PERIYAR UNIVERSITY

(NAAC 'A++' Grade with CGPA 3.61 (Cycle - 3)
State University - NIRF Rank 59 - NIRF Innovation Band of 11-50)
SALEM - 636 011

CENTRE FOR DISTANCE AND ONLINE EDUCATION

(CDOE)

BACHELOR OF BUSINESS ADMINISTRATION

SEMESTER - I

CORE – I: BUSINESS STATISTICS

(Candidates admitted from 2024 onwards)

CENTRE FOR DISTANCE AND ONLINE EDUCATION (CDOE)

BACHELOR OF BUSINESS ADMINISTRATION

SEMESTER - I

CORE – I: BUSINESS STATISTICS

(Candidates admitted from 2024 onwards)

Prepared by

Dr. K. Alagirisamy
Assistant Professor
Department of Statistics
Periyar University
Salem – 636011.

List of Contents

BUSINESS STATISTICS

OBJECTIVE:

The objective of data collection is to gather accurate and relevant information that can be used to answer specific questions, test hypotheses, and evaluate outcomes. The process is fundamental to research, decision-making, and strategic planning across various fields, including business, science, healthcare, and social sciences.

1.1 STATISTICS – AN INTRODUCTION:

In the modern world of computers and information technology, the importance of statistics is very well recogonised by all the disciplines. Statistics has originated as a science of statehood and found applications slowly and steadily in Agriculture, Economics, Commerce, Biology, Medicine, Industry, planning, education and so on. As on date there is no other human walk of life, where statistics cannot be applied.

1.1.1 Origin and Growth of Statistics:

The word ' Statistics' and ' Statistical' are all derived from the Latin word Status, means a political state. The theory of statistics as a distinct branch of scientific method is of comparatively recent growth. Research particularly into the mathematical theory of statistics is rapidly proceeding and fresh discoveries are being made all over the world.

1.1.2 Meaning of 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 inparticular area.
2. The study of ways of collecting, analysing and interpretingthe facts.

1.1.3 Definitions:

Statistics is defined differently by different authors over a period of time. In the olden days statistics was confined to only state affairs but in modern days it embraces almost every sphere of human activity. Therefore a number of old definitions, which was confined to narrow field of enquiry were replaced by more definitions, which are much more comprehensive and exhaustive. Secondly, statistics has been defined in two different ways – Statistical data and statistical methods.

1.1.4 Descriptive Statistics

This involves summarizing and organizing data so it can be easily understood. Techniques include:

Measures of Central Tendency: Mean, median, and mode.

Measures of Dispersion: Range, variance, and standard deviation.

Data Visualization: Creating charts and graphs to represent data visually.

1.1.5 Inferential Statistics

This involves making predictions or inferences about a population based on a sample of data. Techniques include:

Hypothesis Testing: Assessing if there is enough evidence to support a certain belief about a population.

Regression Analysis: Examining the relationship between variables.

Confidence Intervals: Estimating the range within which a population parameter lies, with a certain level of confidence.

The following are some of the definitions of statistics as numerical data.

1.Statistics are the classified facts representing the conditions of people in a state. In particular they are the facts, which can be stated in numbers or in tables of numbers or in any tabular or classified arrangement.

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

1.1.6 Definitions by A.L. Bowley:

Statistics are numerical statement of facts in any department of enquiry placed in relation to each other. - A.L. Bowley Statistics may be called the science of counting in one of the departments due to Bowley, obviously this is an incomplete definition as it takes into account only the aspect of collection and ignores other aspects such as analysis, presentation and interpretation.

Bowley gives another definition for statistics, which states ‘ statistics may be rightly called the scheme of averages’ . This definition is also incomplete, as averages play an important role in understanding and comparing data and statistics provide more measures.

1.1.7 Definition by 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:

1.**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.

2.**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.

3.**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.,

4.**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.

1.1.8. Definition by 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. The above definition seems to be the most comprehensive and exhaustive.

1.2 COLLECTION OF DATA.

Everybody collects, interprets and uses information, much of it in a 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.

1.To describe the methods of collecting primary statistical information.
2.To consider the status involved in carrying out a survey.
3.To analyse the process involved in observation and interpreting.
4.To define and describe sampling.
5.To analyse the basis of sampling.
6.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 precautions 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:

1.Time series data.

2.Spatial data

3.Spacio-temporal data.

1.2.1 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 2001,2002,2003,2004**.**

Year	Food	Educa tion	Others	Total
2001	3000	2000	3000	8000
2002	3500	3000	4000	10500
2003	4000	3500	5000	12500
2004	5000	5000	6000	16000

1.2.2 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

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

2.District wise rainfall in Tamilnadu

3.Prices of silver in four metropolitan cities

Example:

The population of the southern states of India in 1991**.**

State	Population
Tamilnadu	5,56,38,318
Andhra Pradesh	6,63,04,854
Karnataka	4,48,17,398
Kerala	2,90,11,237
Pondicherry	7,89,416

1.2.3 Spacio Temporal Data:

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

Exampl:

Stat e	Population	
	1981	1991
Tamil Nadu	4,82,97,456	5,56,38,318
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

1.2.4 Categories of data:

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

These categories are,

1.Primary data 2. Secondary data

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.

Example:

If a researcher is interested to know the impact of noon- meal scheme for the school children, he has to undertake a survey and collect data on the opinion of parents and children by asking relevant questions. Such a data collected for the purpose is called primary data.

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

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

1.Direct personal interviews:

The persons from whom informations are collected are known as informants. The investigator personally meets them and asks questions to gather the necessary informations. It is the suitable method for intensive rather than extensive field surveys. It suits best for intensive study of the limited field.

Merits:

1.People willingly supply informations because they are approached personally. Hence, more response noticed in this method than in any other method.

2.The collected informations are likely to be uniform and accurate. The investigator is there to clear the doubts of the informants.

3.Supplementary informations on informant’ s personal aspects can be noted. Informations on character and environment may help later to interpret some of the results.

4.Answers for questions about which the informant is likely to be sensitive can be gathered by this method.

5.The wordings in one or more questions can be altered to suit any informant. Explanations may be given in other languages also. Inconvenience and misinterpretations are thereby avoided.

Limitations:

1.It is very costly and time consuming.

2.It is very difficult, when the number of persons to be interviewed is large and the persons are spread over a wide area.

3.Personal prejudice and bias are greater under this method.

2.Indirect Oral Interviews:

Under this method the investigator contacts witnesses or neighbours or friends or some other third parties who are capable of supplying the necessary information. This method is preferred if the required information is on addiction or cause of fire or theft or murder etc., If a fire has broken out a certain place, the persons living in neighbourhood and witnesses are likely to give information on the cause of fire. In some cases, police interrogated third parties who are supposed to have knowledge of a theft or a murder and get some clues. Enquiry committees appointed by governments generally adopt this method and get people' s views and all possible details of facts relating to the enquiry. This method is suitable whenever direct sources do not exists or cannot be relied upon or would be unwilling to part with the information.

The validity of the results depends upon a few factors, such as the nature of the person whose evidence is being recorded, the ability of the interviewer to draw out information from the third parties by means of appropriate questions and cross examinations, and the number of persons interviewed. For the success of this method one person or one group alone should not be relied upon.

3.Information from correspondents:

The investigator appoints local agents or correspondents in different places and compiles the information sent by them. Informations to Newspapers and some departments of Government come by this method. The advantage of this method is that it is cheap and appropriate for extensive investigations. But it may not ensure accurate results because the correspondents are likely to be negligent, prejudiced and biased. This method is adopted in those cases where informations are to be collected periodically from a wide area for a long time.

4.Mailed questionnaire method:

Under this method a list of questions is prepared and is sent to all the informants by post. The list of questions is technically called questionnaire. A covering letter accompanying the questionnaire explains the purpose of the investigation and the importance of correct informations and request the informants to fill in the blank spaces provided and to return the form within a specified time. This method is appropriate in those cases where the informants are literates and are spread over a wide area.

Merits:

1.It is relatively cheap.

2.It is preferable when the informants are spread over the wide area.

Limitations:

1.The greatest limitation is that the informants should be literates who are able to understand and reply the questions.

2.It is possible that some of the persons who receive the questionnaires do not return them.

3.It is difficult to verify the correctness of the informations furnished by the respondents.

With the view of minimizing non-respondents and collecting correct information, the questionnaire should be carefully drafted. There is no hard and fast rule. But the following general principles may be helpful in framing the questionnaire. A covering letter and a self addressed and stamped envelope should accompany the questionnaire. The covering letter should politely point out the purpose of the survey and privilege of the respondent who is one among the few associated with the investigation. It should assure that the informations would be kept confidential and would never be misused. It may promise a copy of the findings or free gifts or concessions etc.,

Characteristics of a good questionnaire:

1.Number of questions should be minimum.

2.Questions should be in logical orders, moving from easy to more difficult questions.

3.Questions should be short and simple. Technical terms and vague expressions capable of different interpretations should be avoided.

4.Questions fetching YES or NO answers are preferable. There may be some multiple-choice questions requiring lengthy answers are to be avoided.

5.Personal questions and questions which require memory power and calculations should also be avoided.

6.Question should enable cross check. Deliberate or unconscious mistakes can be detected to an extent.

7.Questions should be carefully framed so as to cover the entire scope of the survey.

8.The wording of the questions should be proper without hurting the feelings or arousing resentment.

9.As far as possible confidential informations should not be sought.

10.Physical appearance should be attractive, sufficient space should be provided for answering each questions.

5.Schedules sent through Enumerators:

Under this method enumerators or interviewers take the schedules, meet the informants and filling their replies. Often distinction is made between the schedule and a questionnaire. A schedule is filled by the interviewers in a face-to-face situation with the informant. A questionnaire is filled by the informant which he receives and returns by post. It is suitable for extensive surveys.

Merits:

1.It can be adopted even if the informants are illiterates.

2.Answers for questions of personal and pecuniary nature can be collected.

3.non-response is minimum as enumerators go personally and contact the informants.

4.The information collected are reliable. The enumerators can be properly trained for the same.

5.It is most popular methods.

Limitations:

1.It is the costliest method.

2.Extensive training is to be given to the enumerators for collecting correct and uniform informations.

3.Interviewing requires experience. Unskilled investigators are likely to fail in their work.

Before the actual survey, a pilot survey is conducted. The questionnaire/Schedule is pre-tested in a pilot survey. A few among the people from whom actual information is needed are asked to reply. If they misunderstand a question or find it difficult to answer or do not like its wordings etc., it is to be altered. Further it is to be ensured that every questions fetches the desired answer.

Merits and Demerits of primary data:

1.The collection of data by the method of personal survey is possible only if the area covered by the investigator is small. Collection of data by sending the enumerator is bound to be expensive. Care should be taken twice that the enumerator record correct information provided by the informants.

2.Collection of primary data by framing a schedules or distributing and collecting questionnaires by post is less expensive and can be completed in shorter time.

3.Suppose the questions are embarrassing or of complicated nature or the questions probe into personnel affairs of individuals, then the schedules may not be filled with accurate and correct information and hence this method is unsuitable.

4.The information collected for primary data is mere reliable than those collected from the secondary data.

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:

1.Published sources,

2.Unpublished sources,

1.Published Sources

The various sources of published data are:

1**.**Reports and official publications of

(i)International bodies such as the International Monetary Fund, International Finance Corporation and United Nations Organisation.

(ii)Central and State Governments such as the Report of the Tandon Committee and Pay Commission.

2.Semi-official publication of various local bodies such as Municipal Corporations and District Boards.

3.Private publications-such as the publications of –

(i)Trade and professional bodies such as the Federation of Indian Chambers of Commerce and Institute of Chartered Accountants.

(ii)Financial and economic journals such as ‘ Commerce’ , ‘ Capital’ and ‘ Indian Finance’

(iii)Annual reports of joint stock companies.

(iv)Publications brought out by research agencies, research scholars, etc.

It should be noted that the publications mentioned above vary with regard to the periodically of publication. Some are published at regular intervals (yearly, monthly, weekly etc.,) whereas others are ad hoc publications, i.e., with no regularity about periodicity of publications.

Note: A lot of secondary data is available in the internet. We can access it at any time for the further studies.

2.Unpublished Sources

All statistical material is not always published. There are various sources of unpublished data such as records maintained by various Government and private offices, studies made by research institutions, scholars, etc. Such sources can also be used where necessary

Precautions in the use of Secondary data

The following are some of the points that are to be considered in the use of secondary data

1.How the data has been collected and processed

2.The accuracy of the data

3.How far the data has been summarized

4.How comparable the data is with other tabulations

5.How to interpret the data, especially when figures collected for one purpose is used for another

Generally speaking, with secondary data, people have to compromise between what they want and what they are able to find.

Merits and Demerits of Secondary Data:

1.Secondary data is cheap to obtain. Many government publications are relatively cheap and libraries stock quantities of secondary data produced by the government, by companies and other organisations.

2.Large quantities of secondary data can be got through internet.

3.Much of the secondary data available has been collected for many years and therefore it can be used to plot trends.

4.Secondary data is of value to:

- The government – help in making decisions and planning future policy.
- Business and industry – in areas such as marketing, and sales in order to appreciate the general economic and social conditions and to provide information on competitors.
- Research organisations – by providing social, economical and industrial information.

1.3. Tabulation:

Tabulation 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.

1.3.1 Advantages of Tabulation:

Statistical data arranged in a tabular form serve following objectives:

1.It simplifies complex data and the data presented are easily understood.

2.It facilitates comparison of related facts.

3.It facilitates computation of various statistical measures like averages, dispersion, correlation etc.

4.It presents facts in minimum possible space and unnecessary repetitions and explanations are avoided. Moreover, the needed information can be easily located.

5Tabulated data are good for references and they make it easier to present the information in the form of graphs and diagrams.

1.3.2 Preparing a 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:

1.Table number

2.Title of the table

3.Captions or column headings

4Stubs or row designation

5.Body of the table

6.Footnotes

7.Sources of data

1.Table Number:

A table should be numbered for easy reference and identification. This number, if possible, should be written in the centre at the top of the table. Sometimes it is also written just before the title of the table.

2.Title:

A good table should have a clearly worded, brief but unambiguous title explaining the nature of data contained in the table. It should also state arrangement of data and the period covered. The title should be placed centrally on the top of a table just below the table number (or just after table number in the same line).

3.Captions or column Headings:

Captions in a table stands for brief and self explanatory headings of vertical columns. Captions may involve headings and sub-headings as well. The unit of data contained should also be given for each column. Usually, a relatively less important and shorter classification should be tabulated in the columns.

4.Stubs or Row Designations:

Stubs stands for brief and self explanatory headings of horizontal rows. Normally, a relatively more important classification is given in rows. Also a variable with a large number of classes is usually represented in rows. For example, rows may stand for score of classes and columns for data related to sex of students. In the process, there will be many rows for scores classes but only two columns for male and female students.

A model structure of a table is given below:

Table Number Title of the Table

Sub Heading	Caption Headings	Total
	Caption Sub-Headings	
Stub Sub-Headings	Body	
Total		

Foot notes:

Sources Note:

5.Body:

The body of the table contains the numerical information of frequency of observations in the different cells. This arrangement of data is according to the discription of captions and stubs.

6.Footnotes:

Footnotes are given at the foot of the table for explanation of any fact or information included in the table which needs some explanation. Thus, they are meant for explaining or providing further details about the data, that have not been covered in title, captions and stubs.

7.Sources of data:

Lastly one should also mention the source of information from which data are taken. This may preferably include the name of the author, volume, page and the year of publication. This should also state whether the data contained in the table is of ‘ primary or secondary’ nature.

Requirements of a Good 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 he points to be stressed.

Though, there is no hard and fast rule for forming a table yet a few general point should be kept in mind:

1.A table should be formed in keeping with the objects of statistical enquiry.

2.A table should be carefully prepared so that it is easily understandable.

3.A table should be formed so as to suit the size of the paper. But such an adjustment should not be at the cost of legibility.

4.If the figures in the table are large, they should be suitably rounded or approximated. The method of approximation and units of measurements too should be specified.

5.Rows and columns in a table should be numbered and certain figures to be stressed may be put in ‘ box’ or ‘ circle’ or in bold letters.

6.The arrangements of rows and columns should be in a logical and systematic order. This arrangement may be alphabetical, chronological or according to size.

7.The rows and columns are separated by single, double or thick lines to represent various classes and sub-classes used. The corresponding proportions or percentages should be given in adjoining rows and columns to enable comparison. A vertical expansion of the table is generally more convenient than the horizontal one.

8.The averages or totals of different rows should be given at the right of the table and that of columns at the bottom of the table. Totals for every sub-class too should be mentioned.

9.In case it is not possible to accommodate all the information in a single table, it is better to have two or more related tables.

Type of Tables:

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 may be classified as follows:

1. Simple or one-way table

2. Two way table

3. Manifold table

1.Simple or one-way Table:

A simple or one-way table is the simplest table which contains data of one characteristic only. A simple table is easy to construct and simple to follow. For example, the blank table given below may be used to show the number of adults in different occupations in a locality.

The umber of adults in a locality in respect of occupation and sex

Occupations	No. Of Adults
Total	

2.Two-way Table:

A table, which contains data on two characteristics, is called a two way table. In such case, therefore, either stub or caption is divided into two co-ordinate parts. In the given table, as an example the caption may be further divided in respect of ‘ sex’ . This subdivision is shown in two-way table, which now contains two characteristics namely, occupation and sex.

Occupation	No. of Adults		Total
	Male	Female	
Total			

3.Manifold Table:

Thus, more and more complex tables can be formed by including other characteristics. For example, we may further classify the caption sub-headings in the above table in respect of “marital status”, “ religion” and “socio-economic status” etc. A table ,which has more than two characteristics of data is considered as a manifold table. For instance , table shown below shows three characteristics namely, occupation, sex and marital status.

Occupation	No. of Adults						Total
	Male			Female			
	M	**U**	**Total**	**M**	**U**	**Total**	
Total							

Foot note: M Stands for Married and U stands for unmarried.

Manifold tables, though complex are good in practice as these enable full information to be incorporated and facilitate analysis of all related facts. Still, as a normal practice, not more than four characteristics should be represented in one table to avoid confusion. Other related tables may be formed to show the remaining characteristics

1.4. DIAGRAMATIC AND GRAPHICAL REPRESENTATION

In the previous chapter, we have discussed the techniques of classification and tabulation that help in summarising the collected data and presenting them in a systematic manner. However, these forms of presentation do not always prove to be interesting to the common man. 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.

1.4.1 Diagrams:

A 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.

1.They are attractive and impressive.

2.They make data simple and intelligible.

3.They make comparison possible

4.They save time and labour.

5.They have universal utility.

6.They give more information.

7.They have a great memorizing effect.

1.4.2. 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.

1.A diagram should be neatly drawn and attractive.

2.The measurements of geometrical figures used in diagram should be accurate and proportional.

3.The size of the diagrams should match the size of the paper.

4.Every diagram must have a suitable but short heading.

5.The scale should be mentioned in the diagram.

6.Diagrams should be neatly as well as accurately drawn with the help of drawing instruments.

7.Index must be given for identification so that the reader can easily make out the meaning of the diagram.

8.Footnote must be given at the bottom of the diagram.

9.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:

1.One-dimensional diagrams

2.Two-dimensional diagrams

3.Three-dimensional diagrams

4.Pictograms and Cartograms

1.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

1.Line Diagram

2.Simple Diagram

3.Multiple Bar Diagram

4.Sub-divided Bar Diagram

5.Percentage Bar Diagram

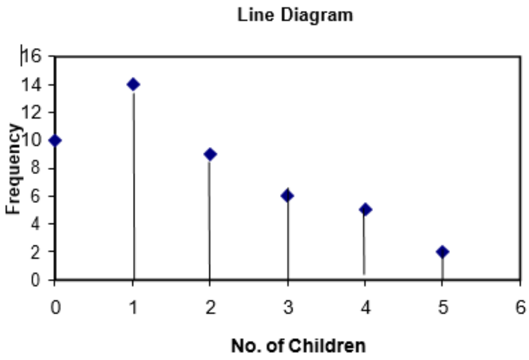
1.Line Diagram:

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



2.Simple Bar Diagram:

Simple bar diagram can be drawn either on horizontal or vertical base, but bars on horizontal base more common. Bars must be uniform width and intervening space between bars must be equal.While constructing a simple bar diagram, the scale is determined on the basis of the highest value in the series. To make the diagram attractive, the bars can be coloured. Bar diagram are used in business and economics.

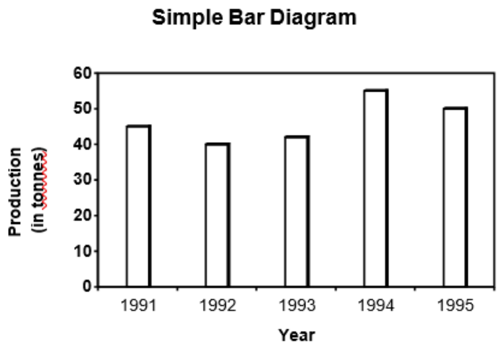
However, an important limitation of such diagrams is that they can present only one classification or one category of data. For example, while presenting the population for the last five decades, one can only depict the total population in the simple bar diagrams, and not its sex-wise distribution.

Example :

Represent the following data by a bar diagram

Year	Production (in tones)
1991	45
1992	40
1993	42
1994	55
1995	50

Solution:



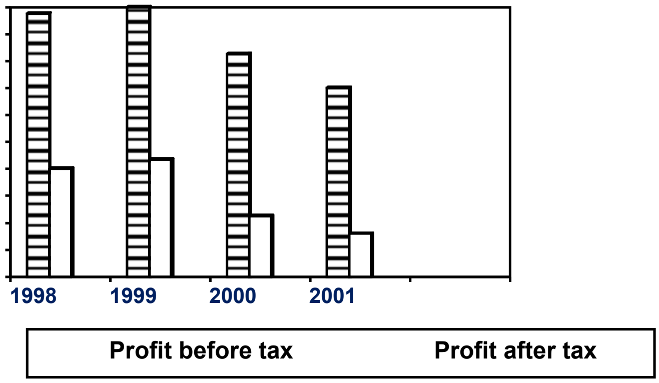
3.Multiple Bar Diagram:

Multiple bar diagram is used for comparing two or more sets of statistical data. Bars are constructed side by side to represent the set of values for comparison. In order to distinguish bars, they may be either differently coloured or there should be different types of crossings or dotting, etc. An index is also prepared to identify the meaning of different colours or dottings.

Example :

Draw a multiple bar diagram for the following data.

Year	Profit before tax (in lakhs of rupees)	Profit after tax (in lakhs of rupees)
1998	195	80
1999	200	87
2000	165	45
2001	140	32



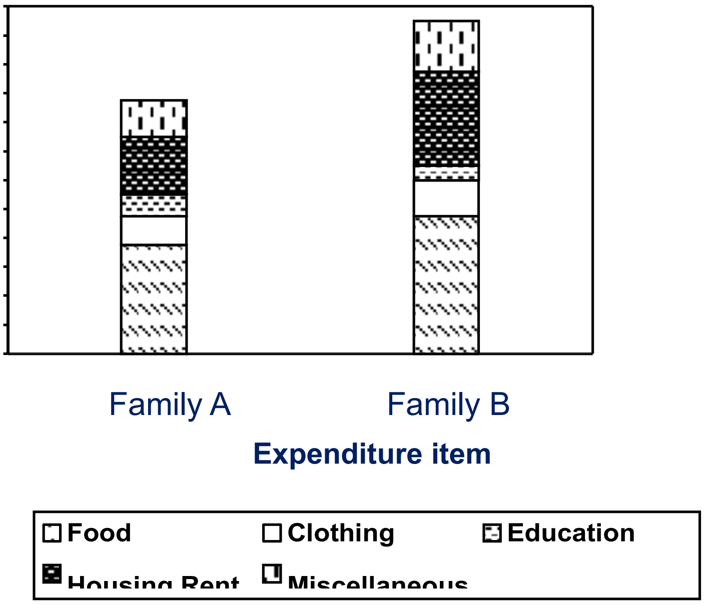
4.Sub-divided Bar Diagram:

In a sub-divided bar diagram, the bar is sub-divided into various parts in proportion to the values given in the data and the whole bar represent the total. Such diagrams are also called Component Bar diagrams. The sub divisions are distinguished by different colours or crossings or dottings. The main defect of such a diagram is that all the parts do not have a common base to enable one to compare accurately the various components of the data.

Example:

Represent the following data by a sub-divided bar diagram.

Expenditure items	Monthly expenditure (in Rs.)	
	Family A	Family B
Food	75	95
Clothing	20	25
Education	15	10
Housing Rent	40	65
Miscellaneous	25	35



5.Percentage bar diagram:

This is another form of component bar diagram. Here the components are not the actual values but percentages of the whole. The main difference between the sub-divided bar diagram and percentage bar diagram is that in the former the bars are of different heights since their totals may be different whereas in the latter the bars are of equal height since each bar represents 100 percent. In the case of data having sub-division, percentage bar diagram will be more appealing than sub-divided bar diagram.

Example :

Represent the following data by a percentage bar diagram

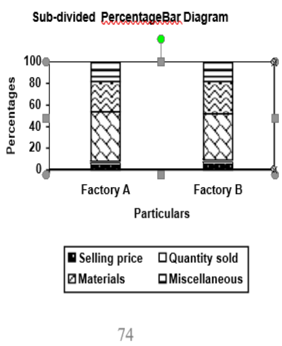
Particular	Factory A	Factory B
Selling Price	400	650
Quantity Sold	240	365
Wages	3500	5000
Materials	2100	3500
Miscellaneous	1400	2100

Solution:

Convert the given values into percentages as follows:

Particulars	Factory A		Factory B	
	Rs.	%	Rs.	%
Selling Price	400	5	650	6
Quantity Sold	240	3	365	3
Wages	3500	46	5000	43
Materials	2100	28	3500	30
Miscellaneous	1400	18	2100	18

Total	7640	100	11615	100



1.4.3. Pie Diagram or Circular Diagram:

Another way of preparing a two-dimensional diagram is in the form of circles. In such diagrams, both the total and the component parts or sectors can be shown. The area of a circle is proportional to the square of its radius.

While making comparisons, pie diagrams should be used on a percentage basis and not on an absolute basis. In constructing a pie diagram the first step is to prepare the data so that various components values can be transposed into corresponding degrees on the circle.

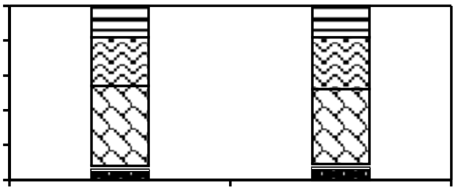
The second step is to draw a circle of appropriate size with a compass. The size of the radius depends upon the available space and other factors of presentation. The third step is to measure points on the circle and representing the size of each sector with the help of a protractor.

Example :

Draw a Pie diagram for the following data of production of sugar in quintals of various countries.

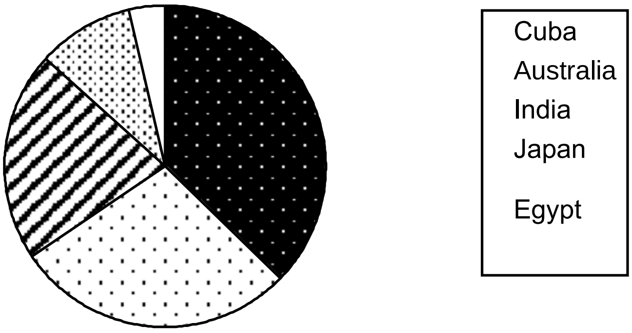
Country	Production of Sugar (in quintals)
Cuba	62
Australia	47
India	35
Japan	16
Egypt	6

Solution:



The values are expressed in terms of degree as follows.

Country	Production of Sugar	
	In Quintals	In Degrees
Cuba	62	134
Australia	47	102
India	35	76
Japan	16	35
Egypt	6	13
Total	166	360



1.5. Graphs:

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.

However here we shall discuss only some important types of graphs which are more popular and they are

1.Histogram 2. Frequency Polygon 3.Frequency Curve 4. Ogive 5. Lorenz Curve

1.Histogram:

A histogram 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.

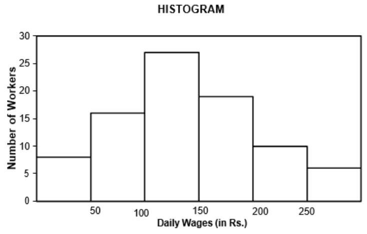
However, we cannot construct a histogram for distribution with open-end classes. It is also quite misleading if the distribution has unequal intervals and suitable adjustments in frequencies are not made.

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

Solution:



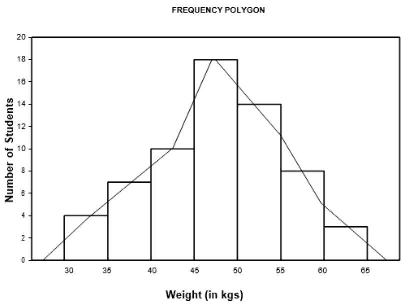
2.Frequency Polygon:

If we mark the midpoints of the top horizontal sides of the rectangles in a histogram and join them by a straight line, the figure so formed is called a Frequency Polygon. This is done under the assumption that the frequencies in a class interval are evenly distributed throughout the class. The area of the polygon is equal to the area of the histogram, because the area left outside is just equal to the area included in it.

Example:

Draw a frequency polygon for the following data.

Weight (in kg)	Number of Students
30-35	4
35-40	7
40-45	10
45-50	18
50-55	14
55-60	8
60-65	3



3.Frequency Curve:

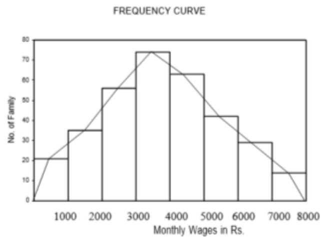
If the middle point of the upper boundaries of the rectangles of a histogram is corrected by a smooth freehand curve, then that diagram is called frequency curve. The curve should begin and end at the base line.

Example 12:

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

Solution:



4.Ogives:

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 a 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:

1.The ‘ less than ogive’ method

2.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 :

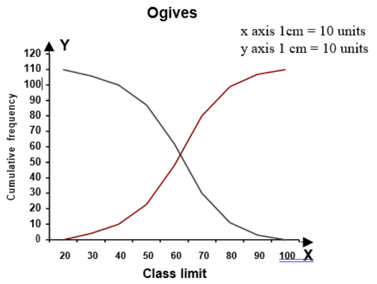
Draw the Ogives for the following data.

Class interval	Frequency
20-30	4
30-40	6
40-50	13
50-60	25
60-70	32
70-80	19

80-90	8
90-100	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



5.Lorenz Curve:

Lorenz curve is a graphical method of studying dispersion. It was introduced by Max.O.Lorenz, a great Economist and a statistician, to study the distribution of wealth and income. It is also used to study the variability in the distribution of profits, wages, revenue, etc.

It is specially used to study the degree of inequality in the distribution of income and wealth between countries or between different periods. It is a percentage of cumulative values of one variable in combined with the percentage of cumulative values in other variable and then Lorenz curve is drawn.

The curve starts from the origin (0,0) and ends at (100,100). If the wealth, revenue, land etc are equally distributed among the people of the country, then the Lorenz curve will be the diagonal of the square. But this is highly impossible.

The deviation of the Lorenz curve from the diagonal, shows how the wealth, revenue, land etc are not equally distributed among people.

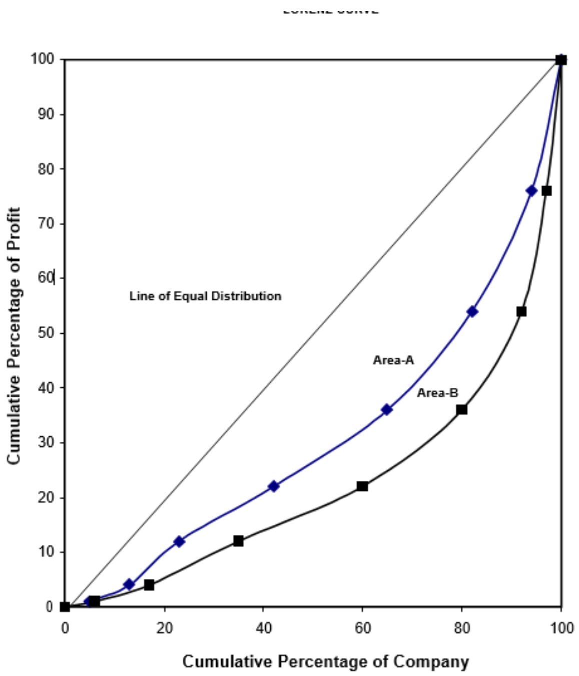
Example:

In the following table, profit earned is given from the number of companies belonging to two areas A and B. Draw in the same diagram their Lorenz curves and interpret them.

Profit earned (in thousands)	Number of Companies	
	Area A	Area B
5	7	13
26	12	25
65	14	43
89	28	57
110	33	45
155	25	28
180	18	13
200	8	6

Solution:

Profits			Area A			Area B		
In Rs.	Cumulative profit	Cumulative percentage	No. of companies	Cumulative number	Cumulative percentage	No. of companies	Cumulative number	Cumulative percentage
5	5	1	7	7	5	13	13	6
26	31	4	12	19	13	25	38	17
65	96	12	14	33	23	43	81	35
89	185	22	28	61	42	57	138	60
110	295	36	33	94	65	45	183	80
155	450	54	25	119	82	28	211	92
180	630	76	18	137	94	13	224	97
200	830	100	8	145	100	6	230	100



1.6. Measures of Central Tendency

In the study of a population with respect to one in which we are interested we may get a large number of observations. 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."

Characteristics for a good or an ideal average:

The following properties should possess for an ideal average.

1. It should be rigidly defined.

2. It should be easy to understand and compute.

3. It should be based on all items in the data.

4. Its definition shall be in the form of a mathematical formula.

5. It should be capable of further algebraic treatment.

6. It should have sampling stability.

7. It should be capable of being used in further statistical computations or processing.

Besides the above requisites, a good average should represent maximum characteristics of the data, its value should be nearest to the most items of the given series

1.6.1. Arithmetic mean or mean :

Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. If the variable x assumes n values x_1, x_2 …x_n then the mean, $\bar{x}$, is given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots xn}{n}$$

This formula is for the ungrouped or raw data.

Example:

Calculate the mean for 2, 4, 6, 8, 10

Solution:

$$\bar{x} = \frac{2+4+6+8+10}{5}$$

= 30/5

=6

1.6.2.Short-Cut method:

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

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

where, A = the assumed mean or any value in x

d = the deviation of each value from the assumed mean

Solution:

X	**d=x-A**
75	7
A **68**	0
80	12
92	24
56	-12
Total	31

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

$$\bar{x} = 68 + \frac{31}{5}$$

$$= 74.2$$

Grouped Data :

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

$$\bar{x} = \frac{\Sigma f x}{N}$$

where x = the mid-point of individual class

f = the frequency of individual class

N = the sum of the frequencies or total frequencies.

$$\bar{x} = A + \frac{\Sigma f d}{N} \text{ x c}$$

Where d= $\frac{X-A}{c}$

A = any value in x
N = total frequency
c = width of the class interval

Example :

Given the following frequency distribution, calculate the arithmetic mean

Marks Number of

Marks : 64 63 62 61 60 59

Number of Students : 8 18 12 9 7 6

X	F	fx	d=x-A	fd
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	□1	□9
60	7	420	□2	□1
59	6	354	□3	4 □1 8
	60	3713		- 7

Direct method:

$$\frac{\Sigma fx}{N} = \frac{3713}{60} = 61.88$$

Short-cut method

$$\bar{x} = A + \frac{\Sigma d}{n} = 62 \ - \ 7/60 = 61.88$$

Merits and demerits of Arithmetic mean :

Merits:

1.It is rigidly defined.

2.It is easy to understand and easy to calculate.

3.If the number of items is sufficiently large, it is more accurate and more reliable.

4.It is a calculated value and is not based on its position in the series.

5.It is possible to calculate even if some of the details of the data are lacking.

6.Of all averages, it is affected least by fluctuations of sampling.

7.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 calculated for open-end classes.

6.It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

1.7. Median:

The median is that value of the variate which divides the group into two equal parts, one part comprising all values greater, and the other, all values less than median.

Ungrouped or Raw data :

Arrange the given values in the increasing or decreasing order. If the number of values are odd, median is the middle valueIf the number of values are even, median is the mean of middle two values.

By formula

Median = Md=($\frac{n+1}{2}$)th item

Example:

When odd number of values are given. Find median for the following data

25, 18, 27, 10, 8, 30, 42, 20, 53

Solution:

Arranging the data in the increasing order 8, 10, 18, 20, 25, 27, 30, 42, 53

The middle value is the 5th item i.e., 25 is the median

Using formula

Median = Md=($\frac{n+1}{2}$)th item

= 9+1/2

=10/2

=5th item

= 25

Example:

When even number of values are given. Find median for the following data

5, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the datain the increasing order 2, 5, 8, 10, 12, 18, 22, 30

Here median is the mean of the middle two items (ie) mean of (10,12) ie

= 10+12 / 2 = 11

Median = 11

Using the formula

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

2

$$= \left(\frac{8+1}{2}\right)^{\text{th}} \text{ item.}$$

$$= \left(\frac{9}{2}\right)^{\text{th}} \text{ item } = 4.5^{\text{th}} \text{ item}$$

$$= 4^{\text{th}} \text{ item} + \left(\frac{1}{2}\right)(5^{\text{th}} \text{ item} - 4^{\text{th}} \text{ item})$$

$$= 10 + \left(\frac{1}{2}\right)[12\text{-}10]$$

$$= 10 + \left(\frac{1}{2}\right) \times 2$$

$$= 10 + 1$$

$$= 11$$

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.

1.7.1.Cumulative frequency :(cf)

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

Discrete Series:

Step1: Find cumulative frequencies

Step 2 :Find N+1 / 2

Step3: See in the cumulative frequencies the value just greater than

Step4: Then the corresponding value of x is median.

Example:

The followingdata pertaining to the number of members in a family. Find median size of the family.

Number of members **x**	1	2	3	4	5	6	7	8	9	10	11	12
Frequency F	1	3	5	6	10	13	9	5	3	2	2	1

Solution:

X	f	cf
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
	60	

Median = size of $(n+1/2)^{th}$ item

= size of $(60+1/2)^{th}$ item

= 30.5^{th} item

The cumulative frequencies just greater than 30.5 is 38.and the value of x corresponding to 38 is 6. Hence the median size is 6 members per family

Note:

It is an appropriate method because a fractional value given by mean does not indicate the average number of members in a family.

1.7.2.Continuous Series:

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

Step1: Find cumulative frequencies.

Step 2 : Find(N/2)

Step3: See in the cumulative frequency the value first greater than

N/2, Then the corresponding class interval is called the Median class. Then apply the formula

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

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:

The following table gives the frequency distribution of 325 workers of a factory, according to their average monthly income in a certain year.

Income group (in Rs)	Number of workers
Below 100	1
100-150	20
150-200	42
200-250	55
250-300	62
300-350	45
350-400	30
400-450	25
450-500	15
500-550	18
550-600	10
600 and above	2
	325

Calculate median income

Solution:

Income group (Class-interval)	Number of workers (Frequency)	Cumulative frequency c.f
Below 100	1	1
100-150	20	21
150-200	42	63
200-250	55	118
250-300	62	180
300-350	45	225
350-400	30	255
400-450	25	280
450-500	15	295
500-550	18	313
550-600	10	323
600 and above	2	325
	325	

$$\frac{N}{2} = \frac{325}{2} = 162.5$$

Here $l = 250$, N = 325, f = 62, c = 50, m = 118

$$\text{Md} = 250 + \left(\frac{162.5 - 118}{62}\right) \times 50$$

$$= 250 + 35.89$$

$$= 285.89$$

Merits of Median :

1. Median is not influenced by extreme values because it is apositional average.
2. Median can be calculated in case of distribution with openend intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median :

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series,median is an estimated value other than any value in theseries.
3. It is not suitable for further mathematical treatment except its use in mean deviation.
4. It is not taken into account all the observations.

1.8. Mode :

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it.

According to Croxton and Cowden " The mode of a distribution is the value at the point around which the items tend to be most heavily concentrated. It may be regarded at the most typical of a series of values". It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order for shoes or ready-made garments the modal size helps because this sizes and other sizes around in common demand.

Computation of the mode:

Ungrouped or Raw Data:

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

Example :

2 , 7, 10, 15, 10, 17, 8, 10, 2

Mode = M_0 =10

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

Example:

1. 12, 10, 15, 24, 30 (no mode)

2. 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

the modes are 7 and 10

Grouped Data:

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

1.8.1Continuous distribution:

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

$$\text{Mode} = M_0 = l + \frac{\triangle_1}{\triangle_1 + \triangle_2} \times C$$

$$l = \text{Lower limit of the model class}$$

$$\triangle_1 = f_1 - f_0$$
$$\triangle_2 = f_1 - f_2$$

f1 = frequency of the modal class
f0 = frequency of the class preceding the modal class
f2 = frequency of the class succeeding the modal class

The above formula can also be written as

$$\text{Mode} = l \quad \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

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. If mode lies in the first class interval, then f0 is taken as zero.
3. The computation of mode poses no problem in distributions with open-end classes, unless the modal value lies in the open-end class.

4.

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

Example:

Calculate mode for the following :

C- I	f
0-50	5
50-100	14
100-150	40
150-200	91
200-250	150
250-300	87
300-350	60
350-400	38
400 and above	15

Solution:

The highest frequency is 150 and corresponding class interval is 200 – 250, which is the modal class.

Here l=200, f1=150, f0=91, f2=87, C=50

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

$$= 200 + \frac{150\text{-}91}{2 \times 150 - 91 - 87} \times 50$$

$$= 200 + \frac{2950}{122}$$

$$= 200 + 24.18 = 224.18$$

Merits of Mode:

1.It is easy to calculate and in some cases it can be located mere inspection

2.Mode is not at all affected by extreme values.

3.It can be calculated for open-end classes.

4.It is usually an actual value of an important part of the series.

5.In some circumstances it is the best representative of data.

Demerits of mode:

1.It is not based on all observations.

2.It is not capable of further mathematical treatment.

3.Mode is ill-defined generally, it is not possible to find mode in some cases.

4.As compared with mean, mode is affected to a great extent, by sampling fluctuations.

5.It is unsuitable in cases where relative importance of items has to be considered.

1.9. 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 x1,x2…..xn are n observations,

$$\text{H.M} = \frac{n}{\sum_{i=1}^{n}\left(\frac{1}{x_i}\right)}$$

For a frequency distribution

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

Example:

From the given data calculate H.M 5,10,17,24,30

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

$$\text{H.M} = \frac{n}{\Sigma\left[\frac{1}{x}\right]}$$

$$= \frac{5}{0.4338} = 11.526$$

Example:

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

Solution:

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

$$\text{H.M} = \frac{N}{\sum f\left[\frac{1}{x}\right]}$$

$$= \frac{18}{0.1968} = 21.91$$

Merits of H.M:

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 H.M:

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

1.10. Geometric mean :

The geometric mean of a series containing n observations is the nth root of the product of the values. If x1,x2…, xn are observations then

$$\text{G.M} = \sqrt[n]{x_1 . x_2 \ldots x_n}$$

$$= (x_1.x_2 \,..\, x_n)^{1/n}$$

$$\log \text{GM} = \frac{1}{n} \log(x_1.x_2 \,..\, x_n)$$

$$= \frac{1}{n} (\log x_1 + \log x_2 + ..+ \log x_n$$

$$= \frac{\sum \log x_i}{n}$$

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

For grouped data

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

Calculate the geometric mean of the following series of monthly income of a batch 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

$$\text{GM} = \text{Antilog} \left[\frac{\sum \log x}{n} \right]$$

$$= \text{Antilog} \frac{13.5107}{5}$$

$$= \text{Antilog } 2.7021 = 503.6$$

Calculate the average income per head from the data given below .Use geometric mean.

Class of people	Number of families	Monthly income per head (Rs)
Landlords	2	5000
Cultivators	100	400

Landless – labours	50	200
Money – lenders	4	3750
Office Assistants	6	3000
Shop keepers	8	750
Carpenters	6	600
Weavers	10	300

Solution:

Class of people	Annual income (Rs) X	Number of families (f)	Log x	f logx
Landlords	5000	2	3.6990	7.398
Cultivators	400	100	2.6021	260.210
Landless – labours	200	50	2.3010	115.050
Money – lenders	3750	4	3.5740	14.296
Office Assistants	3000	6	3.4771	20.863
Shop keepers	750	8	2.8751	23.2008
Carpenters	600	6	2.7782	16.669
Weavers	300	10	2.4771	24.771
		186		482.257

$$\text{GM} = \text{Antilog}\left[\frac{\sum f\log x}{N}\right]$$

$$= \text{Antilog}\left[\frac{482.257}{186}\right]$$

$$= \text{Antilog}\ (2.5928)$$

$$= \text{Rs } 391.50$$

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 thereis 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

Unit I

Self-Assessment

1.Determine the median from the following data

25, 20, 15, 45, 18, 7, 10, 38, 12

2. Find the combined mean from the following data

$\bar{x}_1 = 210$ $n_1=50$, $\bar{x}_2 = 150$ $n_2 = 100$

3. The monthly income of ten families(in rupees) in a certain locality are given below.

Family	A	B	C	D	E	F	G
Income (inrupees)	30	70	60	100	200	150	300

Calculate the arithmetic average by Short-cut method

4. The monthly income of 8 families is given below. Find GM

Family	A	B	C	D	E	F	G	H
income	70	10	500	75	8	250	8	42

Suggested Activities:

1. Measure the heights and weights of your class students.

Find the mean, median, mode and compare

2. Find the mean marks of your class students in various subjects.

Ans:

1.MD= 18

2.170

3.130

4.45.27

MEASURES OF DISPERSION –SKEWNESS AND KURTOSIS

OBJECTIVE:

Measures of dispersion are essential for understanding the spread and variability in a dataset, providing context to central tendency measures, identifying patterns and outliers, and supporting more informed decision-making and risk assessment.

2.1. INTRODUCTION :

The measure of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. 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. For example consider the following marks of two students.

Student I	Student II
68	85
75	90
65	80
67	25
70	65

Both have got a total of 345 and an average of 69 each. The fact is that the second student has failed in one paper. When the averages alone are considered, the two students are equal. But first student has less variation than second student. Less variation is a desirable characteristic.

Characteristics of a good measure of dispersion:

An ideal measure of dispersion is expected to possessthe 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

1.Absolute measure of dispersion

2.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	**Relative measure**
1. Range	1.Co-efficient of Range
2.Quartile deviation	. 2.Co-efficient of Quartile deviation
3.Mean deviation	3. Co-efficient of Mean deviation
4.Standard deviation	4.Co-efficient of variation

Range and coefficient of Range:

1.Range:

This is the simplest possible measure of dispersion and is defined as the difference between the largest and smallest values of the variable.

In symbols,

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:

Co – efficient of range = L - S / L+S

Example:

Find the value of range and its co-efficient for the following data. 7, 9, 6, 8, 11, 10, 4

Solution:

L=11, S = 4.

Range = L– S = 11- 4 = 7

Co-efficient of Range = L - S / L+S

= 11-4 / 11+4

= 7 / 15

= 0.4667

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, et c., 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

2.2. Quartile Deviation and Co efficient of Quartile Deviation:

Quartile Deviation (Q.D): Definition: Quartile Deviation is half of the difference between the first and third quartiles. Hence, it is called Semi Inter Quartile Range. In symbols

Q.D = $\frac{Q_3-Q_1}{2}$ Among the quartiles Q1, Q2 2 and Q3, the range Q3 − Q1 is called inter quartile range and Q3 − Q1 / 2 , Semi inter quartile range

Co-efficient of Quartile Deviation :

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

Example :

Find the Quartile Deviation for the following data: 391, 384, 591, 407, 672, 522, 777, 733, 1490, 2488

Solution: Arrange the given values in ascending order. 384, 391, 407, 522, 591, 672, 733, 777, 1490, 2488

Position of Q_1 is $\frac{n+1}{4} = \frac{10+1}{4} =$ 2.75th item

Q_1 = 2nd value + 0.75 (3rd value– 2nd value)

= 391 + 0.75 (407– 391)

= 391 + 0.75× 16

= 391 + 12 = 403

Position of Q_1 is 3 $\frac{n+1}{4} =$ 8.25th item

Q3= 8th value + 0.25 (9th value– 8th value)

= 777 + 0.25 (1490– 777)

= 777 + 0.25 (713)

= 777 + 178.25 = 955.25

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

$= \frac{955.25-403}{2}$

= 276.125

Example:

Weekly wages of labours are given below. Calculated Q.D and Coefficient of Q.D.

Weekly Wage (Rs.)	:100	200	400	500	600
No. of Weeks	: 5	8	21	12	6

Solution:

Weekly Wage (Rs.)	No. of Weeks	Cum. No. of Weeks
100	5	5
200	8	13
400	21	34
500	12	46
600	6	52
Total	N=52	

Position of Q1 in $\frac{N+1}{4} = \frac{52+1}{4} = 13.25^{th}$ item

Q_1 $= 13^{th}$ value + 0.25 (14^{th} Value – 13^{th} value)

$= 13^{th}$ value + 0.25 (400 – 200)

$= 200 + 0.25\ (400 – 200)$

$= 200 + 0.25\ (200)$

$= 200 + 50 = 250$

Position of Q_3 is 3 $\left(\frac{N+1}{4}\right) = 3\ 13.75^{th}$ item

Q_3 $= 39^{th}$ value $+ 0.75\ (40^{th}$ value $– 39^{th}$ value)

$= 500 + 0.75\ (500 – 500)$

$= 500 + 0.75 \times 0$

$= 500$

$$\text{Q.D.} = \frac{Q_3 -}{Q_1 2} = \frac{500 - 250}{2} = \frac{250}{\quad} = 125$$

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

$$= \frac{500 - 250}{500 + 250}$$

$$= \frac{250}{750} = 0.3333$$

Example:

For the date given below, give the quartile deviation andcoefficient 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	47	800.5- 950.5	372
951- 1100	28	950.5- 1100.5	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

$l_1 = 500.5$, $m_1 = 48$, $f_1 = 189$, $c_1 = 150$

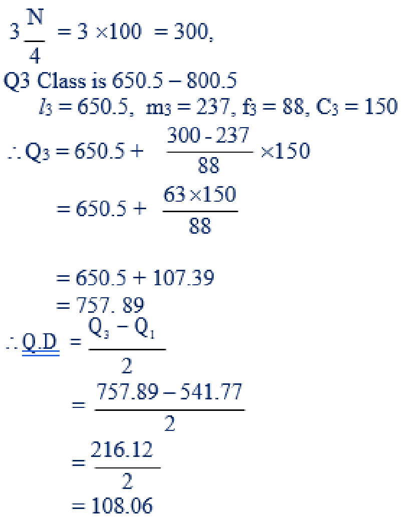
$$\therefore\ Q_1 = 500.5 + \frac{100 - 48}{189} \times 150$$

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

$= 500.5 + 41.27$

$= 541.77$

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



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

Q3 Class is 650.5 – 800.5

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

$$\therefore Q_3 = 650.5 + \frac{300 - 237}{88} \times 150$$

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

$= 650.5 + 107.39$

$= 757.\ 89$

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

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

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

$= 108.06$

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

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

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

$$= 0.1663$$

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.

2.3. Mean Deviation and Coefficient of Mean Deviation:

Mean Deviation:

The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation is measure of dispersion based on all items in a distribution.

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. Some times 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.

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

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

deviation $= \dfrac{\text{Mean deviation}}{\text{Mean or Median or Mode}} \text{X } 100$

2.3.1.Computation of mean deviation – Individual Series :

1. Calculate the average mean, median or mode of theseries.
2. Take the deviations of items from average ignoringsigns and denote these deviations by |D|.
3. Compute the total of these deviations, i.e., □|D|
4. Divide this total obtained by the number of items.

Symbolically: M.D.= $\dfrac{\sum|D|}{n}$

Example :

Calculate mean deviation from mean and median for thefollowing data:

100,150,200,250,360,490,500,600,671 also calculate co-efficients of M.D.

Solution:

Mean = $\bar{\mathbf{x}} = \dfrac{\sum \mathbf{x}}{\mathbf{n}} = \dfrac{3321}{9} = 369$

Now arrange the data in ascending order

100, 150, 200, 250, 360, 490, 500, 600, 671

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

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

= Value of 5th item

= 360

X	$\lvert D \rvert = \lvert x - \bar{x} \rvert$	$\lvert D \rvert = \lvert x - Md \rvert$
100	269	260
150	219	210
200	169	160
250	119	110
360	9	0
490	121	130
500	131	140
600	231	240
671	302	311
3321	1570	1561

M.D from mean $= \frac{\sum \lvert D \rvert}{n}$

$= \frac{1570}{9} = 174.44$

Co-efficient of M.D = $\frac{M.D}{\bar{x}}$

$= \frac{174.44}{369} = 0.47$

M.D from median $= \frac{\sum \lvert D \rvert}{n}$

$= \frac{1561}{9} = 173.44$

Co-efficient of M.D.= $\frac{M.D}{Median} = \frac{173.44}{360} = 0.48$

2.3.2. Mean Deviation – Discrete series:

Steps: 1. Find out an average (mean, median or mode)

2. Find out the deviation of the variable values from theaverage, ignoring signs and denote them by |D|

3. Multiply the deviation of each value by its respectivefrequency and find out the total

$\sum f \lvert D \rvert$

Divide this total obtained by the number of items. Symbolically:

M.D. = ∑|D| / 2

Example:

Compute Mean deviation from mean and median from the following data

Height in cms	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 person sf	d= x- A A =162	fd	|D| = |X- mean|	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

$$\bar{x} = A + \frac{\sum fd}{N}$$

$$= 162 + \frac{-95}{195} = 162 - 0.49 = 161.51$$

$$\text{M.D.} = \frac{\sum f\,|D|}{N} = \frac{338.59}{195} = 1.74$$

Coefficient of M.D.= $\frac{\text{M.D}}{\overline{X}} = \frac{1.74}{161.51} = 0.0108$

Height x	No. of persons f	c.f.	D = $\lvert X - \text{Median} \rvert$	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

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

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

$$= \text{Size of } 98^{th} \text{ item}$$

$$= 161$$

$$\text{M.D} = \frac{\sum f\,|D|}{N} = \frac{334}{195} = 1.71$$

Coefficient of M.D. = $\frac{\text{M.D}}{\text{Median}} = \frac{1.71}{161} = .0106$

2.3.3 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

$$\text{M.D} = \frac{\sum f\,|D|}{N}$$

Where D = m - average
M = Mid point

Example:

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

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

Also compute co-efficient of mean deviation.

Solution

X	m	f	$d = \frac{m - A}{c}$ (A=35,C=10)	fd	$\lvert D\rvert = \lvert m - \bar{x}\rvert$	$\lvert f\ D\rvert$
0-10	5	20	-3	-60	31.5	630.0
10-20	15	25	-2	-50	21.5	537.5
20-30	25	32	-1	-32	11.5	368.0
30-40	**35**	40	0	0	1.5	60.0
40-50	45	42	1	42	8.5	357.0
50-60	55	35	2	70	18.5	647.5
60-70	65	10	3	30	28.5	285.0
70-80	75	8	4	32	38.5	308.0
		212		32		3193.0

$$\bar{x} = A + \frac{\sum fd}{} \times cN$$

$$= 35 + \frac{32}{212} \times 10 \quad = 35 + \frac{320}{212} = 35 + 1.5 = 36.5$$

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

Calculation of median and M.D. from median

X	m	f	c.f	|D| = |m-Md|	f |D|
0-10	5	20	20	32.25	645.00
10-20	15	25	45	22.25	556.25
20-30	25	32	77	12.25	392.00
30-40	35	40	117	2.25	90.00
40-50	45	42	159	7.75	325.50
50-60	55	35	194	17.75	621.25
60-70	65	10	204	27.75	277.50
70-80	75	8	212	37.75	302.00
				Total	3209.50

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

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

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

$$= 30 + \frac{106 - 77}{40} \times 10$$

$$= 30 + \frac{29}{4}$$

$$= 30 + 7.25 = 37.25$$

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

$$= \frac{3209.5}{212} = 15.14$$

$$\text{Coefficient of M.D} = \frac{\text{M.D}}{\text{Median}}$$

$$= \frac{15.14}{37.25} = 0.41$$

Merits and Demerits of M.D :

Merits:

1. It is simple to understand and easy to compute.
2. It is rigidly defined.
3. It is based on all items of the series.
4. It is not much affected by the fluctuations of sampling.
5. It is less affected by the extreme items.
6. It is flexible, because it can be calculated from anyaverage.
7. It is better measure of comparison.

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. It is mathematically unsound and illogical.

2.4.Standard Deviation and Coefficient of variation:

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.

2.4.1.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 □(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.

Steps:

1. Find out the actual mean of the series ($\overline{x}$)
2. Find out the deviation of each value from the mean

   $(x = X - \overline{X})$

3.Square the deviations and take the total of squared deviations $\sum x^2$

4. Divide the total ($\sum x^2$) by the number of observation $\left(\frac{\sum x^2}{n}\right)$

The square root of $\left(\frac{\sum x^2}{n}\right)$ is standard deviation.

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

b) Deviations taken from assumed mean:

This method is adopted when the arithmetic mean isfractional 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. Theformula is:

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

Where d-stands for the deviation from assumed mean = (X-A)

Steps:

1. Assume any one of the item in the series as an average (A)
2. Find out the deviations from the assumed mean; i.e., X-Adenoted by d and also the total of the deviations □d
3. Square the deviations; i.e., d^2 and add up the squares ofdeviations, i.e, $\sum d^2$
4. Then substitute the values in the following formula:

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

Note: We can also use the simplified formula for standarddeviation.

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

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.14, 22, 9, 15, 20, 17, 12, 11

Solution:

Deviations from actual mean.

Values (X)	$X - \overline{X}$	$(X - \overline{X})^2$
14	-1	1
22	7	49
9	-6	36
15	0	0
20	5	25
17	2	4
12	-3	9
11	-4	16
120		140

$$\overline{X} = \frac{120}{\quad} = 15$$

8

$$\sigma = \sqrt{\frac{\Sigma(x - \overline{x})^2}{n}}$$

$$= \sqrt{\frac{140}{8}}$$

$$= \sqrt{17.5} = 4.18$$

Example 10:
The table below gives the marks obtained by 10 students instatistics. Calculate standard deviation.

Student Nos	:	1	2	3	4	5	6	7	8	9	10
Marks	:	43	48	65	57	31	60	37	48	78	59

Solution: (Deviations from assumed mean)

Nos.	Marks (x)	d=X-A (A=57)	d^2
1	43	-14	196
2	48	-9	81
3	65	8	64
4	57	0	0
5	31	-26	676
6	60	3	9
7	37	-20	400
8	48	-9	81
9	78	21	441
10	59	2	4
n = 10		$\sum d=-44$	$\sum d^2 =1952$

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

$$= \sqrt{\frac{1952}{10} - \left(\frac{-44}{10}\right)^2}$$

$$= \sqrt{195.2 - 19.36}$$

$$\sqrt{175.84}$$

$$= \qquad = 13.26$$

2.4.2.Calculation of standard deviation:

Discrete Series:

There are three methods for calculating standard deviationin discrete series:

(a) Actual mean methods

(b) Assumed mean method

(c) Step-deviation method.

(a) Actual mean method:

Steps:

1. Calculate the mean of the series.
2. Find deviations for various items from the means i.e.,

   $x - \bar{x} = d$.
3. Square the deviations ($= d^2$) and multiply by the respective frequencies(f) we get fd^2
4. Total to product ($\sum fd^2$) Then apply the formula:

$$\sigma = \sqrt{\frac{\sum fd^2}{\sum f}}$$

If the actual mean in fractions, the calculation takes lot of time and labour; and as such this method is rarely used in practice.

(b) Assumed mean method:

Here deviation are taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.

Steps:

1. Assume any one of the items in the series as an assumed mean and denoted by A.
2. Find out the deviations from assumed mean, i.e, X-A and denote it by d.
3. Multiply these deviations by the respective frequencies and get the ∑fd
4. Square the deviations (d^2).
5. Multiply the squared deviations (d^2) by the respective frequencies (f) and get ∑fd2.
6. Substitute the values in the following formula:

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

Where $d = X - A$, $N = \sum f$.

Example:

Calculate Standard deviation from the following data.

X :	20	22	25	31	35	40	42	45
f :	5	12	15	20	25	14	10	6

Solution:

Deviations from assumed mean

x	f	d = x –A (A = 31)	d^2	fd	fd^2
20	5	-11	121	-55	605
22	12	-9	81	-108	972
25	15	-6	36	-90	540
31	20	0	0	0	0
35	25	4	16	100	400
40	14	9	81	126	1134
42	10	11	121	110	1210
45	6	14	196	84	1176
	N=107			∑fd=167	$\sum fd^2$ =6037

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

$$= \sqrt{\frac{6037}{107} - \left(\frac{167}{107}\right)^2}$$

$$= \sqrt{56.42 - 2.44}$$

$$= \sqrt{53.98} = 7.35$$

(c) Step-deviation method:

If the variable values are in equal intervals, then we adoptthis method.

Steps:

1. Assume the center value of the series as assumed mean A
2. Find out d $\frac{x-A}{C}$, where C is the interval between each value
3. Multiply these deviations d’ by the respective frequencies and get ∑fd
4. Square the deviations and get d^2
5. Multiply the squared deviation (d^2) by the respective frequencies (f) and obtain the total ∑fd
6. Substitute the values in the following formula to get thestandard deviation

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

Example :

Compute Standard deviation from the following data

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

Solution:

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

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

$$= \sqrt{\frac{109}{60} - \left(\frac{5}{60}\right)^2} \times 10$$

$$= \sqrt{1.817 - 0.0069} \times 10$$

$$= \sqrt{1.8101} \times 10$$

$$= 1.345 \times 10$$

$$= 13.45$$

2.4.3.Calculation of Standard Deviation –Continuous series:

In the continuous series the method of calculating standard deviation is almost the same as in a discrete series. But in a continuous series, mid-values of the class intervals are to be found out. The step- deviation method is widely used.

The formula is

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

$$d = \frac{m - A}{C}$$, C- Class interval

Steps:

1. Find out the mid-value of each class.

2. Assume the center value as an assumed mean and denoteit by A

3. Find out d = $\frac{m - A}{C}$

4. Multiply the deviations d by the respective frequencies and get Σfd

5. Square the deviations and get d^2

6. Multiply the squared deviations (d^2) by the respective frequencies and get $\sum fd^2$

7. Substituting the values in the following formula to get the standard deviation

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

Example :

The daily temperature recorded in a city in Russia in a yearis given below.

Temperature C ^{0}	No. of days
-40 to –30	10
-30 to –20	18
-20 to –10	30
-10 to 0	42
0 to 10	65
10 to 20	180
20 to 30	20
	365

Calculate Standard Deviation.

Solution:

Temperature	Mid value (m)	No. of days f	d = $\frac{m-(-5^n)}{10^n}$	fd	fd ^{2}
-40 to -30	-35	10	-3	-30	90
-30 to -20	-25	18	-2	-36	72
-20 to -10	-15	30	-1	-30	30
-10 to - 0	-5	42	0	0	0
0 to 10	5	65	1	65	65
10 to 20	15	180	2	360	720
20 to 30	25	20	3	60	180
		N=365		Σfd = 389	Σfd^2 =1157

$$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times C$$

$$\sqrt{\frac{1157}{365} - \frac{389}{365}}$$

$$\sqrt{3.1699 = - \ 1.1358} \qquad \times 10$$

$$\sqrt{} = 2.0341 \times 10$$

$$= 1.4262 \times 10$$

$$= 14.26°c$$

2.4.4.Combined Standard Deviation:

If a series of N_1 items has mean $\bar{x}_1$ and standard deviation σ_1 and another series of N_2 items has mean X_2 and standard deviation σ_2 , we can find out the combined mean and

$$\overline{X}_{12} = \frac{N1\,\overline{X}_1 + N_2\,\overline{X}_2}{N_1 + N_2}$$

combinedstandard deviation by using the formula.

mean and combinedstandard deviation by using the formula.

$$\sigma_{12} = \sqrt{\frac{N_1\,\sigma_1^2 + N_2\,\sigma_2^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

Where $d_1 = \overline{X}_1 - \overline{X}_{12}$

$d_2 = \overline{X}_2 - \overline{X}_{12}$

Example :

Particulars regarding income of two villages are given below.

	Village	
	A	**B**
No.of people	600	500

Average income	175	186
Standard deviation of income	10	9

Compute combined mean and combined Standard deviation.

Solution:

Given $N_1 = 600$, $\overline{X}_1 = 175$, $\sigma_1 = 10 N_2 = 500$, $\overline{X}_2 = 186$,

$\overline{\sigma_2} = 9$

Combined mean

$$\overline{X}_{12} = \frac{N1\,\overline{X}_1 + N_2\,\overline{X}_2}{N_1 + N_2}$$

$$= \frac{600 \times 175 + 500 \times 186}{600 + 500}$$

$$= \frac{105000 + 93000}{1100}$$

$$= \frac{198000}{1100} = 180$$

Combined Standard Deviation:

$$\sigma_{12} = \sqrt{\frac{N_1\,\sigma_1{}^2 + N_2\,\sigma_2{}^2 + N_1 d_1^2 + N_2 d_2^2}{N_1 + N_2}}$$

$d_1 = \overline{X}_1 - \overline{X}_{12}$

$= 175 - 180$

$= -5$

$d_2 = \overline{X}_2 - X_{12}$

$= 186 - 180$

$= 6$

$$\sigma_{12} = \sqrt{\frac{600 \times 100 + 500 \times 81 + 600 \times 25 + 500 \times 36}{600 + 500}}$$

$$= \sqrt{\frac{60000 + 40500 + 15000 + 18000}{1100}}$$

$$= \sqrt{\frac{133500}{1100}}$$

$$= \sqrt{121.364}$$

$$= 11.02.$$

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 ofarithmetic mean.
3. It is the most important and widely used measure ofdispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling and hencestable.
6. It is the basis for measuring the coefficient of correlationand sampling

Demerits:

1. It is not easy to understand and it is difficult to calculate.
2. It gives more weight to extreme values because the valuesare squared up.
3. As it is an absolute measure of variability, it cannot be usedfor the purpose of comparison.

2.5. Coefficient of Variation :

The Standard deviation is an absolute measure of dispersion. It is expressed in terms of units in which the original figures are collected and stated. The standard deviation of heights of students cannot be compared with the standard deviation of weights of students, as both are expressed in different units, i.e heights in centimeter and weights in kilograms. Therefore the standard deviation must be converted into a relative measure of dispersion for the purpose of comparison. The relative measure is known as the coefficient of variation.

The coefficient of variation is obtained by dividing the standard deviation by the mean and multiply it by 100. symbolically,

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

If we want to compare the variability of two or more series, we can use C.V. The series or groups of data for which the C.V. is greater indicate that the group is more variable, less stable, less uniform, less consistent or less homogeneous. If the C.V. is less, it indicates that the group is less variable, more stable, more uniform, more consistent or more homogeneous.

Example:

In two factories A and B located in the same industrial area, the average weekly wages (in rupees) and the standard deviations are as follows:

Factory	Average	Standard Deviation	No. of workers
A	34.5	5	476
B	28.5	4.5	524

Which factory A or B pays out a larger amount as weeklywages?
Which factory A or B has greater variability in individualwages?

Solution:

Given $N_1 = 476, \overline{X}_1 = 34.5, \sigma_1 = 5$

$N_2 = 524, \overline{X}_2 = 28.5, \sigma_2 = 4.5$

Total wages paid by factory A

$= 34.5 \times 476$

= Rs.16.422

Total wages paid by factory B

$= 28.5 \times 524$

= Rs.14,934.

Therefore factory A pays out larger amount as weekly wages.

C.V. of distribution of weekly wages of factory A and B are

$$\text{C.V.(A)} = \frac{\sigma_1}{\overline{\overline{X}}_1} \times 100$$

$$= \frac{5}{34.5} \times 100$$

$$= 14.49$$

$$\text{C.V (B)} = \frac{\sigma_2}{\overline{\overline{X}}_2} \times 100$$

$$= \frac{4.5}{28.5} \times 100$$

$$=15.79$$

Factory B has greater variability in individual wages, since C.V. of factory B is greater than C.V of factory A

Example:

Prices of a particular commodity in five years in two cities aregiven 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 X=20 dx	dx^2	Prices (Y)	Deviations from Y =15 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
$\Sigma x=100$	$\Sigma dx=0$	$\Sigma dx^2=30$	$\Sigma y=75$	$\Sigma dy=0$	Σdy^2 =68

City A: $\overline{X}=\frac{\Sigma x}{n} = \frac{100}{5} = 20$

$$\sigma_x = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n}} = \sqrt{\frac{\Sigma dx^2}{n}}$$

$$= \sqrt{\frac{30}{5}} = \sqrt{6} = 2.45$$

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

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

$= 12.25$ %

City B: $\overline{Y}=\frac{\Sigma y}{n} = \frac{75}{5} = 15$

$$\sigma_y = \sqrt{\frac{\Sigma(y-\overline{y})^2}{n}} = \sqrt{\frac{\Sigma dy^2}{n}}$$

$$= \sqrt{\frac{68}{5}} = \sqrt{13.6} = 3.69$$

$$C.V.(y) = \frac{\sigma y}{\bar{y}} \text{ x } 100$$

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

$$= 24.6\ \%$$

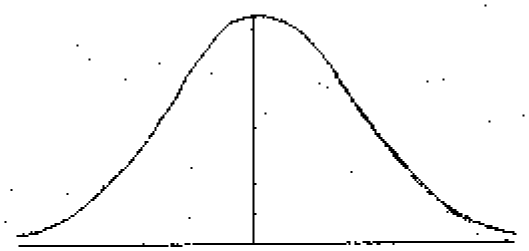
City A had more stable prices than City B, because the coefficient of variation is less in City A.

2.6.Skewness:

Meaning:

Skewness means ‘ lack of symmetry’ . We study skewness tohave an idea about the shape of the curve which we can draw with the help of the given data.If in a distribution mean = median = mode, then that distribution is known as symmetrical distribution.If in a distribution mean $\neq$ median $\neq$ mode , then it is not a symmetrical distribution and it is called a skewed distribution and such a distribution could either be positively skewed or negatively skewed

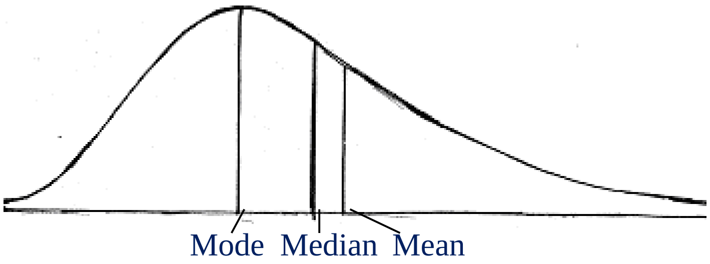
a) Symmetrical distribution:



Mean = Median = Mode

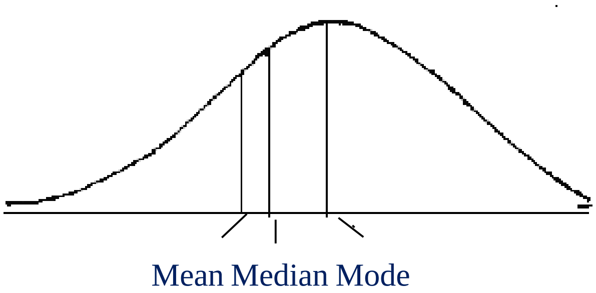
It is clear from the above diagram that in a symmetrical distribution the values of mean, median and mode coincide. The spread of the frequencies is the same on both sides of the center point of the curve.

Positively skewed distribution:



It is clear from the above diagram, in a positively skewed distribution, the value of the mean is maximum and that of the mode is least, the median lies in between the two. In the positively skewed distribution the frequencies are spread out over a greater range of values on the right hand side than they are on the left hand side.

Negatively skewed distribution



It is clear from the above diagram, in a negatively skewed distribution, the value of the mode is maximum and that of the mean is least. The median lies in between the two. In the negatively skewed distribution the frequencies are spread out over a greater range of values on the left hand side than they are on the right hand side.

Measures of skewness:

The important measures of skewness are

(i) Karl– Pearason's coefficient of skewness

(ii) Bowley's coefficient of skewness (iii)Measure of skewness based on moments

2.7. 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:

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

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

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

Example :

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

25, 15, 23, 40, 27, 25, 23, 25, 20

Solution:
Computation of Mean and Standard deviation :

Size	Deviation from A=25 D	d^2
25	0	0
15	-10	100
23	- 2	4
40	15	225
27	2	4
25	0	0
23	-2	4
25	0	0
20	- 5	25
N=9	∑d=-2	$\sum d^2=362$

Mean $= A + \frac{\sum d}{n}$

$$= 25 + \frac{-2}{9}$$

$$= 25 - 0.22 = 24.78$$

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

$$= \sqrt{\frac{362}{9} - \left(\frac{-2}{9}\right)^2}$$

$$= \sqrt{40.22 - 0.05}$$

$$= \sqrt{40.17} = 6.3$$

Mode = 25, as this size of item repeats 3 timesKarl – Pearson' s coefficient of skewness

$$= \frac{\text{Mean - Mode}}{\text{S.D}}$$

$$= \frac{24.78 - 25}{6.3}$$

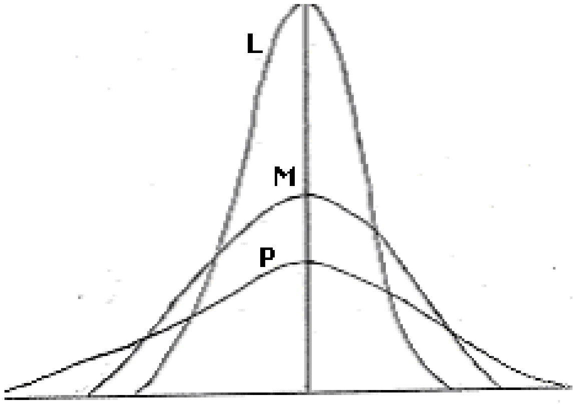
$$= \frac{-0.22}{6.3}$$

$$= -0.03$$

2.8.Kurtosis:

The expression ‘Kurtosis’ is used to describe the peakedness of a curve. The three measures– central tendency, dispersion and skewness describe the characteristics of frequency distributions. But these studies will not give us a clear picture of the characteristics of a distribution.

As far as the measurement of shape is concerned, we have two characteristics– skewness which refers to asymmetryof a series and kurtosis which measures the peakedness of a normal curve. All the frequency curves expose different degrees of flatness or peakedness. This characteristic of frequency curve is termed as kurtosis. Measure of kurtosis denote the shape of top of a frequency curve. Measure of kurtosis tell us the extent to which a distribution is more peaked or more flat topped than the normal curve, which is symmetrical and bell-shaped, is designated as Mesokurtic. If a curve is relatively more narrow and peaked at the top, it is designated as Leptokurtic. If the frequency curve is more flat than normal curve, it is designated as platykurtic



$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

L = Lepto Kurtic M = Meso Kurtic P = Platy Kurtic

Measure of Kurtosis:

The measure of kurtosis of a frequency distribution basedmoments is denoted by β_2 and is given by

If $\beta_2 = 3$, the distribution is said to be normal and the curveis 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	20	15	10	5

Solution:

[**Hint:** Refer Example of page 172 and get the values of first four central moments and then proceed to find β_1 and β_2]

$$\mu_1 = 0 \qquad \mu_2 = \frac{\sum fd^2}{N} = \frac{500}{125} = 4$$

$$\mu_3 = \frac{\sum fd^3}{N} = 0 \qquad \mu_4 = \frac{\sum fd^4}{N} = \frac{4700}{125} = 37.6$$

$$\therefore \beta_1 = \frac{\mu^2_3}{\mu^3_2} = \frac{0}{64} = 0$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{37.6}{4^2}$$

$$= \frac{37.6}{16} = 2.35$$

The value of β_2 is less than 3, hence the curve is platykurtic.

Example :

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

X :	30-33	33-36	36-39	39-42	42-45	45-48
f :	2	4	26	47	15	6

$\mu_1 = 0,$ $\mu_2 = 8.76$ $\mu_3 = -2.91,$ $\mu_4 = 291.454$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

$$\beta_1 = \frac{(-2.91)^2}{(8.76)^3} = \frac{8.47}{672.24} = 0.0126$$

$$\beta_2 = \frac{\mu_4}{\mu_2{}^2}$$

$$\beta_2 = \frac{291.454}{(8.76)^2} = 3.70$$

Since β_2 >3, the curve is leptokurtic.

Unit II

Self – Assessment

1.Compute quartile deviation from the following data

Height in inches	58	59	60	61	62	63	64	65	66
No. of students	15	20	32	35	33	22	20	10	8

2. Calculate mean deviation from mean from the following data

X	2	4	6	8	10
Y	1	4	6	4	1

3. Calculate the S.D of the following

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

4.From the following table calculate the Karl – Pearson’ s coefficient of skewness

Daily Wages (in RS)	150	200	250	300	350	400	450
No. of People	3	25	19	16	4	5	6

Suggested Activity

1.Select any two groups of any size from your class calculate mean, S.D and C.V for statistics marks. Find which group is more consistent.

Ans:

1. Q.D = 1.5
2. M.D = 1.5
3. S.D = 12.3
4. Sk = 0.88

TIME SERIES

OBJECTIVE:

The objectives of time series analysis are to gain insights into time-dependent data, make accurate forecasts, identify and analyze trends and patterns, and support data-driven decision-making.

3.1. INTRODUCTION:

Arrangement of statistical data in chronological order ie., in accordance with occurrence of time, is known as “Time Series”. Such series have a unique important place in the field of Economic and Business statistics. An economist is interested in estimating the likely population in the coming year so that proper planning can be carried out with regard to food supply, job for the people etc. Similarly, a business man is interested in finding out his likely salesin the near future, so that the businessman could adjust his production accordingly and avoid the possibility of inadequate production to meet the demand. In this connection one usually deal with statistical data, which are collected, observed or recorded at successive intervals of time. Such data are generally referred to as‘ time series’

3.2.Definition:

According to Mooris Hamburg “A time series is a set of statistical observations arranged in chronological order”

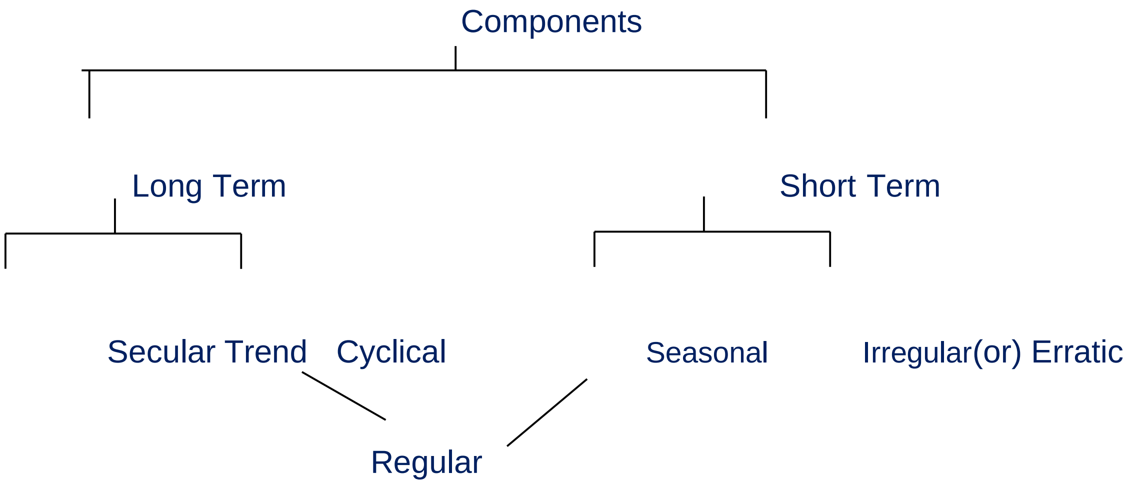
Ya-Lun- chou defining the time series as “A time series may be defined as a collection of readings belonging to different time periods, of some economic variable or composite of variables. A time series is a set of observations of a variable usually at equal intervals of time. Here time may be yearly, monthly, weekly, daily or even hourly usually at equal intervals of time.

Hourly temperature reading, daily sales, monthly production are examples of time series. Number of factors affect the observations of time series continuously, some with equal intervals of time and others are erratic studying, interpreting analyzing the factors is called Analysis of Time Series.

The Primary purpose of the analysis of time series is to discover and measure all types of variations which characterise a time series. The central objective is to decompose the various elements present in a time series and to use them in business decision making.

3.3. Components of Time series:

The components of a time series are the various elements which can be segregated from the observed data. The following are the broad classification of these components.



In time series analysis, it is assumed that there is a multiplicative relationship between these four components. Symbolically,

$$Y = T \times S \times C \times I$$

Where Y denotes the result of the four elements; T = Trend ;S = Seasonal component; C = Cyclical components; I = Irregular component

In the multiplicative model it is assumed that the four components are due to different causes but they are not necessarily independent and they can affect one another.

Another approach is to treat each observation of a time series as the sum of these four components. Symbolically

$$Y = T + S + C + I$$

The additive model assumes that all the components of thetime series are independent of one another.

1) Secular Trend or Long - Term movement or simply Trend
2) Seasonal Variation
3) Cyclical Variations
4) Irregular or erratic or random movements(fluctuations)

3.4.Secular Trend:

It is a long term movement in Time series. The general tendency of the time series is to increase or decrease or stagnate during a long period of time is called the secular trend or simply trend. Population growth, improved technological progress, changes in consumers taste are the various factors of upward trend. We may notice downward trend relating to deaths, epidemics, due to improved medical facilities and sanitations. Thus a time series shows fluctuations in the upward or downward direction in the long run.

Methods of Measuring Trend:

Trend is measured by the following mathematical methods.

1. Graphical method
2. Method of Semi-averages
3. Method of moving averages
4. Method of Least Squares

1.Graphical Method:

This is the easiest and simplest method of measuring trend. In this method, given data must be plotted on the graph, taking time on the horizontal axis and values on the vertical axis. Draw a smooth curve which will show the direction of the trend. While fitting a trend line the following important points should be noted to get a perfect trend line.

(i) The curve should be smooth.

(ii) As far as possible there must be equal number of points above and below the trend line.

(iii) The sum of the squares of the vertical deviations from the trend should be as small as possible.

(iv) If there are cycles, equal number of cycles should be above or below the trend line.

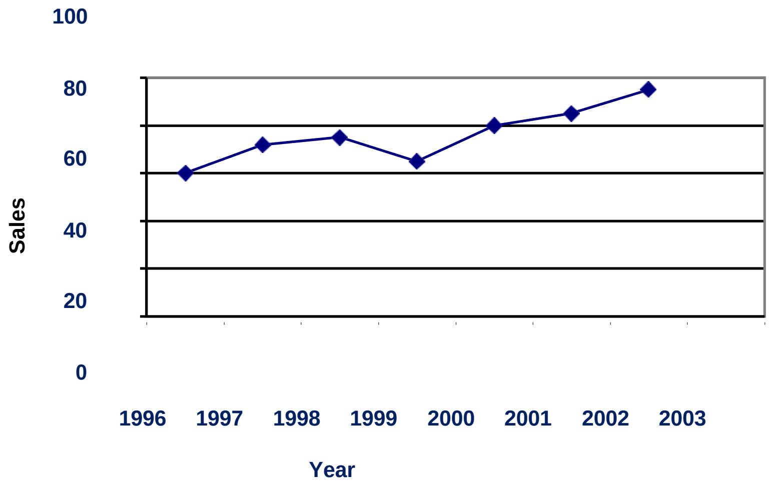
(v) In case of cyclical data, the area of the cycles above and below should be nearly equal.

Example:

Fit a trend line to the following data by graphical method.

Year	1996	1997	1998	1999	2000	2001	2002
Sales (in Rs ‘ 000)	60	72	75	65	80	85	95

Solution:



The dotted lines refers trend lines

Merits:

1. It is the simplest and easiest method. It saves time and labour.
2. It can be used to describe all kinds of trends.
3. This can be used widely in application.
4. It helps to understand the character of time series and to select appropriate trend.

Demerits:

1. It is highly subjective. Different trend curves will be obtained by different persons for the same set of data.
2. It is dangerous to use freehand trend for forecasting purposes.
3. It does not enable us to measure trend in precise quantitative terms.

2.Method of semi averages:

In this method, the given data is divided into two parts, preferably with the same number of years. For example, if we are given data from 1981 to 1998 i.e., over a period of 18 years, the two equal parts will be first nine years, i.e., 1981 to 1989 and from 1990 to 1998. In case of odd number of years like 5,7,9,11 etc, two equal parts can be made simply by omitting the middle year. For example, if the data are given for 7 years from 1991 to 1997, the two equal parts would be from 1991 to 1993 and from 1995 to 1997, the middle year 1994 will be omitted.After the data have been divided into two parts, an average of each parts is obtained. Thus we get two points. Each point is plotted at the mid-point of the class interval covered by respective part and then the two points are joined by a straight line which gives us the required trend line. The line can be extended downwards and upwards to get intermediate values or to predict future values.

Example:

Draw a trend line by the method of semi-averages.

Year	1991	1992	1993	1994	1995	1996
Sales Rs in (1000)	60	75	81	110	106	117

Solution:

Divide the two parts by taking 3 values in each part.

Year	Sales (Rs)	Semi total	Semi average	Trend values
1991	60			59
1992	75	216	72	72
1993	81			85
1994	110			98
1995	106	333	111	111
1996	117			124

Difference in middle periods = 1995 –1992 = 3 years Difference in semi averages = 111 –72 = 39

Annual increase in trend = 39/3 = 13

Trend of 1991 = Trend of 1992 -13

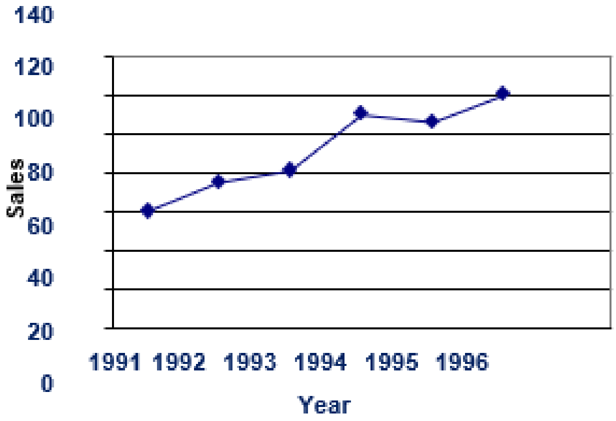
= 72-13 = 59

Trend of 1993 = Trend of 1992 +13

= 72 + 13 = 85

Similarly, we can find all the values

The



following graph will show clearly the trend line

Example:

Calculate the trend value to the following data by the method of semi- averages.

Year	1995	1996	1997	1998	1999	2000	2001
Expenditure(Rs in Lakhs)	1.5	1.8	2.0	2.3	2.4	2.6	3.0

Solution

Year	Expenditure(Rs)	Semi total	Semi average	Trend values
1995	1.5			1.545
1996	1.8	5.3	1.77	1.770
1997	2.0			1.995
1998	2.3			2.220
1999	2.4			2.445
2000	2.6	8.0	2.67	2.670
2001	3.0			2.895

Difference between middle periods = 2000 – 1996

= 4 years

Difference between semi-averages = 2.67 - 1.77

= 0.9

∴ Annual trend values ___ = 0.9 4

= 0.225

Trend of 1995 = Trend of 1996 – 0.225

= 1.77 – 0.225

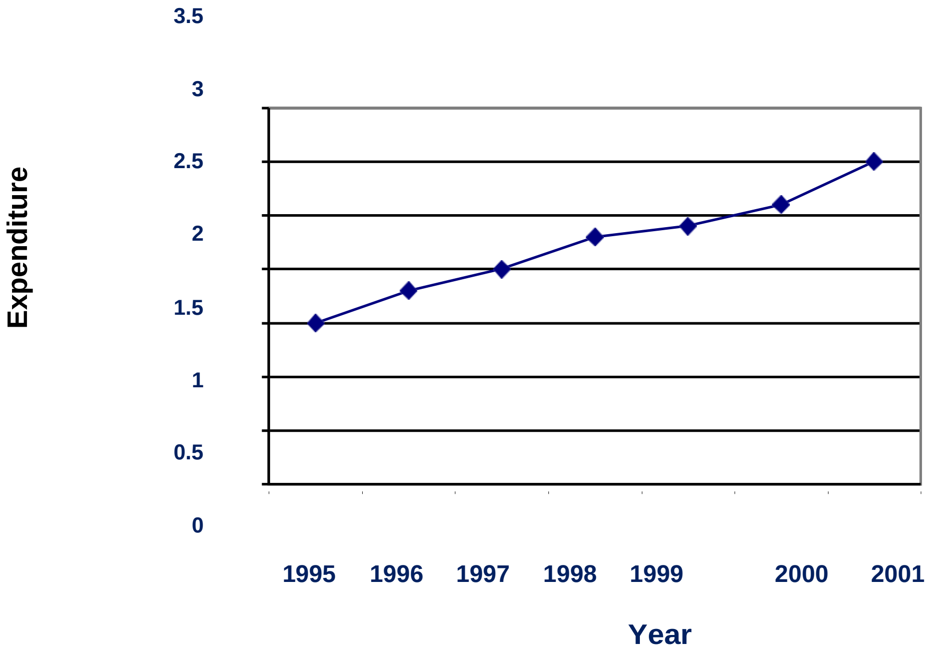
= 1.545

Trend of 1996 = 1.77

Trend of 1997 = 1.77 + 0.225

= 1.995

Similarly we can find all the trend values



Merits:

1. It is simple and easy to calculate
2. By this method every one getting same trend line.
3. Since the line can be extended in both ways, we can find thelater and earlier estimates.

Demerits:

1. This method assumes the presence of linear trend to the valuesof time series which may not exist.
2. The trend values and the predicted values obtained by thismethod are not very reliable.

3.Method of Moving Averages:

This method is very simple. It is based on Arithmetic mean.Theses means are calculated from overlapping groups of successive time series data. Each moving average is based on values coveringa fixed time interval, called “period of moving average” and is shown against the center of the interval.

The method of ‘ odd period of moving average is as follows.
(3 or 5) . The moving averages for three years is

$$\frac{a+b+c}{3}$$

$$\frac{b+c+d}{3} \quad \frac{c+d+e}{3} \quad \text{etc}$$

The formula for five yearly moving average

Is

$$\frac{a+b+c+d+e}{5},$$

$$\frac{b+c+d+e+f}{5} \qquad \frac{c+d+e+f+g}{5} \qquad \text{etc}$$

Steps for calculating odd number of years.

1. Find the value of three years total, place the value against the second year.
2. Leave the first value and add the next three years value (ie 2nd, 3rd and 4th years value) and put it against 3rd year.
3. Continue this process until the last year’ s value taken.
4. Each total is divided by three and placed in the next column.

These are the trend values by the method of moving averages

Example :

Calculate the three yearly average of the following data.

Year	1975	1976	1977	1978	1979	1980	1981	1982	1983	1984
Production in (tones)	50	36	43	45	39	38	33	42	41	34

Solution:

Year	Production (in tones)	3 years moving total	3 years moving average as Trend values
1975	50	-	-
1976	36	129	43.0
1977	43	124	41.3
1978	45	127	42.3
1979	39	122	40.7
1980	38	110	36.7
1981	33	113	37.7
1982	42	116	38.7
1983	41	117	39.0
1984	34	-	-

Even Period of Moving Averages:

When the moving period is even, the middle period of each set of values lies between the two time points. So we must center the moving averages.

The steps are

1. Find the total for first 4 years and place it against the middle ofthe 2nd and 3rd year in the third column.
2. Leave the first year value, and find the total of next four-year and place it between the 3rd and 4th year.
3. Continue this process until the last value is taken.
4. Next, compute the total of the first two four-year totals and place it against the 3rd year in the fourth column.
5. Leave the first four years total and find the total of the next two four years' totals and place it against the fourth year.
6. This process is continued till the last two four years' total is taken into account.
7. Divide this total by 8 (Since it is the total of 8 years) and put it in the fifth column.
8. These are the trend values.

Example :

The production of Tea in India is given as follows.

Calculate the Four-yearly moving averages

Year	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Production (tones)	464	515	518	467	502	540	557	571	586	612

Solution:

Year	Production (in tones)	4 years Moving total	Total of Two four years	Trend Values
1993	464		-	-
		-		
1994	515			
		1964		
1995	518		3966	495.8
		2002		
1996	467		4029	503.6
		2027		
1997	502		4093	511.6
		2066		
1998	540		4236	529.5
		2170		
1999	557		4424	553.0
		2254		
2000	571		4580	572.5
		2326		
2001	586			
		-		
2002	612			

Merits:

1. The method is simple to understand and easy to adopt as compared to other methods.
2. It is very flexible in the sense that the addition of a few more figures to the data, the entire calculations are not changed. We only get some more trend values.
3. Regular cyclical variations can be completely eliminated by a period of moving average equal to the period of cycles.
4. It is particularly effective if the trend of a series is very irregular.

Demerits:

5. It cannot be used for forecasting or predicting future trend, which is the main objective of trend analysis.
6. The choice of the period of moving average is sometimes subjective.
7. Moving averages are generally affected by extreme values of items.
8. It cannot eliminate irregular variations completely.

4.Method of Least Square:

This method is widely used. It plays an important role in finding the trend values of economic and business time series. It helps for forecasting and predicting the future values. The trend line by this method is called the line of best fit.

The equation of the trend line is y = a + bx, where the constants a and b are to be estimated so as to minimize the sum of the squares of the difference between the given values of y and the estimate values of y by using the equation. The constants can be obtained by solving two normal equations.

$$\Sigma y = na + b\Sigma x.............(1)$$

$$\Sigma xy = a\Sigma x \ + \ b\Sigma x^2 .... (2)$$

Here x represent time point and y are observed values. ‘ n’ is the number of pair-values.

When odd number of years are given

Step 1: Writing given years in column 1 and the corresponding sales or production etc in column Step 2: Write in column 3 start with 0, 1, 2 .. against column 1 and denote it as X

Step 3: Take the middle value of X as A

Step 4: Find the deviations u = X – A and write in column 4 Step 5: Find u2 values and write in column 5.

Step 6: Column 6 gives the product uy

Now the normal equations become

$\Sigma y = na + b\Sigma u$ (1) where u = X-A

$\Sigma uy = a\Sigma u + b\Sigma u^2$ (2)Since $\Sigma u = 0$,

From equation (1)

$$a = \frac{\Sigma y}{n}$$

From equation (2)

$$\Sigma uy = b\Sigma u^2$$

b =Σuy / Σu2

The fitted straight line is

$$y = a + bu \ = a + b\ (\ X - A)$$

Example:

For the following data, find the trend values by using the method of Least squares

Year	1990	1991	1992	1993	1994
Production (in tones)	50	55	45	52	54

Estimate the production for the year 1996

Solution:

Year(x)	Production(y)	X= x -1990	u = X-A = X-2	u^2	uy	Trend values
1990	50	0	-2	4	-100	50.2
1991	55	1	-1	1	-55	50.7
1992	45	**2 A**	0	0	0	51.2
1993	52	3	1	1	52	51.7
1994	54	4	2	4	108	52.2
Total	256			10	5	

Where A is an assumed value The equation of straight line is

Y = *a* + *bX*

= *a* + bu , where u = X - 2the normal equations are

$$\Sigma y = na + b\Sigma u\text{......}(1)$$

$$\Sigma uy = a\Sigma u + \text{b}\Sigma u^2 \cdots (2)$$

since $\Sigma u = 0$ from(1) $\Sigma y = na$

$$= 51.2$$

$$a = \frac{\Sigma \text{y}}{\text{n}} = \frac{256}{5}$$

From equation (2)

$$\Sigma uy = b\Sigma u^2$$

$$5 = 10\text{b}$$

$$\text{b} = \frac{5}{10} = 0.5$$

The fitted straight line is

y = a+ bu

y = 51.2 + 0.5 (X-2)

y = 51.2 + 0.5X –1.0

y = 50.2 + 0.5X

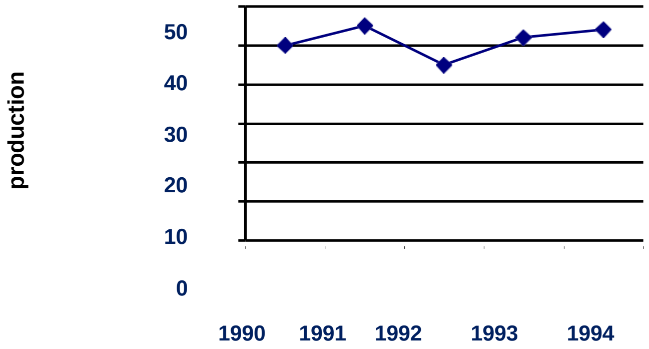
Trend values are, 50.2, 50.7, 51.2, 51.7, 52.2 The estimate production in 1996 is

put X = x – 1990

X = 1996 –1990 = 6

Y = 50.2 + 0.5X = 50.2 +0.5(6)

= 50.2 +3.0 = 53.2 tonnes.



When **even number of years** are given

Here we take the mean of middle two values of X as A Then

u = X – A 1 / 2 = 2 (X-A). The other steps are as given in the odd number of years.

Example 7:

Fit a straight line trend by the method of least squares for the following data.

Year	1983	1984	1985	1986	1987	1988
Sales (Rs. in lakhs)	3	8	7	9	11	14

Also estimate the sales for the year 1991

Solution:

Year (x)	Sales (y)	X = x-1983	u =2X-5	u2	*uy*	Trend values
1983	3	0	-5	25	-15	3.97
1984	8	1	-3	9	-24	5.85
1985	7	2	-1	1	-7	7.73
1986	9	3	1	1	9	9.61
1987	11	4	3	9	33	11.49
1988	14	5	5	25	70	13.37
Total	52		0	70	66	

$$u = \frac{\text{X} - \text{A}}{1/2}$$

$= 2\,(X - 2.5) = 2\text{X} - 5$

The straight line equation is

$y = a + \text{b}X\ = a + bu$

The normal equations are

$\Sigma y = na$ ....(1)

$\Sigma uy = \text{b}\Sigma u^2$ .....(2) From (1) $52 = 6a$

$$a = \frac{52}{6}$$

$= 8.67$

From (2) $66 = 70$ b

$$b = \frac{66}{70}$$

$= 0.94$

The fitted straight line equation is

$y\ =\ a{+}bu$

$y = 8.67{+}0.94(2\text{X}{-}5)$

$y = 8.67 + 1.88\text{X} - 4.7$

$y = 3.97 + 1.88X$ -----------(3)

The trend values are

Put $X = 0$, $y = 3.97$ $X = 1$, $y = 5.85$

$X = 2$, $y = 7.73$ $X = 3$, $y = 9.61$

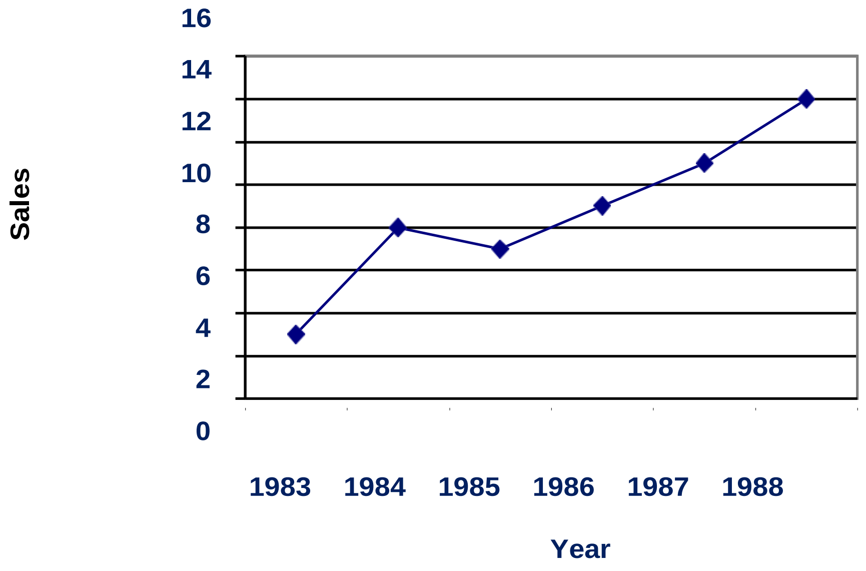
$X = 4$, $y = 11.49$ $X = 5$, $y = 13.37$

The estimated sale for the year 1991 is; put X = x –1983

= 1991 – 1983 = 8

$y = 3.97 + 1.88 \times 8 = 19.01$ lakhs

The following graph will show clearly the trend line.



Merits:

1. Since it is a mathematical method, it is not subjective so iteliminates personal bias of the investigator.
2. By this method we can estimate the future values. As well asintermediate values of the time series.
3. By this method we can find all the trend values.

Demerits:

1. It is a difficult method. Addition of new observations makes re-calculations.
2. Assumption of straight line may sometimes be misleading sinceeconomics and business time series are not linear.

3. It ignores cyclical, seasonal and irregular fluctuations.
4. The trend can estimate only for immediate future and not fordistant future.

3.5.Seasonal Variations:

Seasonal Variations are fluctuations within a year during theseason. The factors that cause seasonal variation are

i) Climate and weather condition.
ii) Customs and traditional habits.

For example the sale of ice-creams increase in summer, the umbrella sales increase in rainy season, sales of woolen clothes increase in winter season and agricultural production depends upon the monsoon etc.,

Secondly in marriage season the price of gold will increase, sale of crackers and new clothes increase in festival times.

So seasonal variations are of great importance to businessmen, producers and sellers for planning the future. The main objective of the measurement of seasonal variations is to study their effect and isolate them from the trend.

3.5.1.Measurement of seasonal variation:

The following are some of the methods more popularly used for measuring the seasonal variations.

1. Method of simple averages.
2. Ratio to trend method.
3. Ratio to moving average method.
4. Link relative method

Among the above four methods the method of simple averages is easy to compute seasonal variations.

1.Method of simple averages

i) The steps for calculations:
ii) Arrange the data season wise
iii) Compute the average for each season.
iv) Calculate the grand average, which is the average of seasonal averages
v) .Obtain the seasonal indices by expressing each season as percentage of Grand average
vi) The total of these indices would be 100n where ‘ n’ is the number of

seasons in the year.

Example:

Find the seasonal variations by simple average method for the data given below.

Quarter

Year	I	II	III	IV
198 9	30	40	36	34
199 0	34	52	50	44
199 1	40	58	54	48
199 2	54	76	68	62
199 3	80	92	86	82

Solution:

Quarter

Year	I	II	III	IV
1989	30	40	36	34
1990	34	52	50	44
1991	40	58	54	48
1992	54	76	68	62
1993	80	92	86	82
Total	238	318	294	270
Average	47. 6	63.6	58.8	54
Seasona l Indices	85	113. 6	105	96.4

Grand average = $\dfrac{47.6+63.6+58.8+54}{4}$

$= \dfrac{224}{4} = 56$

Seasonal Index for I quater

$= \dfrac{First\ quarterly\ Average}{Grand\ Average} \times 100$

$= \dfrac{47.6}{56} \times 100$

$= 85$

Seasonal Index for

II quarter $= \dfrac{Second\ quarterly\ Average}{Grand\ Average} \times 100$

$= \dfrac{63.6}{56} \times 100 = 113.6$

Seasonal Index for

III quarter $= \dfrac{Third\ quarterly\ Average}{Grand\ Average} \times 100$

$= \dfrac{58.8}{56} \times 100 = 105$

Seasonal Index for

IV quarter $= \dfrac{Fourth\ quarterly\ Average}{Grand\ Average} \times 100$

$= \dfrac{54}{56} \times 100 = 96.4$

Example:

Calculate the seasonal indices from the following data using simple average method.

Year

Quarter	1974	1975	1976	1977	1978
I	72	76	74	76	74
II	68	70	66	74	74
III	80	82	84	84	86
IV	70	74	80	78	82

Solution:

Quarter

Year	I	II	III	IV
1974	72	68	80	70
1975	76	70	82	74
1976	74	66	84	80
1977	76	74	84	78
1978	74	74	86	82
Total	372	352	416	384
Average	74.4	70.4	83.2	76.8
Seasonal Indices	97.6	92.4	109.2	100.8

$$\text{Grand Average} = \frac{74.4 + 70.4 + 83.2 + 76.8}{4}$$

$$= \frac{304.8}{4} = 76.2$$

Seasonal Index for

$$\text{I quarter} = \frac{\textit{First quarterly Average}}{\textit{Grand Average}} \times 100$$

$$= \frac{74.4}{} \times 100$$

76.2

= 97.6

Seasonal Index for

$$II \text{ quarter} = \frac{Second\ quarterly\ Average}{Grand\ Average} \times 100$$

$$= \frac{70.4}{76.2} \times 100$$

= 92.4

Seasonal Index for

$$III \text{ quarter} = \frac{Third\ quarterly\ Average}{Grand\ Average} \times 100$$

$$= \frac{83.2}{76.2} \times 100$$

= 109.2

Seasonal Index for

$$IV \text{ quarter} = \frac{Fourth\ quarterly\ Average}{Grand\ Average}\ x\ 100$$

$$= \frac{76.8}{76.2} \times 100$$

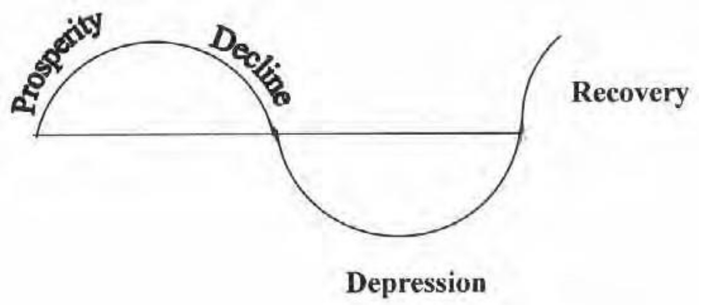
The total of seasonal indices calculated must be equal to 400 here we have = 97.6 + 92.4 + 109.2 + 100.8

= 400 hence verified.

3.6.Cyclical variations:

The term cycle refers to the recurrent variations in time series, that extend over longer period of time, usually two or more years. Most of the time series relating to economic and business show some kind of cyclic variation. A business cycle consists ofthe recurrence of the up and down movement of business activity. Itis a four-phase cycle namely.

1. Prosperity 2. Decline 3. Depression 4. Recovery Each phase changes gradually into the following phase. The following diagram illustrates a business cycle.



The study of cyclical variation is extremely useful in framing suitable policies for stabilising the level of business activities. Businessmen can take timely steps in maintaining business during booms and depression.

Irregular variation:

Irregular variations are also called erratic. These variations are not regular and which do not repeat in a definite pattern. These variations are caused by war, earthquakes, strikes flood, revolution etc. This variation is short-term one, but it affect all the components of series. There is no statistical techniques for measuring or isolating erratic fluctuation. Therefore the residual that remains after eliminating systematic components is taken as representing irregular variations.

3.7FORECASTING

very important use of time series data is towards forecasting the likely value of variable in future. In most cases it is the projection of trend fitted into the values regarding a variable over a sufficiently long period by any of the methods discussed latter. Adjustments for seasonal and cyclical character introduce further improvement in the forecasts based on the simple projection of the trend. The importance of forecasting in business and economic fields lies on account of its role in planning and evaluation. If suitably interpreted, after consideration of other forces, say political, social governmental policies etc., this statistical technique can be of immense help in decision making.

The success of any business depends on its future estimates. On the basis of these estimates a business man plans his production stocks, selling market, arrangement of additional funds etc. Forecasting is different from predictions and projections. Regression analysis, time series analysis, Index numbers are some of the techniques through which the predictions and projections are made. Where as forecasting is a method of foretelling the course of business activity based on the analysis of past and present data mixed with the consideration of ensuring economic policies and circumstances. In particularly forecasting means fore-warning. Forecasts based on statistical analysis are much reliable than a guess work.

According to T.S.Levis and and R.A. Fox, “ Forecasting is using the knowledge we have at one time to estimate what will happen at some future movement of time”.

3.8.Methods of Business forecasting:

There are three methods of forecasting

1. Naive method
2. Barometric methods
3. Analytical Methods

1. Naive method :

It contains only the economic rhythm theory.

2. Barometric methods:

It covers

i) Specific historical analogy

ii) Lead- Lag relationship

iii) Diffusion method

iv) Action –reaction theory

3. Analytical Methods:

It contains

i) The factor listing method

ii) Cross-cut analysis theory

iii) Exponential smoothing

iv) Econometric methods

The economic rhythm theory:

In this method the manufactures analysis the time-series data of his own firm and forecasts on the basis of projections so obtained. This method is applicable only for the individual firm for which the data are analysed, The forecasts under this method are not very reliable as no subjective matters are being considered.

Diffusion method of Business forecasting

The diffusion index method is based on the principle that different factors, affecting business, do not attain their peaks and troughs simultaneously. There is always time-log between them. This method has the convenience that one has not to identify which series has a lead and which has a lag. The diffusion index depicts the movement of broad group of series as a whole without bothering about the individual series. The diffusion index shows the percentage of a given set of series as expanding in a time period. It should be carefully noted that the peaks and troughs of diffusion index are not the peaks troughs of the business cycles. All series do not expand or contract concurrently. Hence if more than 50% are expanding at a given time, it is taken that the business is in the process of booming and vice - versa.

The graphic method is usually employed to work out the diffusion index. The diffusion index can be constructed for a group of business variables like prices, investments, profits etc.

Cross cut analysis theory of Business forecasting:

In this method a thorough analysis of all the factors under present situations has to be done and an estimate of the composite effect of all the factors is being made. This method takes into account the views of managerial staff, economists, consumers etc. prior to the forecasting. The forecasts about the future state of the business is made on the basis of over all assessment of the effect of all the factors.

Unit III

Self - Assessment

1. With the help of graph paper obtain the trend values.

Year	1996	1997	1998	1999	2000	2001	2002
Value	65	85	95	75	100	80	130

2. Draw a trend line by the method of semi-averages

Year	1993	94	95	96	97	98	99	2000
Value	210	200	215	205	220	235	210	235

3. Calculate trend value by taking 5 yearly periods of moving

average from the data given below

Year	1987	88	89	90	91	92	93	94	95	96	97	98	99	2000	01	02
Productions	4	5	6	7	9	6	5	7	8	7	6	8	9	10	7	9

4. Fit a straight line trend by the method of least square to the following data

Year	1996	97	98	99	2000	2001
Profit	300	700	600	800	900	700

Ans:

1. Trend values are 200.94, 205.31, 209.69, 214.06, 218.43, 222.80, 227.19, 231.56

2. 6.2, 6.6, 6.6, 6.8, 7.0, 6.6, 6.6, 7.2, 7.6, 8.0, 8.0

3. 446.67, 546.67, 626.67, 706.67, 786.67, 866.67

INDEX NUMBERS

OBJECTIVE

The objectives of index numbers are to provide a clear and concise measure of changes in economic data, facilitate comparisons over time and across regions, aid in the formulation and evaluation of economic policies, and adjust economic transactions for changes in economic conditions.

4.1.INTRODUCTION:

An index number is a statistical device for comparing the general level of magnitude of a group of related variables in two or more situation. If we want to compare the price level of 2000 with what it was in 1990, we shall have to consider a group of variables such as price of wheat, rice, vegetables, cloth, house rent etc., If the changes are in the same ratio and the same direction, we face no difficulty to find out the general price level. But practically, if we think changes in different variables are different and that too, upward or downward, then the price is quoted in different units i.e milk for litre, rice or wheat for kilogram, rent for square feet, etc

We want one figure to indicate the changes of different commodities as a whole. This is called an Index number. Index Number is a number which indicate the changes in magnitudes. M.Spiegel say, “ An index number is a statistical measure designed to show changes in variable or a group of related variables with respect to time, geographic location or other characteristic”. In general, index numbers are used to measure changes over time in magnitude which are not capable of direct measurement.

On the basis of study and analysis of the definition given above, the following characteristics of index numbers are apparent.

1. Index numbers are specified averages.
2. Index numbers are expressed in percentage.
3. Index numbers measure changes not capable of direct measurement.
4. Index numbers are for comparison.

4.2. Uses of Index numbers

Index numbers are indispensable tools of economic and business analysis. They are particular useful in measuring relative changes. Their uses can be appreciated by the following points.

1. They measure the relative change.
2. They are of better comparison
3. They are good guides.
4. They are economic barometers.
5. They are the pulse of the economy.
6. They compare the wage adjuster.
7. They compare the standard of living.
8. They are a special type of averages.
9. They provide guidelines to policy.
10. To measure the purchasing power of money

4.3.Types of Index numbers:

There are various types of index numbers, but in brief, we shall take three kinds and they are

(a) Price Index,

(b) Quantity Index and

(c) Value Index

a) Price Index:

For measuring the value of money, in general, price index is used. It is an index number which compares the prices for a group of commodities at a certain time as at a place with prices of a base period. There are two price index numbers such as whole sale price index numbers and retail price index numbers. The wholesale price index reveals the changes into general price level of a country, but the retail price index reveals the changes in the retail price of commodities such as consumption of goods, bank deposits, etc.

b) Quantity Index:

Quantity index number is the changes in the volume of goods produced or consumed. They are useful and helpful to study the output in an economy.

c) Value Index:

Value index numbers compare the total value of a certain period with total value in the base period. Here total value is equal to the price of commodity multiplied by the quantity consumed. Notation: For any index number, two time periods are needed for comparison. These are called the Base period and the Current period. The period of the year which is used as a basis for comparison is called the base year and the other is the current year. The various notations used are as given below: P1 = Price of current year q1 = Quantity of current year P0 = Price of base year q0 = Quantity of base year

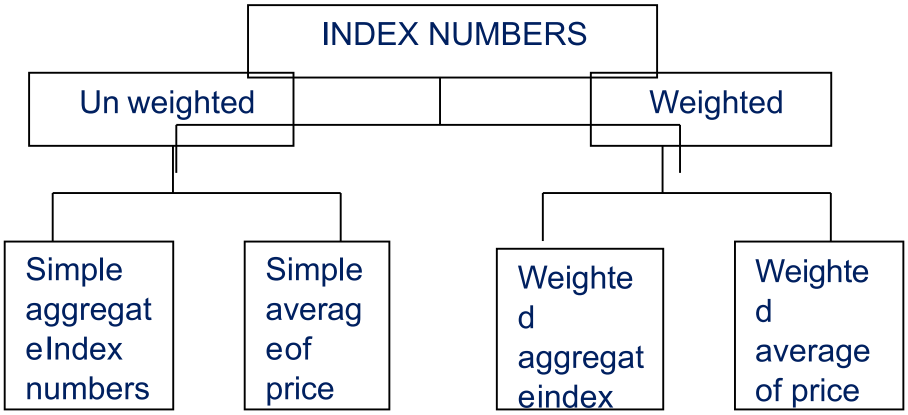
Problems in the construction of index numbers

No index number is an all purpose index number. Hence, there are many problems involved in the construction of index numbers, which are to be tackled by an economist or statistician. They are

1. Purpose of the index numbers
2. Selection of base period
3. Selection of items
4. Selection of source of data
5. Collection of data 6. Selection of average
7. System of weighting

4.4.Method of construction of index numbers:

Index numbers may be constructed by various methods as shown below:


INDEX NUMBERS
Un weighted
Weighted
Simple aggregat eIndex numbers
Simple averag eof price
Weighte d aggregat eindex
Weighte d average of price

4.4.1.Simple Aggregate Index Number

This is the simplest method of construction of index numbers. The price of the different commodities of the current year are added and the sum is divided by the sum of the prices of those commodities by 100. Symbolically,

Simple aggregate price index = $P_{01} = \sum p_1 / \sum p_0 \text{ x } 100$

Where, Σp_1 = total prices for the current year

Σp_0= Total prices for the base year

Example:

Calculate index numbers from the following data by simple aggregate method taking prices of 2000 as base.

Commodity	Price per unit (in Rupees)	
	2000	2004
A	80	95
B	50	60
C	90	100
D	30	45

Solution:

Commodity	Price per unit (in Rupees)	
	2000 (P_0)	2004 (P_1)
A	80	95
B	50	60
C	90	100
D	30	45
Total	250	300

Simple aggregate Price index = $P_{01} = \sum p_1 / \sum p_0 \text{ x } 100$

$= 300/250 \text{ x } 100$

$= 120$

4.4.2.Simple Average Price Relative index:

In this method, first calculate the price relative for the various commodities and then average of these relative is obtained by using arithmetic mean and geometric mean. When arithmetic mean is used for average of price relative, the formula for computing the index is

Simple average of price relative by arithmetic mean $p_{01} = \frac{\Sigma\left(\frac{p_1}{p_0}\times 100\right)}{n}$

P1 = Prices of current year

P0 = Prices of base year

n = Number of items or commodities when geometric mean is used for average of price relative, the formula for obtaining the index is

Simple average of price relative by geometric Mean

$$P_{01} = \text{Antilog}\left(\frac{\Sigma\log\left(\frac{p_1}{p_0}\times 100\right)}{n}\right)$$

Example:

From the following data, construct an index for 1998 taking 1997 as base by the average of price relative using (a) arithmetic mean and (b) Geometric mean

Commodity	Price in 1997	Price in 1998
A	50	70
B	40	60
C	80	100
D	20	30

Solution:

Price relative index number using arithmetic mean

Commodity	Price in 1997 (P_0) —	Price in 1998 (P_1)	$\frac{P_1}{P_0}\times 100$

A	50	70	140
B	40	60	150
C	80	100	125
D	20	30	150
		Total	565

Simple average of price relative index = (P01) =$\frac{\Sigma\left(\frac{p_1}{p0}\times 100\right)}{n}$

= 565/4=141.25

Simple average of price Relative index

$$P_{01} = \text{Antilog}\left(\frac{\Sigma\log\left(\frac{p_1}{p_0}\times 100\right)}{n}\right)$$

P_{01}= Antilog 8.5952 / 4

= Antilog [2.1488] = 140.9

4.4.3.Weighted aggregate index numbers

In order to attribute appropriate importance to each of the items used in an aggregate index number some reasonable weights must be used. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the most important ones are

1. Laspeyre's method
2. Paasche's method
3. Fisher's ideal Method
4. Bowley's Method
5. Marshall- Edgeworth method
6. Kelly's Method

1.Laspeyre' s method:

The Laspeyres price index is a weighted aggregate price index, where the weights are determined by quantities in the based period and is given by

Laspeyre' s price index = $P_{01} = \sum p_1q_0 / \sum p_0q_0 \times 100$

2.Paasche's method

The Paasche's price index is a weighted aggregate price index in which the weight are determined by the quantities in the current year. The formulae for constructing the index is

Paasche's price index number = $P_{01} = \sum p_1q_1 / \sum p_0q_1 \times 100$

Where P0 = Price for the base year
P1 = Price for the current year
q0 = Quantity for the base year
q1 = Quantity for the current year

Fisher' s ideal Method Fisher' s Price index number is the geometric mean of the Laspeyres and Paasche indices Symbolically Fisher' s ideal index number = P01 $^F = \sqrt{LxP}$

$$= \sqrt{\sum p1q0 \; / \; \sum p0q0 \; X \; \sum p_1q_1 / \sum p_0q_1} \quad X \; 100$$

It is known as ideal index number because
(a) It is based on the geometric mean
(b) It is based on the current year as well as the base year
(c) It conform certain tests of consistency
(d) It is free from bias.

3.Fisher' s ideal Method

Fisher' s Price index number is the geometric mean of the Laspeyres and Paasche indices Symbolically

Fisher' s ideal index number = P_{01}^F = $\sqrt{L \times P}$

$$= \sqrt{\sum p1q0 \; / \; \sum p0q0 \; X \; \sum p_1q_1 / \sum p_0q_1} \quad X \; 100$$

It is known as ideal index number because
(a) It is based on the geometric mean
(b) It is based on the current year as well as the base year
(c) It conform certain tests of consistency
(d) It is free from bias.

4.Bowley' s Method:

Bowley' s price index number is the arithmetic mean of Laspeyre' s and Paasche' s method. Symbolically

Bowley' s price index number = $P_{01}{}^{B} = \frac{L+P}{2}$

$$= 1/2\sqrt{\Sigma p1q0\,/\,\Sigma p0q0\ +\ \Sigma p_1q_1\,/\,\Sigma p_0q_1}\quad X\ \ 100$$

5.Marshall- Edgeworth method:

This method also both the current year as well as base year prices and quantities are considered. The formula for constructing the index is

Marshall Edgeworth price index $= P_{01}^{ME} = \frac{\Sigma(q_0+q_1)p_1}{\Sigma(q_0+q_1)p_0} \times 100$

$$= \frac{\Sigma p_1 q_0 + \Sigma p_1 q_1}{\Sigma p_0 q_0 + \Sigma p_0 q_1} \times 100$$

Where = q = $q_0 + q_1\ /\ 2$

Here the average of the quantities of two years is used as weights

Example:

Construct price index number from the following data by applying 1. Laspeyere's Method 2. Paasche's Method 3. Fisher's ideal Method

Commodity	2000		2001	
	Price	Qty	Price	Qty
A	2	8	4	5
B	5	12	6	10
C	4	15	5	12
D	2	18	4	20

Solutions:

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	$p_1\,q_1$
A	2	8	4	5	16	10	32	20
B	5	12	6	10	60	50	72	60
C	4	15	5	12	60	48	75	60
D	2	18	4	20	36	40	72	80
					172	148	251	220

$$\text{Laspeyre' s price index } = P_{01}\square^{L} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$
$$= \frac{251}{172} \times 100 = 145.93$$

$$\text{Paasche price index number } = P_{01}\square^{P} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times 100$$
$$= \frac{220}{148} \times 100$$
$$= 148.7$$
$$\text{Fisher' } s \text{ ideal index number } = \sqrt{L \times P}$$
$$= \sqrt{(145.9) \times (148.7)}$$
$$= \sqrt{21695.33}$$
$$= 147.3$$

Or

$$\text{Fisher' s ideal index number } = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}} \times 100$$
$$= \sqrt{\frac{251}{172} \times \frac{220}{148}} \times 100$$
$$= \sqrt{(1.459) \times (1.487)} \times 100$$
$$= \sqrt{2.170} \times 100$$
$$= 1.473 \times 100 = 147.3$$

Interpretation

The results can be interpreted as follows:

If 100 rupees were used in the base year to buy the given commodities, we have to use Rs 145.90 in the current year to buy the same amount of the commodities as per the Laspeyre’ s formula. Other values give similar meaning .

Example :

Calculate the index number from the following data by applying

(a) Bowley’s price index

(b) (b)Marshall- Edgeworth price index

Commodity	Base year		Current year	
	Quantity	Price	Quantity	Price
A	10	3	8	4
B	20	15	15	20
C	2	25	3	30

Solution:

Commodity	q_0	P_0	q_1	P_1	p_0q_0	p_0q_1	p_1q_0	$p_1\ q_1$
A	10	3	8	4	30	24	40	32
B	20	15	15	20	300	225	400	300
C	2	25	3	30	50	75	60	90
					380	324	500	422

Bowley's price index number

$$= \frac{1}{2}\left[\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} + \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}\right] \times 100$$

$$= \frac{1}{2}\left[\frac{500}{380} + \frac{422}{324}\right] \times 100$$

$$= \frac{1}{2}[1.316 + 1.302] \times 100$$

$$= \frac{1}{2}[2.168] \times 100$$

$$= 1.309 \times 100$$

$$= 130.9$$

Marshall Edgeworths price index Number

$$= P_{01}{\square}^{ME} = \frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} \times 100$$

$$= \left[\frac{500}{380} + \frac{422}{324}\right] \times 100$$

$$= \left[\frac{922}{704}\right] \times 100$$

=131.0

Example:

Calculate a suitable price index from the following data

Commodity	Quantity	Price	
		1996	1997
A	20	2	4
B	15	5	6
C	8	3	2

Solution:

Here the quantities are given in common we can use Kelly's index price number and is given by

Kelly's Price index number = $P_{01}{}^{k} = \Sigma p_1 q / \Sigma p_0 \times 100$

$= 186 / 139 \times 100 = 133.81$

Commodity	q	P_0	P_1	p_0q	P_1q
A	20	2	4	40	80
B	15	5	6	75	90
C	8	3	2	24	16
			Total	139	186

4.4.4.Weighted Average of Price Relative index.

When the specific weights are given for each commodity, the weighted index number is calculated by the formula.

Weighted Average of Price Relative index = $\Sigma pw / \Sigma w$

Where w = the weight of the commodity

P = the price relative index

$= P_1 / P_0 \times 100$

When the base year When the base year value P0q0 is taken as the weight i.e. $W=P_0q_0$ then the formula is

Weighted Averageof Price Relative index

$$= \frac{\Sigma\left(\frac{p_1}{p_0} \times 100\right) \times p_0 q_0}{\Sigma p_0 q_0}$$

$$= \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times 100$$

This is nothing but Laspeyre's formula. When the weights are taken as w = p_0q_1, the formula is

$$= \frac{\Sigma\left(\frac{p_1}{p_0}\times 100\right)\times p_0q_1}{\Sigma p_0q_0}$$

$$= \frac{\Sigma p_1q_1}{\Sigma p_0q_1}\times 100$$

This is nothing but Paasche's Formula

Example:

Compute the weighted index number for the following data

Commodity	Price		Weight
	Current year	Base year	
A	5	4	60
B	3	2	50
C	2	1	30

Solution:

Commodity	P_1	P_0	W	$P= \frac{p_1}{p_0}\times 100$	PW
A	5	4	60	125	7500
B	3	2	50	150	7500
C	2	1	30	200	6000
			140		21000

Weighted Average of Price Relative index = Σ pw / Σ w = 21000 / 140 = 150

4.5.Quantity or Volume index number:

Price index numbers measure and permit comparison of the price of certain goods. On the other hand, the quantity index numbers measure the physical volume of production, employment and etc. The most common type of the quantity index is that of quantity produced.

$$\text{Laspeyre's quantity index number} = Q_{01}{\square}^{L} = \frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times 100$$

$$\text{Paasche's quantity index number} = Q_{01}{\square}^{P} = \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1} \times 100$$

$$Q_{01}{\square}^{F} = \sqrt{L \times P}$$

$$\text{Fisher's quantity index number} \sqrt{\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0} \times \frac{\Sigma q_1 p_1}{\Sigma q_0 p_1}} \times 100$$

These formulae represent the quantity index in which quantities of the different commodities are weighted by their prices

4.6. Tests of Consistency of index numbers:

Several formulae have been studied for the construction of index number. The question arises as to which formula is appropriate to a given problems. A number of tests been developed and the important among these are

1. Unit test

2. Time Reversal test

3. Factor Reversal test

1.Unit test:

The unit test requires that the formula for constructing an index should be independent of the units in which prices and quantities are quoted. Except for the simple aggregate index (unweighted) , all other formulae discussed in this chapter satisfy this test.

2.Time Reversal test:

Time Reversal test is a test to determine whether a given method will work both ways in time, forward and backward. In the words of Fisher, “the formula for calculating the index number should be such that it gives the same ratio between one point of comparison and the other, no matter which of the two is taken as base". Symbolically, the following relation should be satisfied.

P01 × P10 = 1

Where P01is the index for time ' 1' as time ' 0' as base and P10is the index for time ' 0' as time ' 1' as base. If the product is not unity, there is said to be a time bias is the method. Fisher' s ideal index satisfies the time reversal test.

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$\text{Then } P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$\sqrt{1} = 1$$

Therefore Fisher ideal index satisfies the time reversal test

3.Factor Reversal test:

Another test suggested by Fisher is known s factor reversal It holds that the product of a price index and the quantity index should be equal to the corresponding value index. In the words of Fisher, "Just as each formula should permit the interchange of the two times without giving inconsistent results, so it ought to permit interchanging the prices and quantities without giving inconsistent result, ie, the two results multiplied together should give the true value ratio.

In other word, if P_{01} represent the changes in price in the current year and Q_{01} represent the changes in quantity in the current year, then

$$P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Thus based on this test, if the product is not equal to the value ratio, there is an error in one or both of the index number. The Factor reversal test is satisfied by the Fisher' s ideal index

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$P_{10} = \sqrt{\frac{\Sigma q_1 p_0}{\Sigma p_0 q_0} \times \frac{\Sigma q_1 p_1}{\Sigma p_1 q_0}}$$

$$\text{Then } P_{01} \times P_{10} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1} \times \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$$

$$= \sqrt{\left(\frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}\right)^2}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

$$P_{01} \text{ x } Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

the factor reversal test is satisfied by the Fisher's ideal index.

Example:

Construct Fisher' s ideal index for the Following data. Test whether it satisfies time reversal test and factor reversal test.

	Base year		Current year	
Commodity	Quantity	Price	Quantity	Price
A	12	10	15	12
B	15	7	20	5
C	5	5	8	9

Solution

Commodity	q_0	p_0	q_1	p_1	P_0q_0	p_0q_1	p_1q_0	p_1q_1
A	12	10	15	12	120	150	144	180
B	15	7	20	5	105	140	75	100
C	5	5	8	9	25	40	45	72

					250	330	264	352

Fisher ideal index number P_{01}^{F} = $\sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$ x 100

$$= \sqrt{\frac{264}{250} \times \frac{352}{330}} \text{ x } 100$$

$$= \sqrt{(1.056)\ -\ (1.067)} \text{ x } 100$$

$$= \sqrt{1.127} \text{ x } 100$$

= 1.062 x 100 = 106.2

Time Reversal test:

Time Reversal test is satisfied when P01× P10 = 1

$$P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$$

$$= \sqrt{\frac{264}{250} \times \frac{352}{330}}$$

$$P_{10} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_0 q_0}{\Sigma p_1 q_0}}$$

$$= \sqrt{\frac{300}{352} \times \frac{250}{264}}$$

Now P_{01} x P_{10} = $\sqrt{\ \frac{300}{352} \times \frac{250}{264}\ x\ \frac{264}{250} \times \frac{352}{330}}$

$$=\sqrt{1}$$

= 1

Hence Fisher ideal index satisfy the time reversal test

Factor Reversal test:

Factor Reversal test is satisfied when $P_{01} \times Q_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$

Now $P_{01} = \sqrt{\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

$$= \sqrt{\frac{264}{250} \times \frac{352}{330}}$$

$Q_{01} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}}$

$$= \sqrt{\frac{300}{352} \times \frac{250}{264}}$$

Then $P_{01} \times Q_{01} = \sqrt{\frac{\Sigma p_0 q_1}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0} \; x \; \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}}$

$$= \sqrt{\left(\frac{352}{250}\right)^2}$$

$$= \frac{352}{250}$$

$$= \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$$

Hence Fisher ideal index number satisfy the factor reversal test.

4.7.Consumer Price Index

Consumer Price index is also called the cost of living index. It represent the average change over time in the prices paid by the ultimate consumer of a specified basket of goods and services. A change in the price level affects the costs of living of different classes of people differently. The general index number fails to reveal this. So there is the need to construct consumer price index. People consume different types of commodities. People's consumption habit is also different from man to man, place to place and class to class i.e richer class, middle class and poor class.

The scope of consumer price is necessary, to specify the population group covered. For example, working class, poor class, middle class, richer class, etc and the geographical areas must be covered as urban, rural, town, city etc

4.7.1.Use of Consumer Price index

The consumer price indices are of great significance and is given below.

1. This is very useful in wage negotiations, wage contracts and dearness allowance adjustment in many countries.
2. At government level, the index numbers are used for wage policy, price policy, rent control, taxation and general economic policies.
3. Change in the purchasing power of money and real income can be measured.
4. Index numbers are also used for analysing market price for particular kinds of goods and services.

4.7.2.Method of Constructing Consumer price index:

There are two methods of constructing consumer price index. They are

1.Aggregate Expenditure method (or) Aggregate method.

2. Family Budget method (or) Method of Weighted Relative method.

Aggregate Expenditure method:

This method is based upon the Laspeyre's method. It is widely used. The quantities of commodities consumed by a particular group in the base year are the weight. The formula is Consumer Price Index number = $\sum p1\ q1 / \sum p_0\ q_0 \quad x\ 100$

Family Budget method or Method of Weighted Relatives:

Thismethod is estimated an aggregate expenditure of an average family on various items and it is weighted. The formula is

Consumer Price index number = $\sum pw$ / $\sum w$

Where $P = P_1 / P_2 \times 100$ for each item. w = value weight (i.e) p_0q_0

“Weighted average price relative method” which we have studied before and “Family Budget method” are the same for finding out consumer price index.

Example:

Construct the consumer price index number for 1996 on the basis of 1993 from the following data using Aggregate expenditure method.

		Price in	
Commodity	Quantity consumed	1993	1996
A	100	8	12
B	25	6	7
C	10	5	8
D	20	15	18

Solution:

Commodity	q_0	p_0	p_1	p_0q_0	p_1q_0
A	100	8	12	800	1200
B	25	6	7	150	175
C	10	5	8	50	80
D	20	15	18	300	360
			Total	1300	1815

Consumer price index by Aggregate expenditure method

$= \sum p1\ q1 / \sum p_0\ q_0 \quad x\ 100$

= 1815 / 1300 x 100
= 139.6

Example:

Calculate consumer price index by using Family Budget method for year 1993 with 1990 as base year from the following data.

		Price in	
Items	Weights	1990 (Rs.)	1993 (Rs.)

Food	35	150	140
Rent	20	75	90
Clothing	10	25	30
Fuel and lighting	15	50	60
Miscellaneous	20	60	80

Solution:

Items	W	P_0	P_1	$P = \frac{p_1}{p_0} \times 100$	PW
Food	35	150	140	93.33	3266.55
Rent	20	75	90	120.00	2400.00
Clothing	10	25	30	120.00	1200.00
Fuel and lighting	15	50	60	120.00	1800.00
Miscellaneous	20	60	80	133.33	2666.60
	100				11333.15

Consumer price index by Family Budget method = ∑pw / ∑w

= 11333.15 / 113.33

Unit IV

Self – Assessment

1. Calculate the price index number by

(i) Laspeyre' s method

(ii) Paasche' s method

(iii) Fisher's ideal index method.

	1990		1995	
Commodity	price	quantity	Price	Quantity
A	20	15	30	20
B	15	10	20	15
C	30	20	25	10
D	10	5	12	10

2. Calculate Fisher's ideal index for the following data. Also, test whether it satisfies the time reversal test and factor reversal test

	Price		Quantity	
Commodity	2000	2002	2000	2002
A	6	35	10	40
B	10	25	12	30
C	12	15	8	20

Ans:

1.

(i) L = 110

(ii) P = 123.9

(iii) F = 116.7

2. 296

TESTING OF HYPOTHESIS

OBJECTIVE:

The objective of hypothesis testing is to provide a structured method for making inferences about a population based on sample data. It is a fundamental process in statistics and scientific research that allows researchers to make decisions or draw conclusions with a certain level of confidence level.

5.1.Null Hypothesis and Alternative Hypothesis:

Hypothesis testing begins with an assumption called a Hypothesis, that we make about a population parameter. A hypothesis is a supposition made as a basis for reasoning. The conventional approach to hypothesis testing is not to construct a 112 single hypothesis about the population parameter but rather to set up two different hypothesis. So that of one hypothesis is accepted, the other is rejected and vice versa.

5.2.Null Hypothesis:

A hypothesis of no difference is called null hypothesis and is usually denoted by H_0 " Null hypothesis is the hypothesis" which is tested for possible rejection under the assumption that it is true " by Prof. R.A. Fisher. It is very useful tool in test of significance. For example: If we want to find out whether the special classes (for Hr. Sec. Students) after school hours has benefited the students or not. We shall set up a null hypothesis that "H0: special classes after school hours has not benefited the students".

5.3. Alternative Hypothesis:

Any hypothesis, which is complementary to the null hypothesis, is called an alternative hypothesis, usually denoted by H1, For example, if we want to test the null hypothesis that the population has a specified mean μ_0 (say), i.e., :

Step 1: null hypothesis H_0: $\mu = \mu_0$ then

2. Alternative hypothesis may be

i) H_1: $\mu \neq \mu_0$ (ie $\mu > \mu0$ or $\mu < \mu0$)

ii) H_1: $\mu > \mu_0$

iii) H_1: $\mu < \mu_0$

the alternative hypothesis in (i) is known as a two– tailed alternative and the alternative in (ii) is known as right-tailed (iii) is known as left–tailed alternative respectively. The settings of alternative hypothesis is very important since it enables us to decide whether we have to use a single– tailed (right or left) or two tailed test

5.4.Level of significance and Critical value:

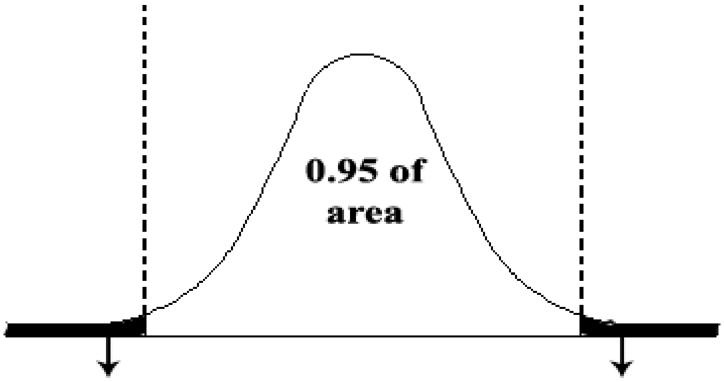
Level of significance:

In testing a given hypothesis, the maximum probability with which we would be willing to take risk is called level of significance of the test. This probability often denoted by "α" is generally specified before samples are drawn.

The level of significance usually employed in testing of significance are 0.05(or 5 %) and 0.01 (or 1 %). If for example a 0.05 or 5 % level of significance is chosen in deriving a test of hypothesis, then there are about 5 chances in 100 that we would reject the hypothesis when it should be accepted. (i.e.,) we are about 95 % confident that we made the right decision. In such a case we say that the hypothesis has been rejected at 5 % level of significance which means that we could be wrong with probability 0.05.

The following diagram illustrates the region in which we could accept or reject the null hypothesis when it is being tested at 5 % level of significance and a two-tailed test is employed.

Accept the null hypothesis if the sample statistics falls in this region



Reject the null hypothesis if the sampleStatistics falls in these two region

Note: Critical Region: A region in the sample space S which amounts to rejection of H0 is termed as critical region or region of rejection.

5.5.Critical Value:

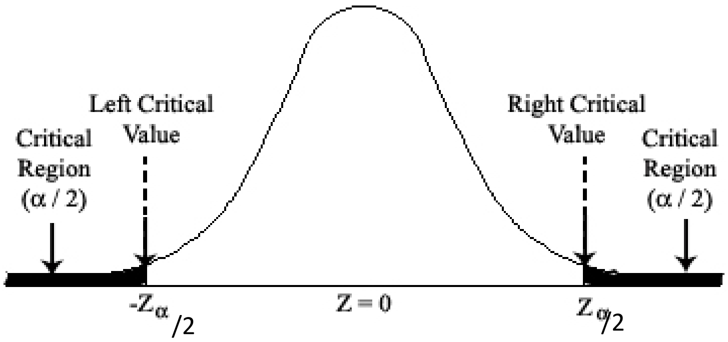
The value of test statistic which separates the critical (or rejection) region and the acceptance region is called the critical value or significant value. It depends upon i) the level of significance used and ii) the alternative hypothesis, whether it is two-tailed or single-tailed For large samples the standard normal variate corresponding to the

Statistics t , $Z = \left|\frac{t-E(t)}{S\cdot E\cdot(t)}\right|$~ N(0,1)

asymptotically as n → ∞

The value of z under the null hypothesis is known as test statistic. The critical value of the test statistic at the level of significance α for a two- tailed test is given by Zα/2 and for a one tailed test by Zα. where Zα is determined by equation P(|Z| >Zα)= α Zα is the value so that the total area of the critical region on both tails is α . ∴ P(Z > Zα) = α / 2 . Area of each tail is. α/2

Zα is the value such that area to the right of Zα and to the left of – Zα is α / 2 as shown in the following diagram



5.6.Chi square statistic:

Various tests of significance described previously have mostly applicable to only quantitative data and usually to the data which are approximately normally distributed. It may also happens that we may have data which are not normally distributed. Therefore there arises a need for other methods which are more appropriate for studying the differences between the expected and observed frequencies. The other method is called Non-parametric or distribution free test. A non- parametric test may be defined as a statistical test in which no hypothesis is made about specific values of parameters. Such non-parametric test has assumed great importance in statistical analysis because it is easy to compute.

5.6.1.Definition:

The Chi- square (χ2) test (Chi-pronounced as ki) is one of the simplest and most widely used non-parametric tests in statistical work. The χ2 test was first used by Karl Pearson in the year 1900. The quantity χ2 describes the magnitude of the discrepancy between theory and observation. It is defined as

$$x^2 = \sum_{i=1}^{n} [\frac{(0_i - E_i)^2}{E_i}]$$

Where O refers to the observed frequencies and E refers to the expected frequencies.

Note: If χ2 is zero, it means that the observed and expected frequencies coincide with each other. The greater the discrepancy between the observed and expected frequencies the greater is the value of χ2.

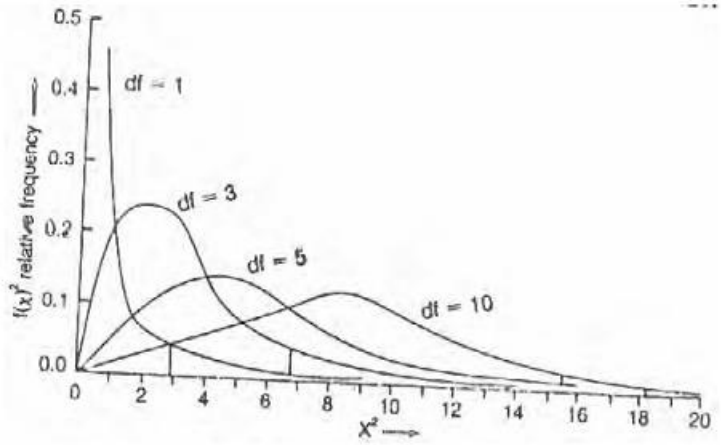
Chi square- Distribution:

The square of a standard normal variate is a Chi-square variate with 1 degree of freedom i.e., If X is normally distributed 2 with mean μ and standard deviation σ,

Then,

$$\left(\frac{x-\mu}{\sigma}\right)^2$$

is a Chi σ square variate (χ2) with 1 d.f. The distribution of Chi-square depends on the degrees of freedom. There is a different distribution for each number of degrees of freedom.



properties of Chi-square distribution:

1. The Mean of χ^2 distribution is equal to the number ofdegrees of freedom (n)

2. The variance of X^2 distribution is equal to 2n
3. The median of χ^2 distribution divides, the area of the curveinto two equal parts, each part being 0.5.
4. The mode of χ^2 distribution is equal to (n-2)
5. Since Chi-square values always positive, the Chi squarecurve is always positively skewed.
6. Since Chi-square values increase with the increase in the degrees of freedom, there is a new Chi-square distributionwith every increase in the number of degrees of freedom.
7. The lowest value of Chi-square is zero and the highest valueis infinity ie $\chi^2 \geq 0$.
8. When Two Chi- squares χ^2_1 and $\chi^2_2$are independent χ^{22} distribution with n_1 and n_2 degrees of freedom and their sum $\chi^2_1 + \chi^2_2$will follow $\square^2$ distribution with $(n_1 + n_2)$ degrees offreedom.
9. When n (d.f) > 30, the distribution of $\sqrt{2x^2}$ approximatelyfollows normal distribution. The mean of the distribution $\sqrt{2x^2}$ is $\sqrt{2n-1}$ and the standard deviation is equal to 1.

5.6.2.Conditions for applying χ^2 test:

The following conditions should be satisfied before applying χ^2 test.

1. N, the total frequency should be reasonably large, say greater than 50.
2. No theoretical cell-frequency should be less than 5. If it is less than 5, the frequencies should be pooled together in order to make it 5 or more than 5.
3. Each of the observations which makes up the sample for this test must be independent of each other.
4. χ^2 test is wholly dependent on degrees of freedom.

5.7.Testing the Goodness of fit (Binomial and Poisson Distribution):

Karl Pearson in 1900, developed a test for testing the significance of the discrepancy between experimental values and the theoretical values obtained under some theory or hypothesis. This test is known as 2-test of goodness of fit and is used to test if the deviation between observation (experiment) and theory may be attributed to chance or if it is really due to the inadequacy of the theory to fit the observed data.

Under the null hypothesis that there is no significant difference between the observed and the theoretical values. Karl Pearson proved that the statistic

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(\mathrm{Oi} - \mathrm{Ei})^2}{\mathrm{Ei}}\right]$$
$$= \frac{(\mathrm{O}_1 - \mathrm{E}_1)^2}{\mathrm{E}_1} + \frac{(\mathrm{O}_2 - \mathrm{E}_2)^2}{\mathrm{E}_2} + \cdots \ldots . \frac{(\mathrm{O}_n - \mathrm{E}_n)^2}{\mathrm{E}_n}$$

Follows χ^2-distribution with v = n – k – 1 d.f. where 0_1, 0_2, ...0_n are the observed frequencies, E_1, E_2..E_n, corresponding to the expected frequencies and k is the number of parameters to be estimated from the given data. A test is done by comparing the computed value with the table value of ^{2} for the desired degrees offreedom.

Example:

Four coins are tossed simultaneously and the number of heads occurring at each throw was noted. This was repeated 240 times with the following results.

No. of heads	0	1	2	3	4
No. of throws	13	64	85	58	20

Fit a Binomial distribution assuming under the hypothesis that the coins are **unbiased.**

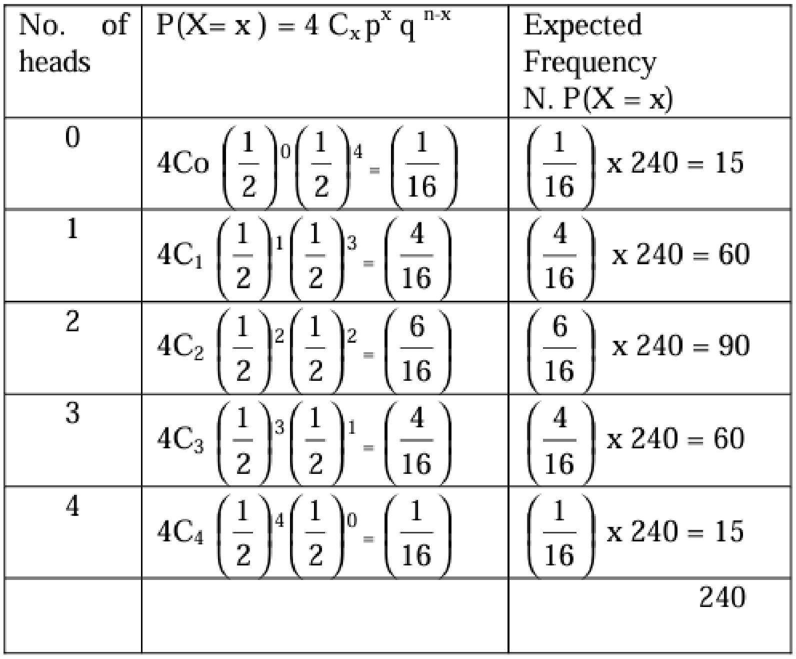
Solution:

Null hypothesis:

H_0:the given data fits the binomial distribution.i,e the coins are unbiased

p = q = 1/2 n = 4 N = 240

Computation of expected frequencies:



No. of heads	$P(X= x) = 4\ C_x\, p^x\, q^{\ n-x}$	Expected Frequency N. P(X = x)
0	$4Co\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^4 = \left(\frac{1}{16}\right)$	$\left(\frac{1}{16}\right)$ x 240 = 15
1	$4C_1\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^3 = \left(\frac{4}{16}\right)$	$\left(\frac{4}{16}\right)$ x 240 = 60
2	$4C_2\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \left(\frac{6}{16}\right)$	$\left(\frac{6}{16}\right)$ x 240 = 90
3	$4C_3\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^1 = \left(\frac{4}{16}\right)$	$\left(\frac{4}{16}\right)$ x 240 = 60
4	$4C_4\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^0 = \left(\frac{1}{16}\right)$	$\left(\frac{1}{16}\right)$ x 240 = 15
		240

Computation of chi square values:

Observed Frequency O	Expected Frequency E	$(O - E)^2$	$\left(\frac{(O - E)^2}{E}\right)$
13	15	4	0.27
64	60	16	0.27
85	90	25	0.28
58	60	4	0.07
20	15	25	1.67
			2.56

$$\chi_0^{\ 2} = \Sigma\left(\frac{(O - E)^2}{E}\right) = 2.56$$

Expected Value:

$$\chi_e^2 = \Sigma\left(\frac{(O - E)^2}{E}\right) \text{ follows } \chi^2\text{-distribution with (n–k–1) d.f.}$$

(Here k = 0, since no parameter is estimated from the data)

$$= 9.488 \quad \text{for} \quad \nu = 5-1= 4 \text{ d.f.}$$

Inference:

Since $\chi_0^2 < \chi^2$ we accept our null hypothesis at 5% level of significance and say that the given data fits Binomial distribution

5.8.T-test

In the previous chapter we have discussed problems relating to large samples. The large sampling theory is based upon two important assumptions such as

a) The random sampling distribution of a statistic is approximately normal and

b) The values given by the sample data are sufficiently close to the population values and can be used in their place for the calculation of the standard error of the estimate.

The above assumptions do not hold good in the theory of small samples. Thus, a new technique is needed to deal with the theory of small samples. A sample is small when it consists of less than 30 items. (n< 30)

Since in many of the problems it becomes necessary to take a small size sample, considerable attention has been paid in developing suitable tests for dealing with problems of small samples. The greatest contribution to the theory of small samples is that of Sir William Gosset and Prof. R.A. Fisher. Sir William Gosset published his discovery in 1905 under the pen name ‘ Student’ and later on developed and extended by Prof. R.A.Fisher. He gave a test popularly known as ‘ t-test’ .

5.8.1.t - statistic definition:

If x1, x2, ……x_n is a random sample of size n from a normal population with mean μ and variance σ^2, then Student’ s t-statistic is defined as

$$t = \frac{\bar{x}-\mu}{S/\sqrt{n}}$$

Where $\bar{x} = \frac{\sum x}{n}$ is the sample mean and $S^2 = \frac{1}{n-1}\Sigma(x-\bar{x})^2$

is an unbiased estimate of the population variance σ^2 It followsstudent' s t-distribution with v = n -1 d.f

Assumptions for students t-test:

1.The parent population from which the sample drawn is normal.

2.The sample observations are random and independent.

3.The population standard deviation σ is not known.

Properties of t- distribution:

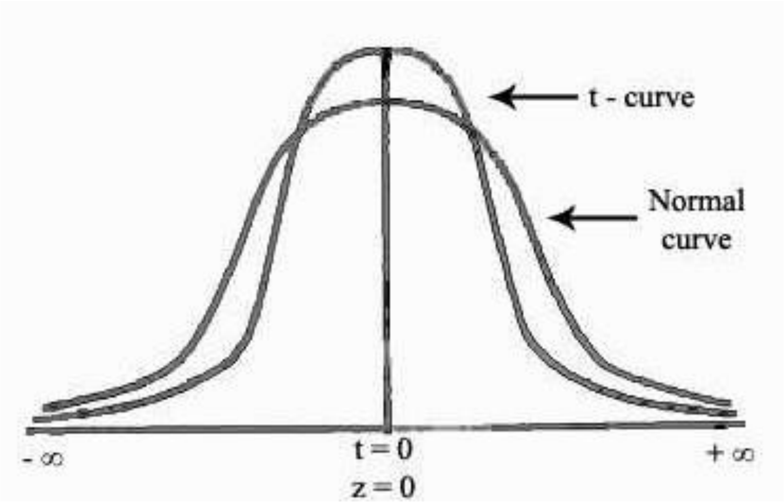
1.t-distribution ranges from □□ to □ just as does a normal distribution.

2.Like the normal distribution, t-distribution also symmetrical and has a mean zero.

3.t-distribution has a greater dispersion than the standard normal distribution.

4.As the sample size approaches 30, the t-distribution, approaches the Normal distribution.

Comparison between Normal curve and corresponding t - curve:



Degrees of freedom (d.f):

Suppose it is asked to write any four number then one will have all the numbers of his choice. If a restriction is applied or imposed to the choice that the sum of these number should be 50. Here, we have a choice to select any three numbers, say 10, 15, 20 and the fourth number is 5: [50 - (10 +15+20)]. Thus our choice of freedom is reduced by one, on the condition that the total be 50. Therefore the restriction placed on the freedom is one and degree of freedom is three. As the restrictions increase, the freedom is reduced.

The number of independent variates which make up the statistic is known as the degrees of freedom and is usually denoted by v(Nu)

The number of degrees of freedom for n observations is n - k where k is the number of independent linear constraint imposed upon them.

For the student' s t-distribution. The number of degrees of freedom is the sample size minus one. It is denoted by v= n -1

The degrees of freedom plays a very important role in χ^2 test of a hypothesis.

When we fit a distribution the number of degrees of freedom is (n– k-1) where n is number of observations and k is number of parameters estimated from the data.

For e.g., when we fit a Poisson distribution the degrees of freedom is

v= n – 1 -1

In a contingency table the degrees of freedom is (r-1) (c -1) where r refers to number rows and c refers to number of columns.

Thus in a 3 x 4 table the d.f are (3-1) (4-1) = 6 d.f In a 2 x 2 contingency table the d.f are (2-1) (2-1) = 1

In case of data that are given in the form of series of variables in a row or column the d.f will be the number of observations in a series less one ie v= n-1

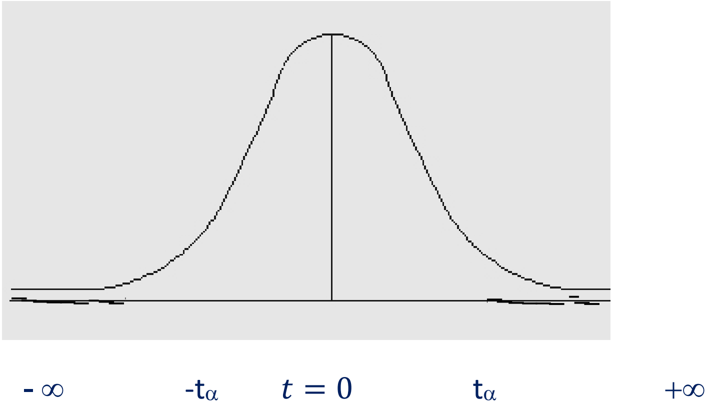
Critical value of t:

The column figures in the main body of the table come under the headings $t_{0.100}$,

$t_{0.50}$, $t_{0.025}$, $t_{0.010}$ and $t_{0.005}$. The subscriptsgive the proportion of the distribution in ‘ tail’ area. Thus for two- tailed test at 5% level of significance there will be two rejection areas each containing 2.5% of the total area and the required column is headed t0.025

For example,

$t_v(.05)$ for single tailed test = $t_v(0.025)$ for two tailed test $t_v(.01)$ for single tailed test = $t_v(0.005)$ for two tailed test Thus for one tailed test at 5% level the rejection area lies in one end of the tail of the distribution and the required column is headed $t_{0.05}$.



Applications of t – distributions:

The t-distribution has a number of applications in statistics, of which we shall discuss the following in the coming sections:

(i) t-test for significance of single mean, population variance being unknown.

(ii) t-test for significance of the difference between two sample means, the population variances being equal but unknown.

(a) Independent samples

(b) Related samples: paired t-test

Test of significance for Mean:

We set up the corresponding null and alternative hypotheses as follows:

H_0: $\mu = \mu_0$; There is no significant difference between the sample mean and population Mean.

H_1: $\mu \neq \mu_0$ ($\mu < \mu_0$ (or) $\mu > \mu_0$)

Level of significance:

5% or 1%

Calculation of statistic:

Under H_0 the test statistic is

$$t_0 = \left| \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \right| \quad \text{or} \quad \left| \frac{\bar{x} - \mu}{s / \sqrt{n-1}} \right|$$

Where

$\bar{x} = \frac{\Sigma x}{n}$ is the sample mean and

$$S^2 = \frac{1}{n-1} \Sigma \left(x - \bar{x} \right) \text{ (or) } s^2 = \frac{1}{n} \Sigma (x - \bar{x})^2$$

Expected value:

$$t_e = \left| \frac{\bar{x} - \mu}{S / \sqrt{n}} \right| \sim \text{student' s t-distribution with (n-1) d.f}$$

Inference :

If $t_0 \le t_e$ it falls in the acceptance region and the null hypothesis is accepted and if $t_0 \le t_e$ the null hypothesis H0 may be rejected at the given level of significance.

Example:

Certain pesticide is packed into bags by a machine. A random sample of 10 bags is drawn and their contents are found to weigh (in kg) as follows

50 49 52 44 45 48 46 45 49 45

Test if the average packing can be taken to be 50 kg.

Solution:

Null hypothesis:

$H_0 : \mu = 50$ kgs in the average packing is 50 kgs

Alternative Hypothesis:

$H_1 : \mu \neq 50$kgs (Two -tailed)

Level of Significance:

Let $\alpha = 0.05$

Calculation of sample mean and S.D

X	d = x –48	d^2
50	2	4
49	1	1
52	4	16
44	–4	16
45	–3	9
48	0	0
46	–2	4
45	–3	9
49	+1	1
45	–3	9
Total	–7	69

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

$$= 48 + \frac{-7}{10}$$

$$= 48–0.7 = 47.3$$

$$S^2 = \frac{1}{n-1}\left[\sum d^2 - \frac{(\sum d)^2}{n}\right|$$

$$= \frac{1}{9}\left[69 - \frac{(7^2)}{10}\right|$$

$$= \frac{64.1}{9} = 7.12$$

Calculation of Statistic:
Under H0 the test statistic is :

$$t_0 = \left|\frac{\bar{x}-\mu}{\sqrt{S^2/n}}\right|$$

$$= \left|\frac{47.3-50.0}{\sqrt{7.12/10}}\right|$$

$$= \frac{2.7}{\sqrt{0.712}} \quad = 3.2$$

Expected value:

$t_e = | \dfrac{\bar{x}-\mu}{S2/\sqrt{n}}$ follows t-distribution with (26-1) = 25d.f

Inference:

Since $t_0 > t_e$, H_0 is rejected at 5% level of significance. Hence we conclude that advertisement is certainly effective in increasing the sales

5.8.2.Related samples –Paired t-test:

In the t-test for difference of means, the two samples were independent of each other. Let us now take a particular situations where

(i) The sample sizes are equal; i.e., $n_1 = n_2 = n$(say), and

(ii) The sample observations $(x_1, x_2, \ldots\ldots x_n)$ and $(y_1, y_2, \ldots\ldots y_n)$ are not completely independent but they are dependent in pairs.

That is we are making two observations one before treatment and another after the treatment on the same individual. For example a business concern wants to find if a particular media of promoting sales of a product, say door to door canvassing or advertisement in papers or through T.V. is really effective. Similarly a pharmaceutical company wants to test the efficiency of a particular drug, say for inducing sleep after the drug is given. For testing of such claims gives rise to situations in (i) and (ii) above, we apply paired t-test.

Paired – t –test:

Let di = Xi – Yi (i = 1, 2, ……n) denote the difference in the observations for the ith unit.

Null hypothesis:

$H_0 : \mu_1 = \mu_2$ ie the increments are just by chance

Alternative Hypothesis:

$H_1 : \mu_1 \neq \mu_2$ $(\mu_1 > \mu_2 \text{(or)} \mu_1 < \mu_2)$

Calculation of test statistics:

Level of significance: Let $\alpha = 0.05$

Typist	d	d^2
A	2	4
B	4	16
C	0	0
D	3	9
E	–1	1
F	4	16
G	–3	9
H	2	4
I	5	25
	$\Sigma d = 16$	$\Sigma d^2 = 84$

$$t_0 = \left|\frac{\bar{d}}{S/\sqrt{n}}\right|$$

$$\text{where } \bar{d} = \frac{\Sigma d}{n} \text{ and } S^2 = \frac{1}{n-1}\Sigma(d-\bar{d})^2 = \frac{1}{n-1}\left[\Sigma d^2 - \frac{(\Sigma d)^2}{n}\right]$$

Expected value:

$$t_e = \left|\frac{\bar{d}}{S/\sqrt{n}}\right|$$

follows t-distribution with n –1 d.f

Inference:

By comparing t0 and te at the desired level of significance, usually 5% or 1%, we reject or accept the null hypothesis.

5.9.F – Statistic Definition:

If X is a χ^2 variate with n_1 d.f. and Y is an independent χ^2 - variate with n_2 d.f., then F - statistic is defined as

$$F = \frac{X/n_1}{Y/n_2}$$

i.e. F - statistic is the ratio of two independent chi-square variates divided by their respective degrees of freedom. This statistic follows G.W. Snedocor' s F-distribution with (n1, n2) d.f.

Testing the ratio of variances:

Suppose we are interested to test whether the two normal population have same variance or not. Let x1, x2, x3 ….. 1 n x , be a random sample of size n1, from the first population with variance σ1 2 and y1, y2, y3 … 2 n y , be random sample of size n2 form the second population with a variance σ2 2. Obviously the two samples are independent.

Null hypothesis:

H0 =σ1 2 =σ2 2 =σ2

i.e. population variances are same. In other words H0 is that the two independent estimates of the common population variance do not differ significantly.

Calculation of statistics:

Under H_0, the test statistic is

$$F_0 = \frac{S_1\square^2}{S_2\square^2}$$

$$S_1\square^2 = \frac{1}{n_1 - 1}\sum(x - \bar{x})^2 = \frac{n_1 s_1\square^2}{n_1 - 1}$$

$$S_2\square^2 = \frac{1}{n_2 - 1}\sum(y - \bar{y})^2 = \frac{n_2 s_2\square^2}{n_2 - 1}$$

It should be noted that numerator is always greater than the denominator in F-ratio

F =Larger Variance / Samller Variance

v1 = d.f for sample having larger variance v2 = d.f for sample having smaller varianc

Expected value :

$F_e = S_1{}^2 \; / \; S_2{}^2$ follows F- distribution withv1 = n1– 1 ,v2 = n2−1 d.f

The calculated value of F is compared with the table value forv1 andv2 at 5% or 1% level of significance If F0 > Fe then we reject H0. On the other hand if F0 < Fe we accept the null hypothesis and it is a inferred that both the samples have come from the population having same variance. Since F- test is based on the ratio of variances it is also known as the variance Ratio test. The ratio of two variances follows a distribution called the F distribution named after the famous statisticians R.A. Fisher.

Example:

The following data refer to yield of wheat in quintals on plots of equal area in two agricultural blocks A and B Block A was a controlled block treated in the same way as Block B expect the amount of fertilizers used

	No of plots	Mean yield	Variance
Block A	8	60	50
Block B	6	51	40

Use F test to determine whether variance of the two blocks differ significantly?

Solution:

We are given that

$n_1 = 8$ $n_2 = 6$ $\bar{x}_1 = 60$ $x_2 = 51$ $s_1^2 = 50$ $s_2^2 = 40$

Null hypothesis:

H_0: $\sigma_1^2 = \sigma_2^2$ ie there is no difference in the variances of yield of wheat.

Alternative Hypothesis:

H1: σ1 2≠ σ2 2(two tailed test) Level of significance: Let α = 0.05 Calculation of statistic

$$S_1 \square^2 = \frac{n_1 s_1 \square^2}{n_1 - 1} = \frac{8 \times 50}{7}$$

$$= 57.14$$

$$S_2 \square^2 = \frac{n_2 s_2 \square^2}{n_2 - 1} = \frac{6 \times 40}{5}$$

$$= 48$$

$$S_1 \square^2 > S_2 \square^2$$

$$F_0 = \frac{S_1 \square^2}{S_2 \square^2} = \frac{57.14}{48} = 1.19$$

Expected value:

$F_e = S_1{}^2 \ / \ S_2{}^2$ follows F- distribution with v1 = 8−1 =7 v2 = 6−1 = 5 d.f

= 4.88

Inference:

Since F0 < Fe, we accept the null hypothesis and hence infer that there is no difference in the variances of yield of wheat

5.10.ANALYSIS OF VARIANCE

The analysis of variance is a powerful statistical tool for tests of significance. The term Analysis of Variance was introduced by Prof. R.A. Fisher to deal with problems in agricultural research. The test of significance based on t-distribution is an adequate procedure only for testing the significance of the difference between two sample means. In a situation where we have three or more samples to consider at a time, an alternative procedure is needed for testing the hypothesis that all the samples are drawn from the same population, i.e., they have the same mean. For example, five fertilizers are applied to four plots each of wheat and yield of wheat on each of the plot is given. We may be interested in finding out whether the effect of these fertilizers on the yields is significantly different or in other words whether the samples have come from the same normal population. The answer to this problem is provided by the technique of analysis of variance. Thus basic purpose of the analysis of variance is to test the homogeneity of several means.

Variation is inherent in nature. The total variation in any set of numerical data is due to a number of causes which may be classified as:

(i) Assignable causes and (ii) Chance causes

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be traced separately.

5.10.1. Definition:

According to R.A. Fisher , Analysis of Variance (ANOVA) is the “ Separation of Variance ascribable to one group of causes from the variance ascribable to other group”. By this technique the total variation in the sample data is expressed as the sum of its non negative components where each of these components is a measure of the variation due to some specific independent source or factor or cause. 186 7.2 Assumptions: For the validity of the F-test in ANOVA the following assumptions are made. (i) (ii) (iii) The observations are independent Parent population from which observations are taken is normal and Various treatment and environmental effects are additive in nature.

5.10.2.One way Classification:

Let us suppose that N observations x_{ij} , i = 1, 2, ……k ; j = 1,2….ni) of a random variable X are grouped on some basis,

into k classes of sizes n_1, n_2 , ….n_k respectively $(N = \sum^{k} n_i)$ as exhibited below

	Mean	Total
x_{11} x_{12} … x_1n_1	$\overline{x_1}\,.$	$T_{1.}$
x_{21} x_{22} … x_2n_2	$\overline{x_2}\,.$	$T_{2.}$
. . .	.	.
. . .	.	.

. . .	.	.
x_{i1} x_{i2} … x_in_i	$\overline{x}_{i.}$	$T_{i.}$

. . .
. . .
. . .
	.
	.
	.
		.
		.
		.

x_{k1} x_{k2} ….. x_kn_k	$\overline{x}_{k}$.	$T_{k.}$
		G

The total variation in the observation xij can be spilit into the following two components :

(i) The variation between the classes or the variation due to different bases of classification, commonly known as treatments.

(ii)The variation within the classes i.e., the inherent variation of the random variable within the observations of a class.

The first type of variation is due to assignable causes which can be detected and controlled by human endeavour and the second type of variation due to chance causes which are beyond the control of human hand. In particular, let us consider the effect of k different rations on the yield in milk of N cows (of the same breed and stock) divided k into k classes of sizes n1, n2 , …..nk respectively. .$N = \sum_{i=1} n_i$

the sources of variation are

(i) Effect of the rations

(ii) Error due to chance causes produced by numerous causes that they are not detected and identified

5.10.3.Test Procedure:

The steps involved in carrying out the analysis are:

1) Null Hypothesis:

The first step is to set up of a null hypothesis

H0: μ1 = μ2 = … = μk

Alternative hypothesis H1:

all μi ‘ s are not equal (i = 1,2,…,k)

2) Level of significance: Let α: 0.05

3) Test statistic:

Various sum of squares are obtained as follows.

a) Find the sum of values of all the (N) items of the given data. Let this grand total represented by ‘ G’ . given data. Let this grand total represented by ‘ G’ .

Then correction factor (C.F) = $\frac{G^2}{N}$

b) Find the sum of squares of all the individual items (xij) and then the Total sum of squares (TSS) is TSS = Σ xi 2 j– C.F

c) Find the sum of squares of all the class totals (or each treatment total) Ti (i:1,2,….k) and then the sum of squares between the classes or between the treatments (SST) is SST=$\sum_{1^1=1}^{k} \frac{T_1}{n_1} - c \cdot F$

d) Where ni (i: 1,2,…..k) is the number of observations in the ith class or number of observations received by ith treatment d) Find the sum of squares within the class or sum of squares due to error (SSE) by subtraction. SSE = TSS- SST

5.10.4.Degrees of freedom (d.f):

The degrees of freedom for total sum of squares (TSS) is (N−1). The degrees of freedom for SST is (k−1) and the degrees of freedom for SSE is (N−k)

Mean sum of squares

The mean sum of squares for treatments is SST/ k -1 and mean sum of squares for error is SSE / N k

ANOVA Table

The above sum of squares together with their respective degrees of freedom and mean sum of squares will be summarized in the following table.

ANOVA Table for one-way classification:

Sources of variation	d.f	S.S	M.S.S	F ratio
Between treatments	K–1	SST	$\frac{SST}{k-1} = MST$	$\frac{MST}{MSE} = F_T$
Error	N–k	SSE	$\frac{SSE}{N-k} = MSE$	
Total	N–1			

Calculation of variance ratio:

Variance ratio of F is the ratio between greater variance and smaller variance, thus

F = Variance between the treatments / Variance within the treatment

= MST/ MSE

If variance within the treatment is more than the variance between the treatments, then numerator and denominator should be interchanged and degrees of freedom adjusted accordingly.

Critical value of F or Table value of F:

The Critical value of F or table value of F is obtained from F table for (k-1, N-k) d.f at 5% level of significance. Inference:

If calculated F value is less than table value of F, we may accept our null hypothesis H0 and say that there is no significant difference between treatments. If calculated F value is greater than table value of F, we reject our H0 and say that the difference between treatments is significant.

Example:

Three processes A, B and C are tested to see whether their outputs are equivalent. The following observations of outputs are made:

A	10	12	13	11	10	14	15	13
B	9	11	10	12	13			
C	11	10	15	14	12	13		

Carry out the analysis of variance and state your conclusion.

Solution:

To carry out the analysis of variance, we form the following tables

									Total	Squares
A	10	12	13	11	10	14	15	13	98	9604
B	9	11	10	12	13				55	3025
C	11	10	15	14	12	13			75	5625
								G = 228		

Squares:

A	100	144	169	121	100	196	225	169
B	81	121	100	144	169			
C	121	100	225	196	144	169		
					Total = 2794			

Test Procedure:

Null Hypothesis: H0: $\mu_1 = \mu_2 = \mu_3$

i.e., There is no significant difference between the three processes.

Alternative Hypothesis H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

Level of significance : Let □: 0.05

Test statistic

Correct factor (c.f) = G^2/N

$= 228^2 / 19$

$= 51984 / 19$

$= 2736$

Total sum of square (TSS) = $\sum\sum x_{ij}^2$ – C.F

$= 2794 – 2736$

$= 58$

Sum of squares due to processes = (SST)

$$= \frac{\sum_{i=1}^{3} T_{i}.^{2}}{n_{i}} - C.F$$

= 9604 / 8 + 3025 / 5 + 5625 / 6 - 2736

= (1200.5 + 605+937.5) – 2736

= 2743 – 2436

= 7

Sum of squares due to error (SSE) = TSS– SST = 58− 7 = 51

ANOVA table

Sources of variation	d.f	S.S	M.S.S	F ratio
Between Processes	3 –1 = 2	7	$\frac{7}{2} = 3.50$	$\frac{3.5}{3.19} = 1.097$
Error	16 —	51	$\frac{51}{16} = 3.19$	
Total	19 –1 = 18			

Table Value:

Table value of Fe for (2,16) d.f at 5% level of significance is 3.63

Inference:

Since calculated F0 is less than table value of Fe, we may accept our H0 and say that there is no significant difference between the three processes.

Example:

A test was given to five students taken at random from the fifth class of three schools of a town. The individual scores are

School I	9	7	6	5	8
School II	7	4	5	4	5
School III	6	5	6	7	6

Carry out the analysis of variance

Solution:

To carry out the analysis of variance, we form the following tables.

						Total	Squares
School I	9	7	6	5	8	35	1225
School II	7	4	5	4	5	25	625
School III	6	5	6	7	6	30	900
					Total	G=90	2750

Squares:

School I	81	49	36	25	64
School II	49	16	25	16	25
School III	36	25	36	49	36
				Total = 568	

Test Procedure**:**

Null Hypothesis: H_0: $\mu_1 = \mu_2 = \mu_3$ i.e., There is no significant difference between the performance of schools.

Alternative Hypothesis: H_1: $\mu_1 \neq \mu_2 \neq \mu_3$

Level of significance: Let α:0.05

Test Statistic:

Correct factor (c.f) $= \frac{G^2}{N}$

$$= \frac{90^2}{15}$$

$$= \frac{8100}{15} = 540$$

Total sum of squares (TSS) $= \Sigma\Sigma x_{ij}{\square}^{2} - C.F$

Sum of squares between schools $= \frac{\Sigma Ti^2}{n_i} - C.F$

$$= \frac{2750}{5} - 540$$

$$= 550 - 540 = 10$$

Sum of squares due to error (SSE) $= TSS - SST$

$$= 28 - 10 = 18$$

ANOVA table

Source of variation	d.f	S.S	M.S.S	F ratio
Between Schools	3-1 = 2	10	$\frac{10}{2} = 5.0$	$\frac{5}{1.5} = 3.33$
Error	12	18	$\frac{18}{12} = 1.5$	
Total	15 -1 = 14			

Table Value:

Table value of Fe for (2,12) d.f at 5% level of significance is 3.8853

Inference:

Since calculated F0 is less than table value of Fe, we may accept our H0 and say that there is no significant difference between the performance of schools

5.10.5.Two way classification:

Let us consider the case when there are two factors which may affect the variate values xij, e.g the yield of milk may be affected by difference in treatments i.e., rations as well as the difference in variety i.e., breed and stock of the cows. Let us now suppose that the N cows are divided into h different groups or classes according to their breed and stock, each group containing k cows and then let us consider the effect of k treatments (i.e., rations given at random to cows in each group) on the yield of milk.

Let the suffix i refer to the treatments (rations) and j refer to the varieties (breed of the cow), then the yields of milk xij (i:1,2, …..k; j:1,2….h) of N = h × k cows furnish the data for the comparison of the treatments (rations). The yields may be expressed as variate values in the following k × h two way table.

	Mean	Total
x_{11} x_{12} $x_{1j} \ldots x_{1h}$	$\overline{x_1}$.	$T_{1.}$
x_{21} x_{22} $x_{2j} \ldots x_{2h}$	$\overline{x_2}$.	$T_{2.}$
. . . .	. .	. .
x_{i1} x_{i2} $x_{ij} \ldots x_{ih}$	$\overline{x_i}$.	$T_{i.}$

. . . .
. . . .
. . . .
	.
	.
	.
		.
		.
		.

x_{k1} x_{k2} $x_{kj} \ldots x_{k}h$	$\overline{x_k}$.	$T_{k.}$
Mean $\overline{x}_{.1}$. $\overline{x}_{.2}$ $\ldots \overline{x}_{.j} \ldots \overline{x}_{.h}$	$\overline{x}$	
Total $T_{.1}$ $T_{.2}$ .... $T_{.j} \ldots T_{.h}$		G

The total variation in the observation xij can be split into the following three components:

(i) The variation between the treatments (rations)

(ii) The variation between the varieties (breed and stock)

(iii) The inherent variation within the observations of treatments and within the observations of varieties.

The first two types of variations are due to assignable causes which can be detected and controlled by human endevour and the third type of variation due to chance causes which are beyond the control of human hand.

Test procedure for Two- way analysis:

The steps involved in carrying out the analysis are:

Null hypothesis:

The first step is to setting up a null hypothesis H_0

Ho : µ1. = µ2. = ...µk. = µ Ho : µ . 1 = µ . 2 = ...µ . h = µ

i.e., there is no significant difference between rations (treatments) and there is no significant difference between varieties (breed and stock)

Level of significance: Let α : 0.05

Test Statistic:

Various sums of squares are obtained as follows:

a) Find the sum of values of all the N (k×h) items of the given data. Let this grand total represented by ‘ G’ Then correction factor (C.F) = G^2 / N

b) Find the sum of squares of all the individual items (x_{ij}) and then the total sum of squares (TSS)

$$\sum_{i=1}^{k} \sum_{j=1}^{k} x^2{}_{ij} - \text{C.F}$$

c) Find the sum of squares of all the treatment (rations) totals, i.e., sum of squares of row totals in the h × k two-way table. Then the sum of squares between treatments or sum of squares between rows is

SST=SSR=$\sum_{i=1}^{k} \frac{Ti\ 2}{n_1} - c \cdot F$

where h is the number of observations in each row d) Find the sum of squares of all the varieties (breed and stock) totals, in the h × k two- way table. Then the sum of squares between varieties or sum of squares between columns is$\frac{\sum_{j=1}^{k} T^2 \cdot j}{k}$ – C.F where k is the number of observations in each column.

d) Find the sum of squares due to error by subtraction: i.e., SSE = TSS– SSR – SSC

Degrees of freedom:

i) The degrees of freedom for total sum of squares is N−1 = hk−1

ii) The degrees of freedom for sum of squares between treatments is k−1

iii) The degree of freedom for sum of squares between varieties is h– 1

iv) The degrees of freedom for error sum of squares is (k−1) (h−1)

Mean sum of squares (MSS):

i) Mean sum of squares for treatments (MST) is SST / K-1

ii) Mean sum of squares for varieties (MSV) is SSV / h-1

iii) Mean sum of squares for error (MSE) is SSE / (h-1) (k-1)

ANOVA table :

The above sum of squares together with their respective degrees of freedom and mean sum of squares will be summarized in the following table.

Sources of variation	d.f	SS	MSS	F_0 - ratio
Between Treatments	k-1	SST	MST	$\frac{MST}{MSE} = F_R$
Between Varieties	h-1	SSV	MSV	$\frac{MSV}{MSE} = F_c$
Error	(h-1) (k-1)	SSE	MSE	
Total	N-1			

Critical values Fe or Table values of F:

(i) The critical value or table value of 'F' for between treatments is obtained from F table for [(k−1, (k−1) (h−1)] d.f at 5% level of significance.

(ii) (ii) The critical value or table value of Fe for between varieties is obtained from F table for [(h−1), (k−1) (h−1)] d.f at 5% level of significance.

Inference:

(i) If calculated F_0 value is less than or greater than the table value of F_e for between treatments (rows) H_0 may be accepted or rejected accordingly.

(ii) If calculated F_0 value is less than or greater than the table value of F_e for between varieties (column), H_0 may be accepted or rejected accordingly

Example:

Three varieties of coal were analysed by four chemists and the ash-content in the varieties was found to be as under

	Chemists			
Varieties	1	2	3	4
A	8	5	5	7
B	7	6	4	4
C	3	6	5	4

Carry out the analysis of variance.

Solution:

To carry out the analysis of variance we form the following tables

Chemists						
Varieties	1	2	3	4	Total	Squares
A	8	5	5	7	25	625
B	7	6	4	4	21	441
C	3	6	5	4	18	324
Total	18	17	14	15	**G = 64**	1390
Squares	324	289	196	225	1034	

Individual squares

Chemists				
Varieties	1	2	3	4
A	64	25	25	49
B	49	36	16	16
C	9	36	25	16

Test Procedure:

Null hypothesis:

$H_0 : \mu_{1.} = \mu_{2.} = \mu_{3.} = \mu$

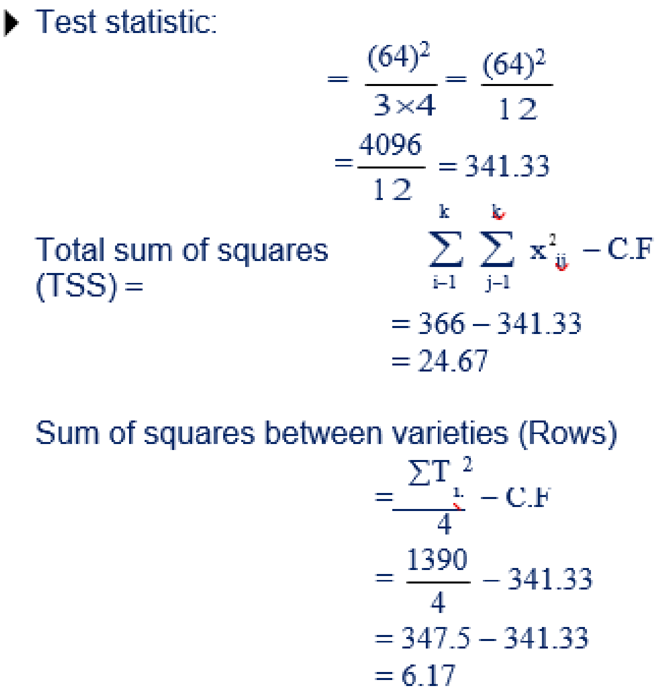
$H_0 : \mu_{.1} = \mu_{.2} = \mu_{.3} = \mu_{.4} = \mu$

(i) i.e., there is no significant difference between varieties (rows)

(ii) i.e., there is no significant difference between chemists (columns)

Level of significance :

Let α: 0.05



- Test statistic:

$$= \frac{(64)^2}{3\times4} = \frac{(64)^2}{12}$$

$$= \frac{4096}{12} = 341.33$$

Total sum of squares (TSS) = $\sum_{i-1}^{k} \sum_{j-1}^{k} x_{ij}^{2} - C.F$

$$= 366 - 341.33$$
$$= 24.67$$

Sum of squares between varieties (Rows)

$$= \frac{\sum T_{i.}^{2}}{4} - C.F$$

$$= \frac{1390}{4} - 341.33$$

$$= 347.5 - 341.33$$
$$= 6.17$$

Sum of squares between chemists (columns)

$$= \frac{\sum T_{.i}^{2}}{3} - C.F$$

$$= \frac{1034}{3} \quad 341.33$$

$$= 344.67 - 341.33$$

$$= 3.34$$

Sum of square due to error (SSE)

$$= TSS - SSR - SSC$$

$$= 24.67 - 6.17 - 3.34$$

$$= 24.67 - 9.51$$

$$= 15.16 \mid$$

ANOVA TABLE

Sources of variation	d.f	SS	MSS	F ratio
Between Varieties	3 –1 = 2	6.17	3.085	$\frac{3.085}{2.527} = 1.22$
Between Chemists	4 –1 = 3	3.34	1.113	$\frac{2.527}{1.113} = 2.27$
Error	6	15.16	2.527	
Total	12–1 = 11			

Table value :

(i) Table value of F_e for (2,6) d.f at 5% level of significance is 5.14

(ii) Table value of F_e for (6,3) d.f at 5% level of significance is 8.94

Inference:

i) Since calculated F_0 is less than table value of F_e, we may accept our H_0 for between varieties and say that there is no significant difference between varieties.

ii) Since calculated F_0 is less than the table value of F_e for chemists, we may accept our H_o and say that there is no significant difference between chemists.

Unit- V

Self – Assessment

1. Ten flower stems are chosen at random from a population and their heights are found to be (in cms) 63, 63, 66, 67, 68, 69, 70, 70, 71 and 71. Discuss whether the mean height of the population is 66 cm.

2. A random sample of size 10 from a normal population gave the following values 65, 72, 68, 74, 77, 61,63, 69, 73, 71

Test the hypothesis that population variance is 32.

3. Two random samples were drawn from two normal populations and their values are

A	66	67	75	76	82	84	88	90	92	-	-
B	64	66	74	78	82	85	87	92	93	95	97

Test whether the two populations have the same variance at 5% level of significance.

4. The standard deviations calculated from two samples of sizes 9 and 13 are 2.1 and 1.8 respectively. May the samples should be regarded as drawn from normal populations with the same standard deviation?

Ans:

1.t = 1.891 H0 is accepted

2. χ2 = 7.3156 H0 is accepted

3. F = 1.415 H0 is accepted

4. F = 1.41 H0 is accepted

Glossary:

UNIT I

Measures of Central Tendency

Mean: The arithmetic average of a set of values, calculated by summing all the values and dividing by the number of values.

Median: The middle value of a dataset when it is ordered from least to greatest. If the dataset has an even number of observations, the median is the average of the two middle numbers.

Mode: The value that appears most frequently in a dataset. A dataset may have one mode, more than one mode, or no mode at all.

Range: The difference between the maximum and minimum values in a dataset.

Variance: A measure of how much the values in a dataset differ from the mean. It is the average of the squared differences from the mean.

UNIT II

Standard Deviation: The square root of the variance, providing a measure of dispersion in the same units as the original data.

Interquartile Range (IQR): The range between the first quartile (25th percentile) and the third quartile (75th percentile). It measures the spread of the middle 50% of the data.

Skewness: A measure of the asymmetry of the data distribution. Positive skewness indicates a right-skewed distribution, while negative skewness indicates a left-skewed distribution.

Kurtosis: A measure of the "tailedness" of the data distribution. High kurtosis indicates heavy tails and a sharp peak, while low kurtosis indicates light tails and a flatter peak.

Quartiles: Values that divide a dataset into four equal parts. The first quartile (Q1) is the 25th percentile, the second quartile (Q2) is the median or 50th percentile, and the third quartile (Q3) is the 75th percentile.

Percentiles: Values below which a certain percentage of the data falls. For example, the 90th percentile is the value below which 90% of the observations may be found.

Minimum: The smallest value in the dataset.

Maximum: The largest value in the dataset.

Frequency: The number of times a value appears in a dataset.

Frequency Distribution: A summary of how often different values occur within a dataset.

Outlier: A data point that is significantly different from the other data points in the dataset. Outliers can affect the mean and standard deviation of the dataset.

UNIT III

Time Series: A sequence of data points recorded at successive, evenly spaced points in time.

Observation: A single data point within a time series.

Lag: The difference between the current time period and a past time period in a time series.

Level: The baseline value of the time series without trends or seasonality.

Trend: The long-term progression or direction of the data.

Seasonality: Regular, repeating patterns or cycles in the data occurring at fixed intervals (e.g., monthly, quarterly).

Cycle: Fluctuations in the time series that occur at irregular intervals, often related to economic or other external factors.

Noise: Random variations in the data that cannot be attributed to trend, seasonality, or cycles.

UNIT IV

Index Number: A statistical measure that shows changes in a variable or group of related variables over time, relative to a base value.

Base Period: The time period against which all other periods are compared when constructing an index number. Often set to 100.

Current Period: The time period being compared to the base period in an index number calculation.

Price Index: Measures the average change in prices over time for a fixed basket of goods and services.

Quantity Index: Measures the average change in quantities of goods and services over time.

Value Index: Measures the average change in the total value (price × quantity) of goods and services over time.

Consumer Price Index (CPI): Measures changes in the price level of a basket of consumer goods and services purchased by households.

Laspeyres Index: Uses the quantities from the base period to calculate the index.

Paasche Index: Uses the quantities from the current period to calculate the index.

Fisher Index: The geometric mean of the Laspeyres and Paasche indices.

Simple Index: An index number calculated for a single item or variable.

Composite Index: An index number calculated for a group of related items or variables.

Weighted Index: An index number where different items are given different levels of importance (weights).

UNIT V

Hypothesis: A statement or assumption about a population parameter that can be tested statistically.

Null Hypothesis ($H0$): The hypothesis that there is no effect or no difference, and it serves as the default or starting assumption.

Alternative Hypothesis ($H1$) The hypothesis that there is an effect or a difference, contrary to the null hypothesis.

One-Tailed Test: Tests the effect in one specific direction (e.g., whether a parameter is greater than a certain value).

Two-Tailed Test: Tests the effect in both directions (e.g., whether a parameter is different from a certain value, either greater or less).

Significance Level (α): The probability of rejecting the null hypothesis when it is actually true, often set at 0.05.

P-Value: The probability of obtaining test results at least as extreme as the observed results, assuming that the null hypothesis is true.

Test Statistic: A standardized value calculated from sample data during a hypothesis test.

Type I Error (α): Rejecting the null hypothesis when it is actually true (false positive).

Type II Error (β): Failing to reject the null hypothesis when it is actually false (false negative).

Power of a Test: The probability of correctly rejecting the null hypothesis when it is false, calculated as 1−β.

T-Test: Used to determine if there is a significant difference between sample and population means when the population variance is unknown and the sample size is small.

Chi-Square Test: Used to determine if there is a significant association between categorical variables.

ANOVA (Analysis of Variance): Used to compare the means of three or more samples to understand if at least one sample mean is significantly different from the others.

F-Test: Used to compare two variances to determine if they come from populations with equal variances.

Reference Books:

1 Statistical Methods" by S.P. Gupta.

2."Fundamentals of Mathematical Statistics" by S.C. Gupta and V.K. Kapoor

3.. "Introduction to the Theory of Statistics" by A.M. Mood, F.A. Graybill, and D.C. Boes

4. "Business Statistics" by J.K. Sharma

5."Basic Statistics" by B.L. Agarwal

Reference links:

https://11thstudymaterials.files.wordpress.com/2018/06/11th-statitics-volume-1-new-school-books-download-english-medium.pdf

Video reference link:

1.https://youtu.be/_D7KHkH15ds?si=FKClD5d5VyY-UzYm

2. https://youtu.be/tQinih99j_c