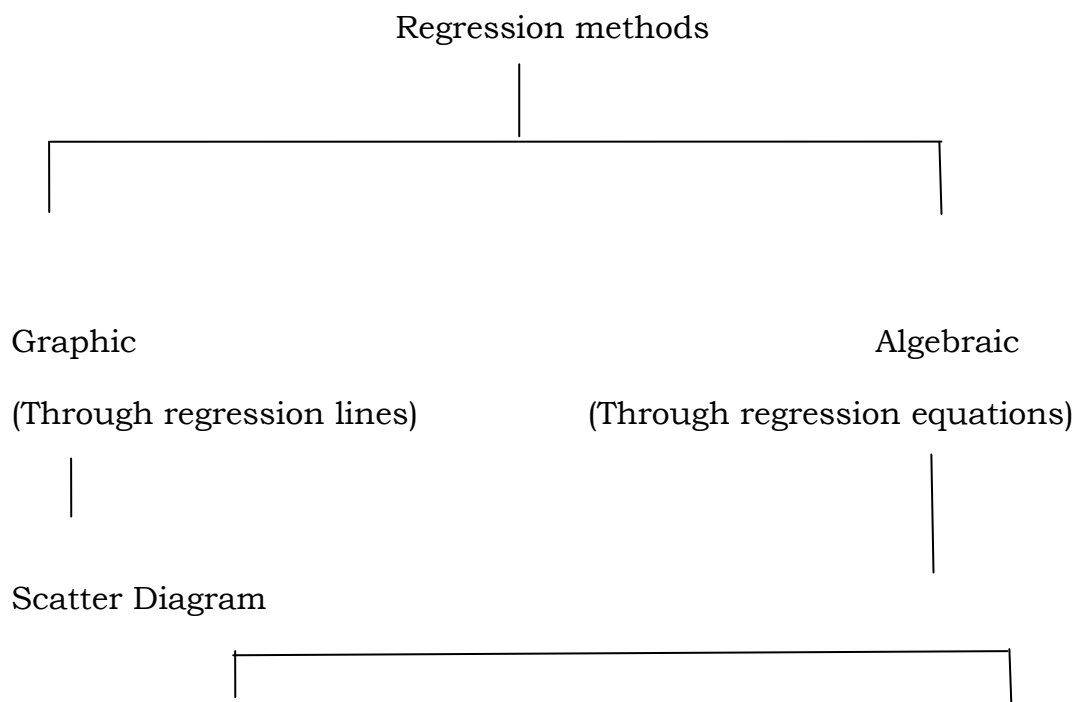
- The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e., $\sum(Y - Y_c)^2 < \sum(Y - A_i)^2$

Where A_i = corresponding values of any other straight line.

The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point are \bar{x}, \bar{y} .

Methods of Regression Analysis:

The various methods can be represented in the form of chart given below:





Regression Equations
(Through normal equations)

Regression Equations
(Through regression coefficient)

Graphic Method:

Scatter Diagram

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

Algebraic Method:

- **Regression Equation**

The two regression equations for X on Y; $X = a + bY$

And for Y on X; $Y = a + bX$

Where X, Y are variables, and a,b are constants whose values are to be determined

For the equation, $X = a + bY$

The normal equations are

$$\sum X = na + b \sum Y \text{ and}$$

$$\sum XY = a \sum Y + b \sum Y^2$$

For the equation, $Y = a + bX$, the normal equations are

$$\sum Y = na + b \sum X \text{ and}$$

$$\sum XY = a \sum X + b \sum X^2$$

From these normal equations the values of a and b can be determined.



5.8 REGRESSION CO-EFFICIENT

The regression equation of Y on X is $y_e = \bar{y} + r \frac{\sigma_y}{\sigma_x} (x - \bar{x})$

Here, the regression Coefficient of Y on X is

$$b_1 = b_{yx} = r \frac{\sigma_y}{\sigma_x}$$
$$y_e = \bar{y} + b_1(x - \bar{x})$$

The regression equation of X on Y is

$$X_e = \bar{x} + r \frac{\sigma_x}{\sigma_y} (y - \bar{y})$$

Here, the regression Co-efficient of X on Y

$$b_2 = b_{xy} = r \frac{\sigma_x}{\sigma_y}$$
$$X_e = \bar{X} + b_2(y - \bar{y})$$

If the deviation are taken from respective means of x and y

$$b_1 = b_{yx} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (X - \bar{X})^2} = \frac{\sum xy}{\sum x^2} \text{ and}$$
$$b_2 = b_{xy} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sum (Y - \bar{Y})^2} = \frac{\sum xy}{\sum y^2}$$

where $x = X - \bar{X}$, $y = Y - \bar{Y}$

If the deviations are taken from any arbitrary values of x and y

(Short – cut method)

$$b_1 = b_{yx} = \frac{n \sum uv - \sum u \sum v}{n \sum u^2 - (\sum u)^2}$$
$$b_2 = b_{xy} = \frac{n \sum uv - \sum u \sum v}{n \sum v^2 - (\sum v)^2}$$

Where u = X - A: v = Y - B

A = any value in X



B = any value in Y

Example

Find the two regression equations from the following data:

x	6	2	10	4	8
y	9	11	5	8	7

Solution:

X	Y	X ²	Y ²	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

Regression equation of Y on X is $Y = a + bX$ and the normal equations are

$$\begin{aligned}\sum Y &= na + b\sum X \\ \sum XY &= a\sum X + b\sum X^2\end{aligned}$$

Substituting the values, we get

$$40 = 5a + 30b \dots (1)$$

$$214 = 30a + 220b \dots (2)$$

Multiplying (1) by 6

$$240 = 30a + 180b \dots (3)$$

$$(2) - (3) \quad -26 = 40b$$

$$\text{Or } b = -\frac{26}{40} = -0.65$$

Now, substituting the value of 'b' in equation (1)



$$40 = 5a - 19.5$$

$$5a = 59.5$$

$$a = \frac{59.5}{5} = 11.9$$

Hence, required regression line Y on X is $Y = 11.9 - 0.65 X$.

Again, regression equation of X on Y is

$$X = a + bY \text{ and}$$

The normal equations are

$$\begin{aligned}\sum X &= na + b\sum Y \\ \sum XY &= a\sum Y + b\sum Y^2\end{aligned}$$

Now, substituting the corresponding values from the above table,

We get

$$30 = 5a + 40b \dots (3)$$

$$214 = 40a + 340b \dots (4)$$

Multiplying (3) by 8, we get

$$240 = 40a + 320b \dots (5)$$

(4) - (5) gives

$$-26 = 20b$$

$$b = -\frac{26}{20} = -1.3$$

Substituting $b = -1.3$ in equation (3) gives

$$30 = 5a - 52$$

$$5a = 82$$

$$a = \frac{82}{5} = 16.4$$

Hence, required regression line of X on Y is

$$X = 16.4 - 1.3Y$$



Properties of Regression Co-efficients:

- Both regression coefficients must have the same sign, ie either it will be positive or negative.
- The correlation coefficient is the geometric mean of two regression coefficients.

Correlation coefficient is the geometric mean of the regression coefficients ie, $r = \pm\sqrt{b_1b_2}$.

- The value of the coefficient of correlation cannot exceed unity i.e. **1**. Therefore, if one of the regression coefficients is greater than unity, the other must be less than unity.
- The correlation coefficient will have the same sign as that of the regression coefficients.
- If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
- The sign of both the regression coefficients will be same, i.e. they will be either positive or negative. Thus, it is not possible that one regression coefficient is negative while the other is positive.
- Regression coefficients are independent of origin but not of scale.
- The coefficient of correlation will have the same sign as that of the regression coefficients, such as if the regression coefficients have a positive sign, then “r” will be positive and vice-versa.
- Arithmetic mean of b_1 and b_2 is equal to or greater than the coefficient of correlation. Symbolically $\frac{b_1 + b_2}{2} \geq r$
- If $r=0$, the variables are uncorrelated, the lines of regression become perpendicular to each other.
- The regression coefficients are independent of the change of origin, but not of the scale. By origin, we mean that there will be no effect on the regression coefficients if any constant is subtracted from the value of X and Y. By scale, we mean that if the value of X and Y is either multiplied or divided by some constant, then the regression coefficients will also change.



- If $r = \pm 1$, the two lines of regression either coincide or parallel to each other
- Angle between the two regression lines is $\theta = \tan^{-1} \left[\frac{m_1 - m_2}{1 + m_1 m_2} \right]$
- Where m_1 and, m_2 are the slopes of the regression lines X on Y and Y on X respectively.
- The angle between the regression lines indicates the degree of dependence between the variables.

QUESTIONS

1. What is the correlation? Distinguish between positive and negative correlation.
2. Define Karl Pearson's coefficient of correlation. Interpret r , when $r = 1, -1$ and 0 .
3. Distinguish between linear and non-linear correlation.
4. Mention important properties of correlation coefficient.
5. What is Rank correlation? What are its merits and demerits?
6. $\text{Cov}(x, y) = 18.6$; $\text{Var}(x) = 20.2$; $\text{Var}(y) = 23.7$. Find 'r'.
7. Rank correlation coefficient $r = 0.8$. $\Sigma D^2 = 33$. Find 'n'.
8. Calculate the coefficient of correlation between X and y for the following

x	1	3	4	5	7	8	10
y	2	6	8	10	14	16	20

9. Find the correlation coefficient between the marks obtained by ten students in economics and statistics.

Marks in (Maths)	70	68	67	55	60	60	75	63	60	72
Marks in (Statistics)	65	65	80	60	68	58	75	62	60	70



10. Two judges gave the following ranks to eight competitors in a beauty contest. Examine the relationship between their judgements.

Judge A	4	5	1	2	3	6	7	8
Judge B	8	6	2	3	1	4	5	7

11. From the following data, calculate the coefficient of rank correlation.

x	36	56	20	65	42	33	44	50	15	60
y	50	35	70	25	58	75	60	45	80	38

12. Calculate spearman's coefficient of Rank correlation for the following data.

x	53	98	95	81	75	71	59	55
y	47	25	32	37	30	40	39	45

13. Obtain the rank correlation coefficient for the following data.

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

14. What is a scatter diagram? How is it useful in the study of Correlation?

15. Explain different types of correlation with examples.

16. Distinguish between Karl Pearson's coefficient of correlation and Spearman's correlation coefficient.

17. Compute the coefficient of correlation of the following score of A and B.

A	5	10	5	11	12	4	3	2	7	1
B	1	6	2	8	5	1	4	6	5	2

18. Calculate coefficient of Correlation between price and supply. Interpret the value of correlation coefficient.

Price	8	10	15	17	20	22	24	25
Supply	25	30	32	35	37	40	42	45



19. Find out Karl Pearson's coefficient of correlation in the following series relating to prices and supply of a commodity.

Price (Rs.)	11	12	13	14	15	16	17	18	19	20
Supply (Rs.)	30	29	29	25	24	24	24	21	18	15

20. Compute the coefficient of correlation from the following data.

Age of workers	40	34	22	28	36	32	24	46	26	30
Days absent	2.5	3	5	4	2.5	3	4.5	2.5	4	3.5

21. Calculate Karl Pearson's coefficient of correlation, for the following data.

Class Interval	0	1	2	3	4	5	6	7	8	Total
20-29	2	1	2	2	-	1	-	1	1	10
30-39	-	2	-	1	-	2	-	1	2	8
40-49	-	2	-	2	-	-	1	-1	1	6
50-59	1	-	2	-	-	-	-	1	-	4
60-69	-	-	-	-	-	1	-	1	-	2

22. Find the coefficient of correlation between the ages of 100 mothers and daughters

Age of daughters in years (Y)

Age of mothers in years (X)	5-10	10-15	15-20	20-25	25-30	Total
15-25	6	3	-	-	-	9
25-35	3	16	10	-	-	29
35-45	-	10	15	7	-	32
45-55	-	-	7	10	4	21
55-65	-	-	-	4	5	9
Total	9	29	32	21	9	100



23. The following table gives class frequency distribution of 45 clerks in a business office according to age and pay. Find correlation between age and pay if any.

Pay

Age	60-70	70-80	80-90	90-100	100-110	Total
20-30	4	3	1	-	-	8
30-40	2	5	2	1	-	10
40-50	1	2	3	2	1	9
50-60	-	1	3	5	2	11
60-70	-	-	1	1	5	7
Total	7	11	10	9	8	45

24. Find the correlation coefficient between two subjects marks scored by 60 candidates.

Marks in Statistics

Marks in economics	5-15	15-25	25-35	35-45	Total
0-10	1	1	-	-	2
10-20	3	6	5	1	15
20-30	1	8	9	2	20
30-40	-	3	9	3	15
40-50	-	-	4	4	8
Total	5	18	27	10	60

25. The following table gives the no. of students having different heights and weights. Do you find any relation between height and weight?



Weight in Kg

Height in cms	50-60	60-65	65-70	70-75	75-80	Total
150-155	1	3	5	7	2	18
155-160	2	4	10	7	4	27
160-165	1	5	12	10	7	35
165-170	-	3	8	6	3	20
Total	4	15	37	28	16	100

26. Apply spearman's Rank difference method and calculate coefficient of correlation between x and y from the data given below.

x	22	28	31	23	29	31	27	22	31	18
y	18	25	25	37	31	35	31	29	18	20

27. Find the rank correlation coefficients.

Marks in Test I	70	68	67	55	60	60	75	63	60	72
Marks in Test II	65	65	80	60	68	58	75	62	60	70

28. Calculate spearman's Rank correlation coefficient for the following table of marks of students in two subjects.

First subject	80	64	54	49	48	35	32	29	20	18	15	10
Second Subject	36	38	39	41	27	43	45	52	51	42	40	52

BOOKS FOR STUDY:

1. Anderson, T.W. and Sclove, S.L. (1978) Introduction to Statistical Analysis of data. Houghton Mifflin, Boston.

2. Bhat, B.R., Srivenkataramna, T. and Madhava Rao, K.S. (1996) Statistics a Beginner's Text, Vol. I, New Age International, New Delhi.



3. Croxton, F.E. and Cowden, D.J. (1969) Applied General Statistics, Prentice Hall, New Delhi.
4. Goon, A.M., M.K. Gupta and B. Das Gupta (2002) Fundamentals of Statistics- Vol. I, World Press Ltd, Kolkata.
5. Gupta, S.C. and V.K. Kapoor (2002) Fundamentals of Mathematical Statistics. Sultan Chand & Sons, New Delhi.
6. Spiegel, M.R. and Stephens, L. (2010) Statistics, Schaum's Outline Series. Mc Graw Hill, New York.

Course Material Prepared by

Dr. P. ARUMUGAM

Associate Professor, Department of Statistics,

Manonmaniam Sundaranar University, Tirunelveli – 627 012.