STATISTICS IN EDUCATION

MA [Education]

EDCN-902 C

[ENGLISH EDITION]

Directorate of Distance Education

TRIPURA UNIVERSITY

Reviewer

R P Hooda
Vice-Chancellor of Maharishi Dayanand University (MDU), Rohtak.

Authors

J.S. Chandan, (Units: 1.2, 1.3-1.4, 2.2) © J.S. Chandan, 2017

C.R. Kothari, (Units: 3.2-3.3.3, 4.2) © C.R. Kothari, 2017

Dr. Harish Kumar, (Units: 4.2.1-4.2.2, 4.3.4-4.3.6, 5.2) © Dr. Harish Kumar, 2017

Dr. (Mrs.) Vasantha R. Patri, (Unit: 5.3) © Dr. (Mrs.) Vasantha R. Patri, 2017

Vikas Publishing House, (Units: 1.0-1.1, 1.2.1-1.2.2, 1.5-1.9, 2.0-2.1, 2.3-2.8, 3.0-3.1, 3.3.4-3.3.5, 3.4-3.8, 4.0-4.1, 4.3-4.3.3, 4.4-4.8, 5.0-5.1, 5.4-5.8) © Reserved, 2017

Books are developed, printed and published on behalf of Directorate of Distance Education, Tripura University by Vikas Publishing House Pvt. Ltd.

Information contained in this book has been published by VIKAS® Publishing House Pvt. Ltd. and has been obtained by its Authors from sources believed to be reliable and are correct to the best of their knowledge. However, the Publisher and its Authors shall in no event be liable for any errors, omissions or damages arising out of use of this information and specifically disclaim any implied warranties or merchantability or fitness for any particular use.

VIKAS®

Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.
E-28, Sector-8, Noida - 201301 (UP)
Phone: 0120-4078900 • Fax: 0120-4078999
Regd. Office: 7361, Ravindra Mansion, Ram Nagar, New Delhi 110 055
• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

SYLLABI-BOOK MAPPING TABLE

Statistics in Education

Syllabi	Mapping in Book
Unit - I Meaning of Statistics: Statistis as a Tool in Educational Research. Statistical Tables, Frequency Distribution, Graphical Representation of Data. Meaning Advantages and Modes of Graphical Representation of Data.	**Unit 1:** Statistics: Nature and Scope **(Pages 3-38)**
Unit - II Measures of Central Tendency. Arithmetic Mean, Median Mode: Calculation, Interpretation and Use of Measures of Central Tendency. Measures of Variability-Meaning of the Measures of Variability, Range, Quartile Deviation, Average Deviation, Standard Deviation. When and Where to Use the Various Mesaures of Variability.	**Unit 2:** Measures of Central Tendency **(Pages 39-90)**
Unit - III Correlation and Regression. Correlation-Meaning and Types. The Calculation of the Correlation by the Product Moment Method. Linear Regression, The Regression Line in Prediction, Partial and Multiple Correlation.	**Unit 3:** Correlation and Regression **(Pages 91-130)**
Unit - IV Normal Distribution: Meaning, Significance, Characteristics of Normal Curve. Computing Percentiles and Percentile Ranks. Standard Errors of Measurement Measuring Divergence from Normality. Need and Importance of Significance of the Differene between Means and other Statistics. Null Hypothesis, Level of Confidence, One-tailed and Two-tailed tests of Significance. The Significance of the Difference between Means, Percentages and Correlation Coefficients.	**Unit 4:** Normal Distribution **(Pages 131-176)**
Unit - V Analysis of Variance, Non-parametric Tests. When to Use Parametric and Non-Parametric Test in Education. Median Test, Mann-Whitney ‘U’ Test, Chi-square Test, Rank-difference Correlation.	**Unit 5:** Analysis of Variance **(Pages 177-202)**

CONTENTS

INTRODUCTION

NOTES

Statistics is considered a mathematical science pertaining to the collection, analysis, interpretation or explanation and presentation of data. Statistical analysis is very important for taking decisions and is widely used by academic institutions, natural and social sciences departments, governments and business organizations. The word *statistics* is derived from the Latin word *status* which means a political state or government. It was originally applied in connection with kings and monarchs collecting data on their citizenry which pertained to state wealth, collection of taxes, study of population, and so on.

The subject of statistics is primarily concerned with making decisions about various disciplines of market and employment, such as stock market trends, unemployment rates in various sectors of industries, demographic shifts, interest rates and inflation rates over the years, as well as in education. Statistics is also considered a science that deals with numbers or figures describing the state of affairs of various situations with which we are generally and specifically concerned. To a layman, it often refers to a column of figures or perhaps tables, graphs and charts relating to areas, such as population, national income, expenditures, production, consumption, supply, demand, sales, imports, exports, births, deaths, accidents, and so on. Similarly, statistical records kept at universities may reflect the number of students, percentage of female and male students, number of divisions and courses in each division, number of professors, tuition received, expenditures incurred, and so on.

Hence, the subject of statistics deals primarily with numerical data gathered from surveys or collected using various statistical methods. Its objective is to summarize such data, so that the summary gives us a good indication about some characteristics of a population or phenomenon that we wish to study. To ensure that our conclusions are meaningful, it is necessary to subject our data to scientific analysis so that rational decisions can be made. Thus, the field of statistics is concerned with proper collection of data, organizing this data into manageable and presentable form, analysing and interpreting the data into conclusions for useful purposes.

This book is written in a self-instructional format and is divided into five units. Each unit begins with an Introduction to the topic followed by an outline of the Unit objectives. The content is then presented in a simple and easy-to-understand manner, and is interspersed with Check Your Progress questions to test the reader's understanding of the topic. A list of Questions and Exercises is also provided at the end of each unit, and includes short-answer as well as long-answer questions. The Summary and Key Terms section are useful tools for students and are meant for effective recapitulation of the text.

UNIT 1 STATISTICS: NATURE AND SCOPE

NOTES

Structure

1.0 INTRODUCTION

Statistics has become an integral part of our daily lives. Every day, we are confronted with some form of statistical information through newspapers, magazines and other forms of communication. Such statistical information has become highly influential in our lives. Indeed, the famous science fiction writer H.G. Wells had predicted nearly a century ago that statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write. Thus, the subject of statistics in itself, has gained considerable importance in affecting the processes of our thinking and decision-making.

In this unit, you will learn about the nature and scope of statistics. You will study statistics as a tool in educational research. You will also learn about statistical tables, frequency distribution and the graphical representation of data.

1.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Discuss the meaning, nature and scope of statistics
- Assess statistics as a tool in educational research
- Explain frequency distribution and the construction of a frequency distribution
- Describe the advantages and the modes of graphical representation of data

1.2 MEANING OF STATISTICS

In order for the quantitative and numerical data to be identified as statistics, it must possess certain identifiable characteristics. Some of these characteristics are described below.

1. **Statistics are aggregates of facts:** Single or isolated facts or figures cannot be called statistics as these cannot be compared or related to other figures within the same framework. Accordingly, there must be an aggregate of these figures. For example, if I say that I earn $30,000 per year, it would not be considered statistics. On the other hand, if I say that the average salary of a professor at our college is $30,000 per year, then this would be considered statistics since the average has been computed from many related figures such as yearly salaries of many professors. Similarly, a single birth in a hospital is not statistics, as it has no significance for analytical purposes. However, when such information about many births in the same hospital or birth information for different hospitals is collected, then this information can be compared and analysed, and thus this data would constitute statistics.
2. **Statistics, generally are not the outcome of a single cause, but are affected by multiple causes:** There are a number of forces working together that affect the facts and figures. For example, when we say that the crime rate in New York city has increased by 15 per cent over the last year, a number of factors might have affected this change. These factors may be: general level of economy such as state of economic recession, unemployment rate, extent of use of drugs, areas affected by crime, extent of legal effectiveness, social structure of the family in the area and so on. While these factors can be isolated by themselves, the effects of these factors cannot be isolated and measured individually. Similarly, a marked increase in food grain production in India may have been due to combined effect of many factors such as better seeds, more extensive use of fertilizers, mechanisation in cultivation, better institutional framework and governmental and banking support, adequate rainfall and so on. It is generally not possible to segregate and study the effect of each of these forces individually.
3. **Statistics are numerically expressed:** All statistics are stated in numerical figures which means that these are quantitative information only. Qualitative statements are not subject to accurate interpretations and hence cannot be called statistics. For example, qualitative statements such as *India is a developing country* or *Jack is very tall* would not be considered statistical statements. On the other hand, comparing per capita income of India with that of America would be considered statistical in nature. Similarly, Jack's height in numbers compared to average height in America would also be considered statistics.
4. **Statistical data is collected in a systematic manner:** The procedures for collecting data should be predetermined and well planned and such data

collection should be undertaken by trained investigators. Haphazard collection of data can lead to erroneous conclusions.

5. **Statistics are collected for a predetermined purpose:** The purpose and objective of collecting pertinent data must be clearly defined, decided upon and determined prior to data collection. This would facilitate the collection of proper and relevant data. For example, data on the heights of students would be irrelevant if considered in connection with the ability to get admission in a college, but may be relevant when considering qualities of leadership. Similarly, collective data on the prices of commodities in itself does not serve any purpose unless we know, for the purpose of comparison, the type of commodities under investigation and whether these relate to producer, distributor, wholesale or retail prices. As another example, if you are collecting data on the number of in-patients in the hospital waiting to be X-rayed, then the pre-determined purpose may be to establish the average time for the patients before X-ray and what can be done to reduce this waiting time.
6. **Statistics are enumerated or estimated according to reasonable standard of accuracy:** There are basically two ways of collecting data. One is the actual counting or measuring, which is the most accurate way. For example, the number of people attending a football game can be accurately determined by counting the number of tickets sold and redeemed at the gate. The second way of collecting data is by estimation and is used in situations where actual counting or measuring is not feasible or where it involves prohibitive costs. For example, the crowd at the football game can be estimated by visual observation or by taking samples of some segments of the crowd and then estimating the total number of people on the basis of these samples. Estimates based on samples cannot be as precise and accurate as actual counts or measurements, but these should be consistent with the degree of accuracy desired.
7. **Statistics must be placed in relation to each other:** The main objective of data collection is to facilitate a comparative or relative study of the desired characteristics of the data. In other words, the statistical data must be comparable with each other. The comparisons of facts and figures may be conducted regarding the same characteristics over a period of time from a single source or it may be from various sources at any one given time. For example, prices of different items in a store as such would not be considered statistics. However, prices of one product in different stores constitute statistical data, since these prices are comparable. Also, the changes in the price of a product in one store over a period of time would also be considered statistical data since these changes provide for comparison over a period of time. However, these comparisons must relate to the same phenomenon or subject so that likes are compared with likes and oranges are not compared with apples.

Functions of Statistics

NOTES

Statistics is no longer confined to the domain of mathematics. It has spread to most of the branches of knowledge including social sciences and behavioural sciences. One of the reasons for its phenomenal growth is the variety of different functions attributed to it. Some of the most important functions of statistics are described as follows:

1. **It condenses and summarizes voluminous data into a few presentable, understandable and precise figures:** The raw data, as is usually available, is voluminous and haphazard. It is generally not possible to draw any conclusions from the raw data as collected. Hence, it is necessary and desirable to express this data in few numerical values. For example, the average salary of a policeman is derived from a mass of data from surveys. But just one summarized figure gives us a pretty good idea about the income of police officers. Similarly, stock market prices of individual stocks and their trends are highly complex to comprehend, but a graph of price trends gives us the overall picture at a glance.
2. **It facilitates classification and comparison of data:** Arrangement of data with respect to different characteristics, facilitates comparison and interpretation. For example, data on age, height, sex and family income of college students gives us a much better picture of students when the data is categorized relative to these characteristics. Additionally, simply the statements about these figures don't convey any significant meaning. It is their comparison that helps us draw conclusions.
3. **It helps in determining functional relationships between two or more phenomenon:** Statistical techniques such as *correlational analysis* assist in establishing the degree of association between two or more independent variables. For example, the *coefficient of correlation* between literacy and employment gives us the degree of association between extent of training and industrial productivity. Similarly, correlation between average rainfall and agricultural productivity can be obtained by using such statistical tools. Some statistical methods can also be used in formulating and testing hypothesis about a certain phenomenon. For example, it can be tested whether a credit squeeze is effective in controlling prices of consumer goods or whether tenured professors are more motivated to improve their teaching than untenured professors.
4. **It helps in predicting future trends:** Statistical methods are highly useful tools in analysing the past data and predicting some future trends. For example, the sales for a particular product for the next year can be computed by knowing the sales for the same product over the previous years, the current market trends and the possible changes in the variables that affect the demand of the product.
5. **It helps the central management and the government in formulating policies:** Various governmental policies regarding import and export trade,

taxation, planning, resource allocation and so on are formulated on the basis of data regarding these elements. Many other policies are based upon statistical forecasts made by statisticians, such as policies regarding housing, employment, industrial expansion, food grain production and so on. Some of these policies would be based upon population forecasts for the future years. Also based upon the forecasts of future trends, events or demand, the central organizational management can modify their policies and plan to meet future needs. For example, the oil production in OPEC countries for the next few years would affect the operations of many energy consuming industries in America. Accordingly, these organizations must plan to meet these challenges in the future.

Limitations of Statistics

The field of statistics, though widely used in all areas of human knowledge and widely applied in a variety of disciplines such as business, economics and research, has its own limitations. Some of these limitations are:

1. **It does not deal with individual values:** As discussed earlier, statistics only deals with aggregate values. For example, the marks obtained by one student in a class does not carry any meaning in itself, unless it can be compared with a set standard or with other students in the same class or with his own marks obtained earlier.
2. **It cannot deal with qualitative characteristics:** Statistics is not applicable to qualitative characteristics such as honesty, integrity, goodness, colour, poverty, beauty and so on, since these cannot be expressed in quantitative terms. These characteristics, however, can be statistically dealt with if some quantitative values can be assigned to these with logical criterion. For example, intelligence may be compared to some degree by comparing IQs or some other scores in certain intelligence tests.
3. **Statistical conclusions are not universally true:** Since statistics is not an exact science, as is the case with natural sciences, the statistical conclusions are true only under certain assumptions. Also, the field deals extensively with the laws of probability which at best are educated guesses. For example, if we toss a coin 10 times, where the chances of a head or a tail are 1:1, we cannot say with certainty that there will be 5 heads and 5 tails. Thus the statistical laws are only approximations.
4. **Statistical interpretation requires a high degree of skill and understanding of the subject:** In order to get meaningful results, it is necessary that the data be properly and professionally collected and critically interpreted. It requires extensive training to read and analyse statistics in its proper context. It may lead to fallacious conclusions in the hands of the inexperienced.
5. **Statistics can be misused:** The famous statement that *figures don't lie but the liars can figure,* is a testimony to the misuse of statistics. Thus, inaccurate or incomplete figures, can be manipulated to get desirable

references. For example, the profits for company X being $100,000 in a given year are not necessarily inferior to profits of a company Y being $150,000, unless we know the size of the company and their total sales. Similarly, advertising slogans such as *4 out of 5 dentists recommend brand X tooth paste* gives us the impression that 80 per cent of all dentists recommend this brand. This may not be true since we don't know how big the sample is or whether the sample represents the entire population or not. Accordingly, such statistical conclusions can be highly misleading. Statements like, *statistics can prove anything* and *there are three types of lies — lies, damned lies and statistics*, perhaps, do have a profound basis.

Scope of Statistics

There is hardly any walk of life which has not been affected by statistics—ranging from a simple household to big businesses and the government. Some of the important areas where the knowledge of statistics is usefully applied are as follows:

1. **Government:** Since the beginning of organized society, the rulers and the heads of states have relied heavily on statistics in the form of collecting data on various aspects for formulating sound military and fiscal policies. This data may have involved population, taxes collected, military strength and so on. In the current structure of democratic societies, the government is, perhaps, the biggest collector of data and user of statistics. Various departments of the government collect and interpret vast amount of data and information for efficient functioning and decision-making.
2. **Economics:** Statistics are widely used in economics study and research. The subject of economics is mainly concerned with production and distribution of wealth as well as savings and investments. Some of the areas of economic interest in which statistical tools are used are as follows:
   (a) Statistical methods are extensively used in measuring and forecasting Gross National Product (GNP).
   (b) Economic stability is primarily judged by statistical studies of business cycles.
   (c) Statistical analyses of population growth, unemployment figures, rural or urban population shifts and so on influence much of the economic policy making.
   (d) Econometric models which involve application of statistical methods are used for optimum utilisation of resources available.
   (e) Financial statistics are necessary in the fields of money and banking including consumer savings and credit availability.
3. **Physical, natural and social sciences:** In physical sciences, as an example, the science of meteorology uses statistics in analysing the data gathered by satellites in predicting weather conditions. Similarly, in botany, in the natural sciences, statistics are used in evaluating the effects of temperature and other climatic conditions and types of soil on the health of plants. In the social

sciences, 'statistics are extensively used in all areas of human and social characteristics.'

4. **Statistics and research:** There is hardly any advanced research going on without the use of statistics in one form or another. Statistics are used extensively in medical, pharmaceutical and agricultural research. The effectiveness of a new drug is determined by statistical experimentation and evaluation. In agricultural research, experiments about crop yields, types of fertilizers and types of soils under different types of environments are commonly designed and analysed through statistical methods. In marketing research, statistical tools are indispensable in studying consumer behaviour, effects of various promotional strategies and so on.
5. **Other areas:** Statistics are commonly used by insurance companies, stock brokerage houses, banks, public utility companies and so on. Statistics are also immensely useful to politicians since they can predict their chances for winning through the use of sampling techniques in random selection of voter samples and studying their attitudes on issues and policies.

Statistics in Business and Management

Statistics influence the operations of business and management in many dimensions. Statistical applications include the area of production, marketing, promotion of product, financing, distribution, accounting, marketing research, manpower planning, forecasting, research and development and so on. As the organizational structure has become more complex and the market highly competitive, it has become necessary for executives to base their decisions on the basis of elaborate information systems and analysis instead of intuitive judgement. In such situations, statistics are used to analyse this vast data base for extracting relevant information. Some of the typical areas of business operations where statistics have been extensively and effectively used are as follows:

1. **Entrepreneuring:** If you are opening a new business or acquiring one, it is necessary to study the market as well as the resources from statistical point of view to ensure success of the new venture. A shrewd businessman must make a proper and scientific analysis of the past records and current market trends in order to predict the future course for business conditions. The analysis of the needs and wants of the consumers, the number of competitors in the market and their marketing strategies, availability of resources and general economic conditions and trends would all be extremely helpful to the entrepreneur. A number of new enterprises have failed either due to unreliability of data or due to faulty interpretations and conclusions.
2. **Production:** The production of any item depends upon the demand of that item and this demand must be accurately forecast using statistical techniques. Similarly, decisions as to what to produce and how much to produce are based largely upon the feedback of surveys that are analysed statistically.
3. **Marketing:** An optimum marketing strategy would require a skillful analysis of data on population, shifts in population, disposable income, competition,

social and professional status of target market, advertising, quality of sales people, easy availability of the product and other related matters. These variables and their inter-relationships must be statistically studied and analysed.

4. **Purchasing:** The purchasing department of an organization makes decisions regarding the purchase of raw materials and other supplies from different vendors. The statistical data in the cost structure would assist in formulating purchasing policies as to where to buy, when to buy, at what price to buy and how much to buy at a given time.
5. **Investment:** Statistics have been almost indispensable in making a sound investment whether it be in buying or selling of stocks and securities or real estate. The financial newspapers are full of tables and graphs analysing the prices of stocks and their movements. Based upon these statistical data, a good investor will buy when the prices are at their lowest and sell when the prices are at their highest. Similarly, buying an apartment building would require that an investor take into consideration the rent collected, rate of occupancy, any rent control laws, cost of the mortgage obtained and the age of the building before making a decision about investing in real estate.
6. **Banking:** Banks are highly affected by general economic and market conditions. Many banks have research departments which gather and analyse information not only about general economic conditions but also about businesses in which they may be directly or indirectly involved. They must be aware of money markets, inflation rates, interest rates and so on, not only in their own vicinity but also nationally and internationally. Many banks have lost money in international operations, sometimes in as simple a matter as currency fluctuations because they did not analyse the international economic trends correctly. Many banks have failed because they over-extended themselves in making loans without properly analysing the general business conditions.
7. **Quality control:** Statistics are used in quality control so extensively that even the phenomenon itself is known as *statistical quality control.* Statistical quality control (SQC) consists of using statistical methods to gather and analyse data on the determination and control of quality. This technique primarily deals with the samples taken randomly and as representative of the entire population, then these samples are analysed and inferences made concerning the characteristics of the population from which these random samples were taken. The concept is similar to testing one spoonful from a pot of stew and deciding whether it needs more salt or not. The characteristics of samples are analysed by statistical quality control and the use of other statistical techniques.
8. **Personnel:** Study of statistical data regarding wage rates, employment trends, cost of living indexes, work related accident rates, employee grievances, labour turnover rates, records of performance appraisal and so on and the proper analysis of such data assist the personnel departments in formulating the personnel policies and in the process of manpower planning.

As we have seen, statistics in one form or another, affects every business and every individual. An average individual is involved in statistics, knowingly or unknowingly, every day of his life; whether it be comparing prices during shopping or putting an extra lock on his door as a result of reading the crime rate in the newspapers. Perhaps, it is an exaggeration but basically it is true what an overenthusiastic, statistically aware business executive stated many years ago, *When the history of modern times is finally written, we shall read it as beginning with the age of steam and progressing through the age of electricity to that of statistics.*

1.2.1 Statistics as a Tool in Educational Research

A researcher needs to be familiar with the various statistical methods so as to be able to use the appropriate method in his research study. There are certain basic statistical methods, which can be classified into three groups as follows:

- Descriptive statistics
- Inferential statistics
- Measures of central tendency and dispersion

Descriptive Statistics

According to Smith, descriptive statistics is the formulation of rules and procedures where data can be placed in a useful and significant order. The foundation of applicability of descriptive statistics is the need for complete data presentation. The most important and general methods used in descriptive statistics are as follows:

- **Ratios:** This indicates the relative frequency of the various variables to one another.
- **Percentages:** Percentages (%) can be derived by multiplying a ratio with 100. It is thus a ratio representing a standard unit of 100.
- **Frequency table:** It is a means to tabulate the rate of recurrence of data. Data arranged in such a manner is known as 'distribution'. In case of a large distribution tendency, larger class intervals are used. This facilitates the researcher to acquire a more orderly system.
- **Histogram:** It is the graphical representation of a frequency distribution table. The main advantage of graphical representation of data in the form of histogram is that data can be interpreted immediately.
- **Frequency polygon:** It is used for the representation of data in the form of a polygon. In this method, a dot that represents the highest score is placed in the middle of the class interval. A frequency polygon is derived by linking these dots. An additional class is sometimes added in the end of the line with the purpose of creating an anchor.
- **Cumulative frequency curve:** The procedure of frequency involves adding frequency by starting from the bottom of the class interval, and adding class by class. This facilitates the representation of the number of persons that

perform below the class interval. The researcher can derive a curve from the cumulative frequency tables with the purpose of reflecting data in a graphical manner.

NOTES

Inferential Statistics

Inferential statistics enable researchers to explore unknown data. Researchers can make deductions or statements using inferential statistics with regard to the broad population from which samples of known data has been drawn. These methods are called 'inferential or inductive statistics'. These methods include the following common techniques:

- **Estimation:** It is the calculated approximation of a result, which is usable, even if the input data may be incomplete or uncertain. It involves deriving the approximate calculation of a quantity or a degree or worth. For example, drawing an estimate of cost of a project or deriving a rough idea of how long the project would take.
- **Prediction:** It is a statement or claim that a particular event will surely occur in future. It is based on observation, experience and scientific reasoning of what will happen in given circumstances or situations.
- **Hypothesis testing:** Hypothesis is a proposed explanation, whose validity can be tested. Hypothesis testing attempts to validate or disprove preconceived ideas. In creating hypothesis, one thinks of a possible explanation for a remarked behaviour. The hypothesis dictates the data selected to be analysed for further interpretations.

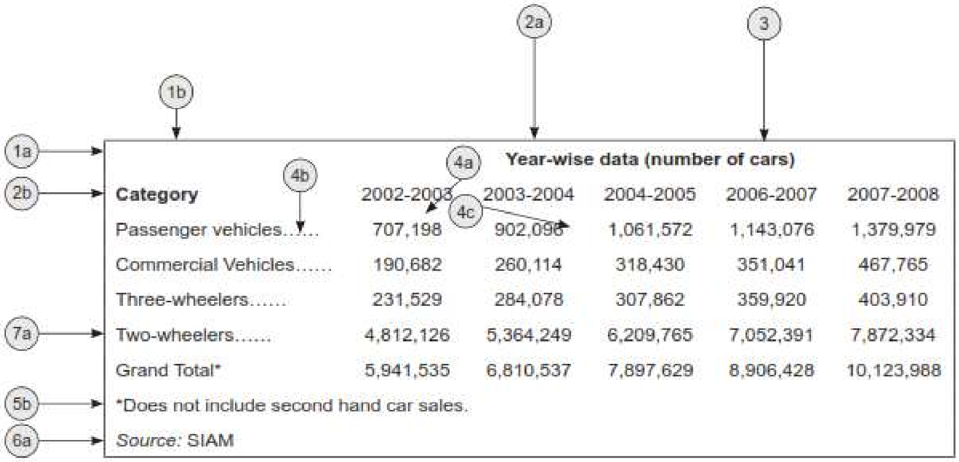
There are also two chief statistical methods based on the tendency of data to cluster or scatter. These methods, known as measures of central tendency and measures of dispersion, have been discussed in the next sub-section.

1.2.2 Statistical Tables

Most research studies involve some form of numerical data, and even though one can discuss this in text, it is best represented in tabular form. The advantage of doing this is that statistical tables present the data in a concise and numeral form, which makes quantitative analysis and comparisons easier. Tables formulated could be general tables following a statistical format for a particular kind of analysis. These are best put in the appendix, as they are complex and detailed in nature. The other kind is simple summary tables, which only contain limited information and yet, are, essentially critical to the report text.

The mechanics of creating a summary table are very simple and are illustrated below with an example (Table 1.1). The illustration has been labelled with numbers which relate to the relevant section.

Table 1.1 *Automobile Domestic Sales Trends*


2a
3
1b
1a
2b
4b
4a
4c
7a
5b
6a

	Year-wise data (number of cars)				
Category	2002-2003	2003-2004	2004-2005	2006-2007	2007-2008
Passenger vehicles......	707,198	902,096	1,061,572	1,143,076	1,379,979
Commercial Vehicles......	190,682	260,114	318,430	351,041	467,765
Three-wheelers......	231,529	284,078	307,862	359,920	403,910
Two-wheelers......	4,812,126	5,364,249	6,209,765	7,052,391	7,872,334
Grand Total*	5,941,535	6,810,537	7,897,629	8,906,428	10,123,988

*Does not include second hand car sales.

Source: SIAM

Table identification details: The table must have a title (1a) and an identification number (1b). The table title should be short and usually would not include any verbs or articles. It only refers to the population or parameter being studied. The title should be briefly yet clearly descriptive of the information provided. The numbering of tables is usually in a series and generally one makes use of Arabic numbers to identify them.

Data arrays: The arrangement of data in a table is usually done in an ascending manner. This could either be in terms of time, as shown in Table 1.1 (column-wise) or according to sectors or categories (row-wise) or locations, e.g., north, south, east, west and central. Sometimes, when the data is voluminous, it is recommended that one goes alphabetically, e.g., country or state data. Sometimes there may be subcategories to the main categories, for example, under the total sales data—a columnwise component of the revenue statement—there could be subcategories of department store, chemists and druggists, mass merchandisers and others. Then these have to be displayed under the sales data head, after giving a tab command as follows:

Total sales

Mass market

Department store

Drug stores

Others (including *paan beedi* outlets)

Measurement unit: The unit in which the parameter or information is presented should be clearly mentioned.

Spaces, leaders and rulings (SLR): For limited data, the table need not be divided using grid lines or rulings, simple white spaces add to the clarity of information presented and processed. In case the number of parameters are too many and the data seems to be bulky to be simply separated by space, it is advisable to use vertical ruling. Horizontal lines are drawn to separate the headings from the main data, as can be seen in Table 1.1. When there are a number of subheadings as in the sales

NOTES

data example, one may consider using leaders (.......) to assist the eye movement in absorbing and processing the information.

NOTES

Total sales

Mass market.........

Department store.........

Drug stores.........

Others (including *paan beedi* outlets).........

Assumptions, details and comments: Any clarification or assumption made, or a special definition required to understand the data, or formula used to arrive at a particular figure, e.g., total market sale or total market size can be given after the main tabled data in the form of footnotes.

Data sources: In case the information documented and tabled is secondary in nature, complete reference of the source must be cited after the footnote, if any.

Special mention: In case some figure or information is significant and the reader should pay special attention to it, the number or figure can be bold or can be highlighted to increase focus.

CHECK YOUR PROGRESS

1. What is descriptive statistics?
2. What does hypothesis testing attempt to validate?
3. What are the table identification details?

1.3 FREQUENCY DISTRIBUTION

Statistical data can be organized into a frequency distribution which simply lists the value of the variable and frequency of its occurrence in a tabular form. A frequency distribution can be defined as the list of all the values obtained in the data and the corresponding frequency with which these values occur in the data.

The frequency distribution can either be ungrouped or grouped. When the number of values of the variable is small, then we can construct an ungrouped frequency distribution which is simply listing the frequency of occurrence against the value of the given variable. As an example, let us assume that 20 families were surveyed to find out how many children each family had. The raw data obtained from the survey is as follows:

0, 2, 3, 1, 1, 3, 4, 2, 0, 3, 4, 2, 2, 1, 0, 4, 1, 2, 2, 3

This data can be classified into an ungrouped frequency distribution. The number of children becomes our variable (*X*) for which we can list the frequency of occurrence (*f*) in a tabular form as follows:

Table 1.2 *Frequency of Occurrence*

Number of Children (*X*)	Frequency (*f*)
0	3
1	4
2	6
3	4
4	3
	Total = 20

The above table is also known as discrete frequency distribution where the variable has discrete numerical values.

However, when the data set is very large, it becomes necessary to condense the data into a suitable number of groups or classes of the variable values and then assign the combined frequencies of these values into their respective classes. As an example, let us assume that 100 employees in a factory were surveyed to find out their ages. The youngest person was 20 years of age and the oldest was 50 years old. We can construct a grouped frequency distribution for this data so that instead of listing frequency by every year of age, we can list frequency according to an age group. Also, since age is a continuous variable, a frequency distribution would be as follows:

Table 1.3 *Frequency Distribution*

Age Group (Years)	Frequency
20 to less than 25	5
25 ” ” ” 30	15
30 ” ” ” 35	25
35 ” ” ” 40	30
40 ” ” ” 45	15
45 ” ” ” 50	10
	Total = 100

In this example, all persons between 20 years (including 20 years old) and 25 years (but not including 25 years old) would be grouped into the first class and so on. The interval of 20 to less than 25 is known as class interval (CI). A single representation of a class interval would be the midpoint (or average) of that class interval. The midpoint is also known as the class-mark.

Constructing a Frequency Distribution

The number of groups and the size of class interval are more or less arbitrary in nature within the general guidelines established for constructing a frequency distribution. The following guidelines for such a construction may be considered:

(i) The classes should be clearly defined and each of the observations should be included in only one of the class intervals. This means that the intervals should be chosen in such a manner that one score cannot belong to more than one class interval, so that there is no overlapping of class intervals.

(ii) The number of classes should neither be too large nor too small. Normally, between 6 and 15 classes are considered to be adequate. Fewer class intervals would mean a greater class interval width with consequent loss of accuracy. Too many class intervals result in a greater complexity.

(iii) All intervals should be of the same width. This is preferred for easy computations. A suitable class width can be obtained by knowing the range of data (which is the absolute difference between the highest value and the lowest value in the data) and the number of classes which are predetermined, so that:

$$\text{The width of the interval} = \frac{\text{Range}}{\text{Number of classes}}$$

In the case of ages of factory workers where the youngest worker was 20 years old and the oldest was 50 years old, the range would be 50–20 = 30. If we decide to make 10 groups then the width of each class would be:

30/10 = 3

Similarly, if we decide to make 6 classes instead of 10, then the width of each class interval would be:

30/6 = 5

(iv) Open-ended cases where there is no lower limit of the first group or no upper limit of the last group should be avoided since this creates difficulty in analysis and interpretation. (The lower and upper values of a class interval are known as lower and upper limits.)

(v) Intervals should be continuous throughout the distribution. For example, in the case of factory workers, we could group them in groups of 20 to 24 years, then 25 to 29 years, and so on, but it would be highly misleading because it does not accurately represent the person who is between 24 and 25 years or between 29 and 30 years, and so on. Accordingly, it is more representative to group them as: 20 years to less than 25 years, 25 years to less than 30 years. In this way, everybody who is 20 years and a fraction less than 25 years is included in the first category and the person who is exactly 25 years and above but a fraction less than 30 years would be included in the second category, and so on. This is especially important for continuous distributions.

(vi) The lower limits of class intervals should be simple multiples of the interval width. This is primarily for the purpose of simplicity in construction and interpretation. In our example of 20 years but less than 25 years, 25 years but less than 30 years, and 30 years but less than 35 years, the lower limit values for each class are simple multiples of the class width which is 5.

Example 1.1: A sample of 30 persons showed their ages (in years) as:

20, 18, 25, 68, 23, 25, 16, 22, 29, 37,
35, 49, 42, 65, 37, 42, 63, 65, 49, 42,
53, 48, 65, 72, 69, 57, 48, 39, 58, 67.

Construct a frequency distribution for this data.

Solution:

Follow the steps as given below:

1. Find the range of the data by subtracting the lowest age from the highest age. The lowest value is 16 and the highest value is 72. Hence, the range is 72–16 = 56.
2. Assume that we shall have 6 classes, since the number of values is not too large. Now we divide the range of 56 by 6 to get the width of the class interval. The width is 56/6 = 9.33. For the sake of convenience, assume the width to be 10 and start the first class boundary with 15 so that the intervals would be 15 and upto 25, 25 and upto 35, and so on.
3. Combine all the frequencies that belong to each class interval and assign this total frequency to the corresponding class interval as follows:

Table 1.4 *Class Intervals*

Class Interval (Years)	Tally	Frequency (f)
15 to less than 25	𝍸	5
25 to less than 35	|||	3
35 to less than 45	𝍸 ||	7
45 to less than 55	𝍸	5
55 to less than 65	|||	3
65 to less than 75	𝍸 ||	7
	Total = 30	

Discrete list conversion to a continuous list

In statistics, calculations are performed by arranging the large raw (ungrouped) data set into grouped data and are represented in tabular form called frequency distribution table. The data to be grouped must be homogenous and comparable. The frequency distribution table gives the size and the number of class intervals. The range of each class is defined by the class boundaries.

The variables constitute either a discrete list or a continuous list. A variable is considered as continuous when it can assume an infinite number of real values and it is considered discrete when it is the finite number of real values. Examples of a continuous variable are distance, age, temperature and height measurements, whereas the examples of a discrete variable are the scores given by the experts or the judgement team for competition examination, basket ball match, cricket match, etc.

For a discrete list of data, the range can be defined as 0 – 4, 5 – 9, 10 – 14, and so on. Similarly, the range of data for a continuous list can be defined as 10 – 20, 20 – 30, 30 – 40, and so on.

In a class interval, the endpoints define the lowest and highest values that a variable can take. In this example, if we consider the data set for age then the class intervals are 0 to 4 years, 5 to 9 years, 10 to 14 years, and 14 years and above. For a discrete variable, the end points, of the first class interval are 0 and 4 but for a continuous variable it will be 0 and 4.999. In this way, the discrete variables can be converted to continuous variables and vice versa.

Conversion of ungrouped list into grouped list

The data collected first-hand for any statistical evaluation is considered as raw or ungrouped data as it is not meaningful and does not present a clear picture. It is then arranged in the ascending or the descending order in a tabular form called array. The following example will make the concept more clear.

Example 1.2: The following table shows the daily wages (in ₹) of 40 workers. Convert the ungrouped data into grouped data and also prepare a discrete frequency table with tally marks.

Ungrouped Data

90	85	50	70	55	86	60	75	80	65
75	78	86	80	60	90	55	95	65	85
55	70	60	85	80	95	90	75	60	86
60	95	85	70	65	55	86	90	80	78

Solution:

After arranging this into grouped data, we get the following table:

95	95	95	90	90	90	90	86	86	86
86	85	85	85	85	80	80	80	80	78
78	75	75	75	70	70	70	65	65	65
60	60	60	60	60	55	55	55	55	50

The discrete frequency distribution of daily wages with tally marks:

Table 1.5 Descrete Frequency Distribution

Daily Wages	Tally Marks	Frequency
95	|||	3
90	||||	4
86	||||	4
85	||||	4
80	||||	4
78	||	2
75	|||	3
70	|||	3
65	|||	3
60	𝍸	5
55	||||	4
50	|	1
	Total	40

Class intervals of unequal width

From the data given in Example 1.2, a table showing class intervals of unequal with is drawn.

Table 1.6 Class Invervals of Unequal

Daily Wages	Tally Marks	Frequency
50–55	𝍸	5
55–60	𝍸	5
60–65	|||	3
65–70	|||	3
70–75	|||	3
75–78	||	2
78–80	||||	4
80–85	||||	4
85–86	||||	4
86–90	||||	4
90–95	|||	3
	Total	40

Cumulative Frequency

While the frequency distribution table tells us the number of units in each class interval, it does not tell us directly the total number of units that lie below or above the specified values of class intervals. This can be determined from a cumulative frequency distribution. When the interest of the investigator focusses on the number of items below a specified value, then this specified value is the upper limit of the class interval. It is known as less than cumulative frequency distribution. Similarly, when the interest lies in finding the number of cases above a specified value, then this value is taken as the lower limit of the specified class interval and is known as more than cumulative frequency distribution. The cumulative frequency simply means summing up the consecutive frequencies as follows (taking the example of ages of 30 workers):

Table 1.7 *Cumulative Frequency Distribution*

Class Interval (Years)	(f)	Cumulative Frequency (Less Than)
15 and upto 25	5	5 (less than 25)
25 and upto 35	3	8 (less than 35)
35 and upto 45	7	15 (less than 45)
45 and upto 55	5	20 (less than 55)
55 and upto 65	3	23 (less than 65)
65 and upto 75	7	30 (less than 75)

Similarly, the following is the greater than cumulative frequency distribution:

Table 1.8 *Greater than Cumulative Frequency Distribution*

Class Interval (Years)	(f)	Cumulative Frequency (Greater Than)
15 and upto 25	5	30 (greater than 15)
25 and upto 35	3	25 (greater than 25)
35 and upto 45	7	22 (greater than 35)
45 and upto 55	5	15 (greater than 45)
55 and upto 65	3	10 (greater than 55)
65 and upto 75	7	7 (greater than 65)

In the preceding greater than cumulative frequency distribution; 30 persons are older than 15, 25 are older than 25, and so on.

Percentage Frequency

The frequency distribution, as defined earlier, is a summary table in which the original data is condensed into groups and their frequencies. But if a researcher would like to know the proportion or the percentage of cases in each group, instead of simply the number of cases, he can do so by constructing a relative frequency distribution table. The relative frequency distribution can be formed by dividing the frequency in each class of the frequency distribution by the total number of observations. It can be converted into a percentage frequency distribution by simply multiplying each relative frequency by 100.

The relative frequencies are particularly helpful when comparing two or more frequency distributions in which the number of cases under investigation is not equal. The percentage distributions make such a comparison more meaningful, since percentages are relative frequencies and hence the total number in the sample or population under consideration becomes irrelevant. Carrying on with the earlier example:

Table 1.9 *Percentage Frequency*

Class Interval (Years)	(*f*)	Rel. Freq.	% Freq.
15 and upto 25	5	5/30	16.7
25 and upto 35	3	3/30	10.0
35 and upto 45	7	7/30	23.3
45 and upto 55	5	5/30	16.7
55 and upto 65	3	3/30	10.0
65 and upto 75	7	7/30	23.3
Total	30		100.0

Cumulative relative frequency distribution

It is often useful to know the proportion or the percentage of cases falling below a particular score point or falling above a particular score point. A less than cumulative relative frequency distribution shows the proportion of cases lying below the upper limit of specific class interval. Similarly, a greater than cumulative frequency distribution shows the proportion of cases above the lower limit of a specified class interval. We can develop the cumulative relative frequency distributions from the less than and greater than cumulative frequency distributions constructed earlier. By following the earlier example, we get:

Table 1.10 *Cumulative Relative Frequency Distribution*

Class Interval (Years)	Cum. Freq. (Less Than)	Cum. Rel. Freq. (Less Than)
Less than 25	5	5/30 or 16.7%
Less than 35	8	8/30 or 26.7%
Less than 45	15	15/30 or 50.0%
Less than 55	20	20/30 or 66.7%
Less than 65	23	23/30 or 76.7%
Less than 75	30	30/30 or 100%

In the above example, 5 out of 30 or 16.7 per cent of the persons are below 25 years of age. Similarly, 15 out of 30 or 50 per cent of the persons are below 45 years of age and so on. Similarly, we can construct a greater than cumulative relative frequency distribution as follows for the same example:

Table 1.11 *Greater than Cumulative Relative Frequency Distribution*

Class Interval (Years)	Cum. Freq. (Greater Than)	Cum. Rel. Freq. (Greater Than)
15 and above	30	30/30 or 100%
25 and above	25	25/30 or 83.3%
35 and above	22	22/30 or 73.3%
45 and above	15	15/30 or 50.0%
55 and above	10	10/30 or 33.3%
65 and above	7	7/30 or 23.3%

In this example, 100 per cent of the persons are above 15 years of age, 73.3 per cent are above 35 years of age and so on. (It should be noted that the less than cumulative frequency distribution is summed up from top downwards and the greater than cumulative frequency distribution is summed from bottom upwards).

CHECK YOUR PROGRESS

4. Define frequency distribution.
5. What is raw data?

1.4 GRAPHICAL REPRESENTATION OF DATA: ADVANTAGES AND MODES

Stem and leaf display is another form of presentation of the data distribution. It allows us to condense data but still retain the individuality of the data. The idea is based on an analogy to plants. The leaves in the stem and leaf diagram are the last digit in each number of observed data. The first digit or digits, as the case may be, are the stem. All the values in the stem are listed in order in a column and a vertical line is drawn beside them and then all the corresponding leaf values are recorded for each stem in a row to the right of the vertical line.

In our example of the ages of 30 workers, the stem and leaf diagram would be displayed as follows:

First, let us put the original data in an ascending order.

16, 18, 20, 22, 23, 25, 25, 29, 35, 37,
37, 39, 42, 42, 42, 48, 48, 49, 49, 53,
57, 58, 63, 65, 65, 65, 67, 68, 69, 72.

Now the stem and leaf diagram:

Table 1.12 *Stem and Leaf*

Stem	Leaves	(*f*)
1	6 8	2
2	0 2 3 5 5 9	6
3	5 7 7 9	4
4	2 2 2 8 8 9 9	7
5	3 7 8	3
6	3 5 5 5 7 8 9	7
7	2	1
		Total = 30

Summing up the frequencies provides a check on whether all the data has been included or not.

Diagrammatic and Graphic Presentation

The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data. These important features are more simply presented in the form of graphs. Also, diagrams facilitate comparisons between two or more sets of data.

The diagrams should be clear and easy to read and understand. Too much information should not be represented through the same diagram; otherwise, it may become cumbersome and confusing. Each diagram should include a brief and self-explanatory title dealing with the subject matter. The scale of the presentation should be chosen in such a way that the resulting diagram is of appropriate size. The intervals on the vertical as well as the horizontal axis should be of equal size; otherwise, distortions would occur.

Diagrams are more suitable to illustrate discrete data, while continuous data is better represented by graphs. The following are the diagrammatic and graphic representation methods that are commonly used.

Diagrammatic Representation

Diagrammatic representation can be of the following types:

(i) Bar diagram

(ii) Pie chart

(iii) Pictogram

(i) Bar diagram: Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram. Additionally, the bars should be equally spaced.

Example 1.3: Suppose that the following were the gross revenues (in $100,000.00) for a company *XYZ* for the years 1989, 1990 and 1991.

Year	Revenue
1989	110
1990	95
1991	65

Construct a bar diagram for this data.

Solution:

The bar diagram for this data can be constructed as follows with the revenues represented on the vertical axis and the years represented on the horizontal axis.

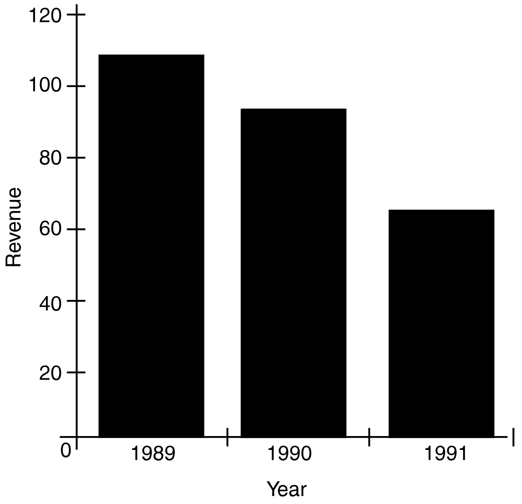


Fig. 1.1 *Bar Diagram*

The bars drawn can be subdivided into components depending upon the type of information to be shown in the diagram. This will be clear by the following example in which we are presenting three components in a bar.

Example 1.4: Construct a subdivided bar chart for the three types of expenditures in dollars for a family of four for the years 1988, 1989, 1990 and 1991 as given as follows:

Year	Food	Education	Other	Total
1988	3000	2000	3000	8000
1989	3500	3000	4000	10500
1990	4000	3500	5000	12500
1991	5000	5000	6000	16000

Solution:

The subdivided bar chart would be as follows:

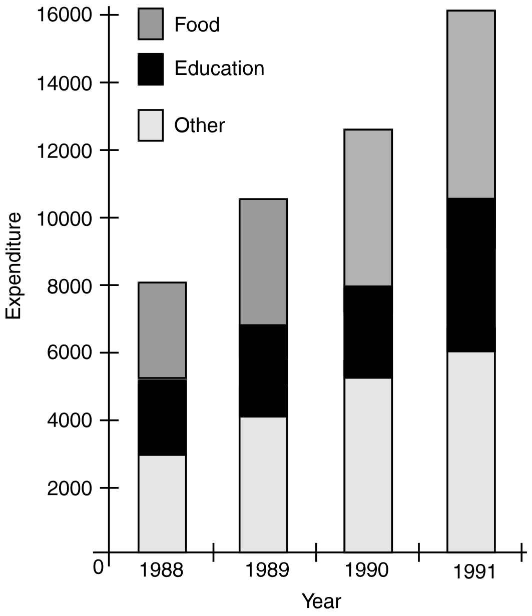


Fig. 1.2 *Sub-Divided Bar Diagram*

(ii) Pie chart: This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

Example 1.5: The following figures relate to the cost of the construction of a house. The various components of cost that go into it are represented as percentages of the total cost.

Item	% Expenditure
Labour	25
Cement, Bricks	30
Steel	15
Timber, Glass	20
Miscellaneous	10

Construct a pie chart for the above data.

Solution:

The pie chart for this data is presented as follows:

NOTES

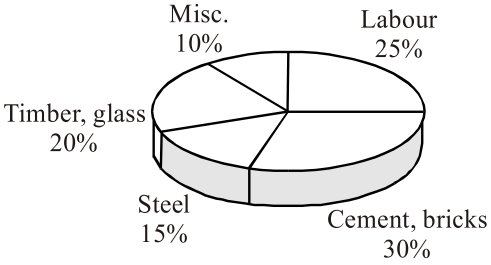


Fig. 1.3 *Pie Chart*

Pie charts are very useful for comparison purposes, especially when there are only a few components. If there are too many components, it may become confusing to differentiate the relative values in the pie.

(iii) Pictogram: Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.

News magazines are very fond of presenting data in this form. For example, in comparing the strength of the armed forces of USA and Russia, they will simply make sketches of soldiers where each sketch may represent 100,000 soldiers. Similar comparison for missiles and tanks is also done.

Graphic Representation

Graphic representation can be of the following types:

(i) Histogram

(ii) Frequency polygon

(iii) Cumulative frequency curve (Ogive)

Each of these is briefly explained and illustrated.

(i) Histogram: A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.

The word histogram is derived from the Greek word *histos* which means 'anything set upright' and *gramma* which means 'drawing, record, writing'. It is considered as the most important basic tool of statistical quality control process.

In this type of representation, the given data are plotted in the form of a series of rectangles. Class intervals are marked along the *X*-axis and the frequencies along the *Y*-axis according to a suitable scale. Unlike the bar chart, which is one-dimensional, meaning that only the length of the bar is important and not the width, a histogram is two-dimensional in which both the length and the width are important. A histogram is constructed from a frequency distribution of a grouped data where the height of the rectangle is proportional to the respective frequency and the width represents the class interval. Each rectangle is joined with the other and any blank spaces between the rectangles would mean that the category is empty and there are no values in that class interval.

As an example, let us construct a histogram for our example of ages of 30 workers. For convenience sake, we will present the frequency distribution along with the mid-point of each interval, where the mid-point is simply the average of the values of the lower and upper boundary of each class interval. The frequency distribution table is shown as follows:

Class Interval (Years)	Mid-point	(f)
15 and upto 25	20	5
25 and upto 35	30	3
35 and upto 45	40	7
45 and upto 55	50	5
55 and upto 65	60	3
65 and upto 75	70	7

The histogram of this data would be shown as follows:

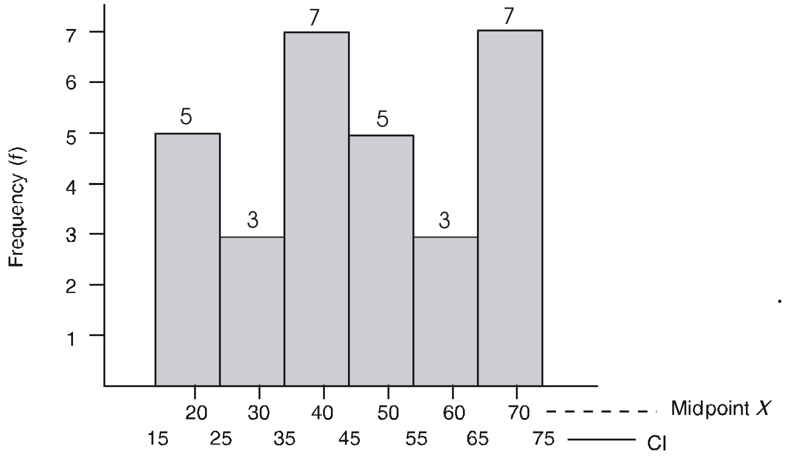


Fig. 1.4 *Histo Gram*

(ii) Frequency polygon: A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies and these plotted points are joined together by straight lines. Since the frequencies generally do not start at zero or end at zero, this diagram as such would not touch the horizontal axis. However, since the area under the entire curve is the same as that of a histogram which is 100 per cent of the data presented, the curve can be enclosed so that the starting point is joined with a fictitious preceding point whose value is zero, so that the start of the curve is at horizontal axis and the last point is joined with a fictitious succeeding point whose value is also zero, so that the curve ends at the horizontal axis. This enclosed diagram is known as the frequency polygon.

We can construct the frequency polygon from the preceding table as follows:

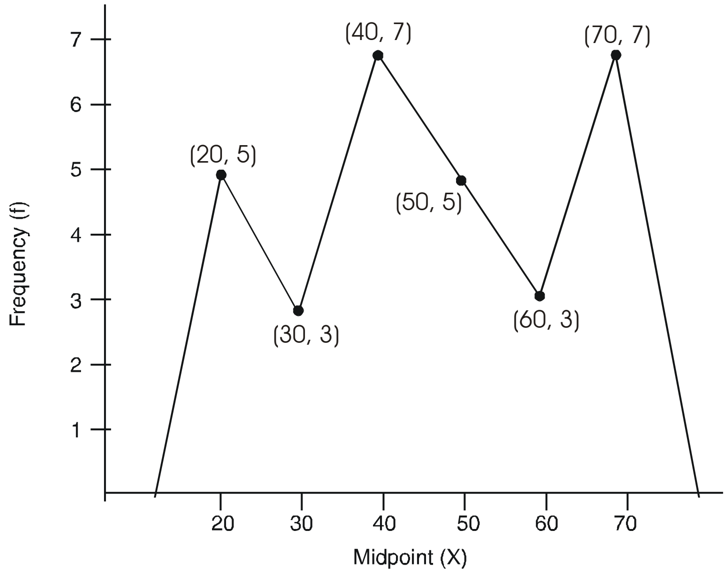


***Fig. 1.5** Frequency Polygon*

(iii) Cumulative frequency curve (Ogive): The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.

Class Interval (Years)	Mid-point	(*f*)	Cum. Freq. (Less Than)	Cum. Freq. (Greater Than)
15 and upto 25	20	5	5 (less than 25)	30 (more than 15)
25 and upto 35	30	3	8 (less than 35)	25 (more than 25)
35 and upto 45	40	7	15 (less than 45)	22 (more than 35)
45 and upto 55	50	5	20 (less than 55)	15 (more than 45)
55 and upto 65	60	3	23 (less than 65)	10 (more than 55)
65 and upto 75	70	7	30 (less than 75)	7 (more than 65)

(a) Less than Ogive: In this case, less than cumulative frequencies are plotted against upper boundaries of their respective class intervals.

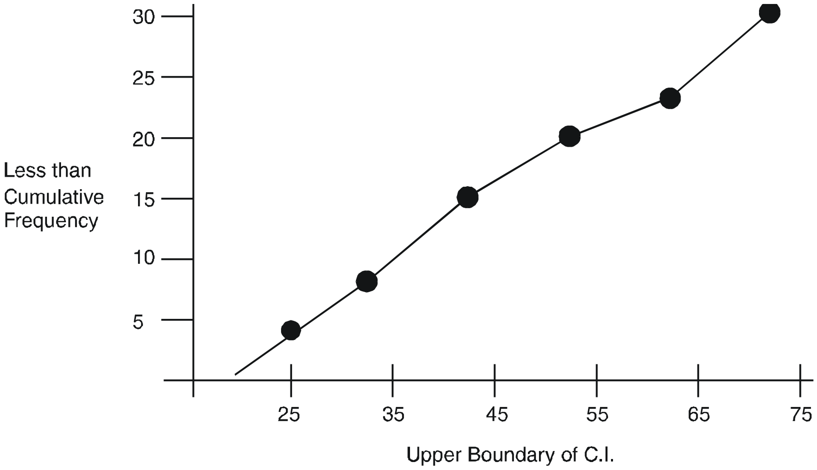


Fig. 1.6 *Less Than Ogive*

(b) Greater than Ogive: In this case, greater than cumulative frequencies are plotted against the lower boundaries of their respective class intervals.

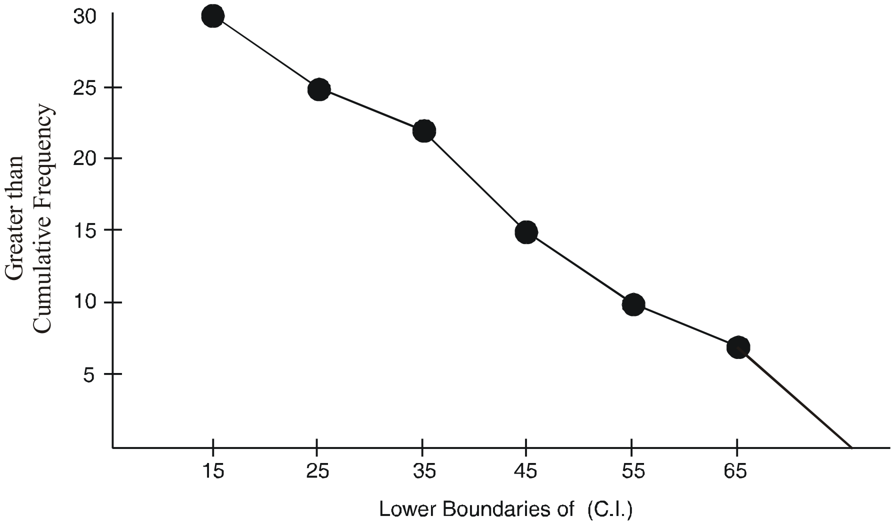


Fig. 1.7 *More Than Ogive*

These ogives can be used for comparison purposes. Several ogives can be drawn on the same grid, preferably with different colours for easier visualization and differentiation.

Although, diagrams and graphs are a powerful and effective media for presenting statistical data, they can only represent a limited amount of information and they are not of much help when intensive analysis of data is required.

Solved Problems

Problem 1: Standard tests were administered to 30 students to determine their IQ scores. These scores are recorded in the following table.

120 115 118 132 135 125 122 140 137 127

129 130 116 119 132 127 133 126 120 125

130 134 135 127 116 115 125 130 142 140

(a) Arrange this data into an ordered array.

(b) Construct a grouped frequency distribution with suitable class intervals.

(c) Compute for this data:

– Cumulative frequency (<)

– Cumulative frequency (>)

(d) Compute:

– Relative frequency

– Cumulative relative frequency (<)

– Cumulative relative frequency (>)

(e) Construct for this data:

– A histogram

– A frequency polygon

– Cumulative relative ogive (<)

– Cumulative relative ogive (>)

Solution:

(a) The ordered array for this data is as follows:

115 115 116 116 118 119 120 120 122 125

125 125 126 127 127 127 129 130 130 132

132 132 133 134 135 135 137 140 140 142

(b) Let there be 6 groupings, so that the size of the class interval be 5. The frequency distribution is shown as follows:

Class Interval (CI)	Frequency (*f*)
115 to less than 120	6
120 ” ” ” 125	3
125 ” ” ” 130	8
130 ” ” ” 135	7
135 ” ” ” 140	3
140 ” ” ” 145	3

(c) The required elements are computed in the following table.

Class Interval	(*f*)	Cum. Freq.(<)	Cum. Freq. (>)
115–120	6	6 (less than 120)	30 (more than 115)
120–125	3	9 (less than 125)	24 (more than 120)
125–130	8	17 (less than 130)	21 (more than 125)
130–135	7	24 (less than 135)	13 (more than 130)
135–140	3	27 (less than 140)	6 (more than 135)
140–145	3	30 (less than 145)	3 (more than 140)

(d) The computed values of relative frequency, cumulative relative frequency (<) and cumulative relative frequency (>) are shown in the following table:

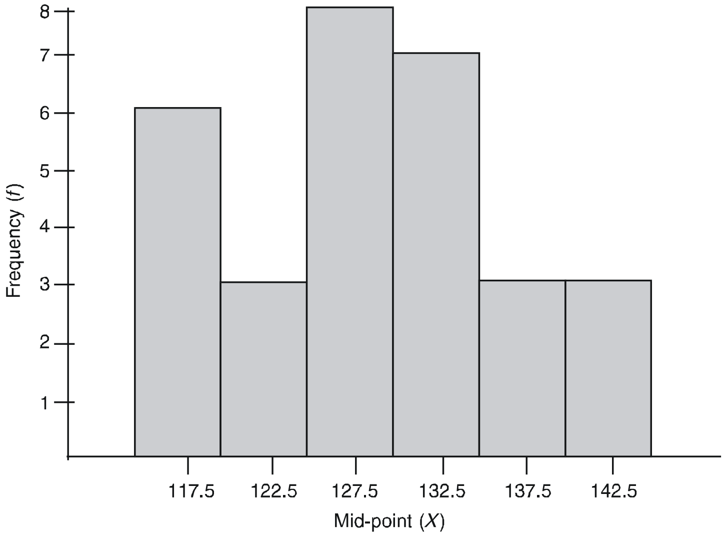
Class Interval	(*f*)	Rel. Freq.	Cum. Rel. Freq. (<)	Cum. Rel. Freq. (>)
115 and upto 120	6	6/30 or 20%	6/30 or 20% (<120)	30/30 or 100% >115)
120 and upto 125	3	3/30 or 10%	9/30 or 30% <125)	24/30 or 80% (>120)
125 and upto 130	8	8/30 or 26.7%	17/30 or 56.7% (<130)	21/30 or 70% (>125)
130 and upto 135	7	7/30 or 23.3%	24/30 or 80% (<135)	13/30 or 43.3% (>130)
135 and upto 140	3	3/30 or 10%	27/30 or 90% (<140)	6/30 or 20% (>135)
140 and upto 145	3	13/30 or 10%	30/30 or 100% (<145)	3/ 30 or 10% (>140)
Total = 30				

(e) Before we construct the histogram and other diagrams, let us first determine the midpoint (*X*) of each class interval.

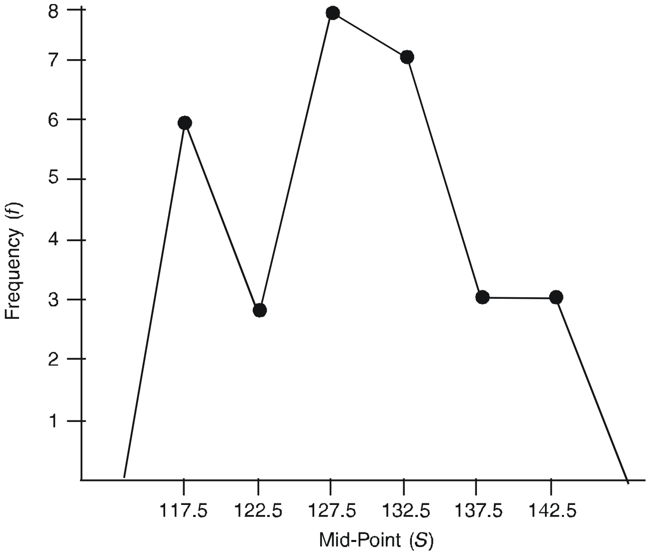
Class Interval	(*f*)	Mid-point (*X*)
115–120	6	117.5
120–125	3	122.5
125–130	8	127.5
130–135	7	132.5
135–140	3	137.5
140–145	3	142.5

NOTES

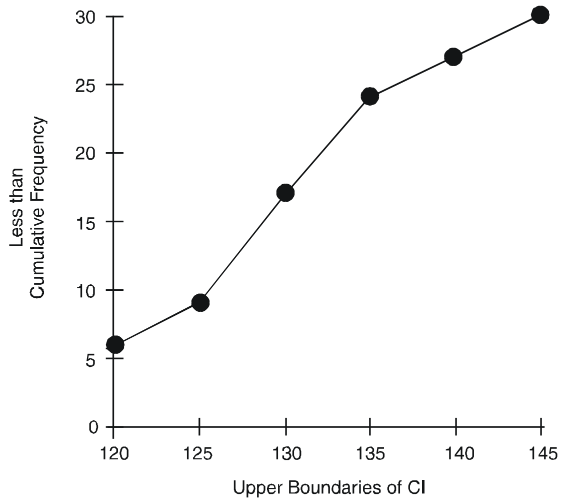
A histogram



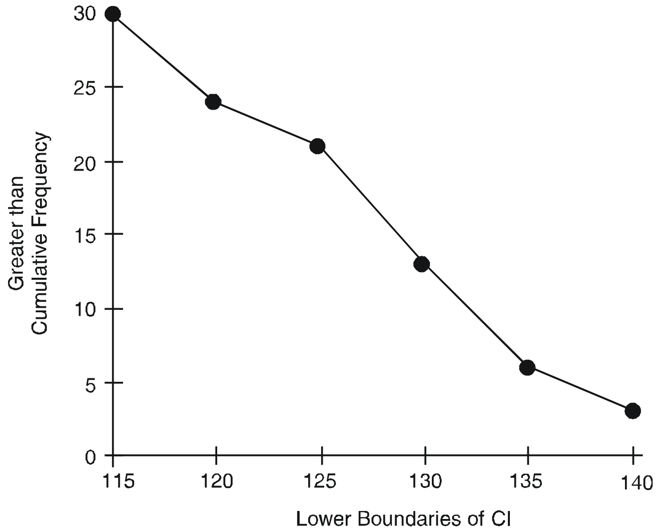
A frequency polygon



A cumulative frequency ogive (<)



A cumulative frequency ogive (>)



Problem 2: Construct a stem and leaf display for the data of IQ scores presented in the preceding example.

Solution:

The IQ scores of the given 30 students are presented in an ordered array, as follows:

115 115 116 116 118 119 120 120 122 125

125 125 126 127 127 127 129 130 130 132

132 132 133 134 135 135 137 140 140 142

NOTES

The stem would consist of the first two digits and the leaf would consist of the last digit.

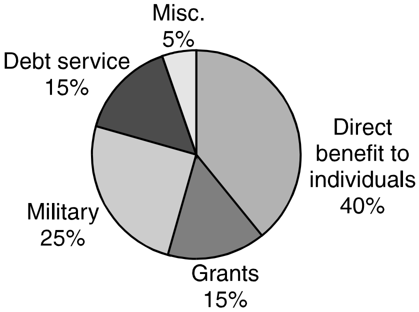
Stem	Leaves
11	5 5 6 6 8 9
12	0 0 2 5 5 5 6 7 7 7 9
13	0 0 2 2 2 3 4 5 5 7
14	0 0 2

Problem 3: Suppose the Office of the Management and Budget (OMB) has determined that the Federal Budget for 2008 would be utilized for proportionate spending in the following categories. Construct a pie chart to represent this data.

Category	Per cent Allocation
Direct benefit to individuals	40
State, local grants	15
Military spending	25
Debt service	15
Misc. operations	5
	Total 100%

Solution:

The pie chart is presented as follows. Care must be taken so that the percentage allocation of budget is represented by the appropriate proportion of the pie.



Check Your Progress

6. What is stem and leaf display?
7. What is a pictogram?
8. What is a frequency polygon?

1.5 SUMMARY

NOTES

- In order for the quantitative and numerical data to be identified as statistics, it must possess certain identifiable characteristics. Some of these characteristics include: (a) Statistics are aggregates of facts (b) Statistics are numerically expressed (c) Statistical data is collected in a systematic manner.
- Statistics is no longer confined to the domain of mathematics. It has spread to most of the branches of knowledge including social sciences and behavioural sciences. One of the reasons for its phenomenal growth is the variety of different functions attributed to it.
- The field of statistics, though widely used in all areas of human knowledge and widely applied in a variety of disciplines such as business, economics and research, has its own limitations. Some of the limitations include: (a) It does not deal with individual values (b) Statistical conclusions are not universally true (c) Statistics can be misused.
- There is hardly any walk of life which has not been affected by statistics—ranging from a simple household to big businesses and the government. Some of the important areas where the knowledge of statistics is usefully applied are government, economics, physical, natural and social sciences, statistics and research.
- Statistics influence the operations of business and management in many dimensions. Statistical applications include the area of production, marketing, promotion of product, financing, distribution, accounting, marketing research, manpower planning, forecasting, research and development and so on.
- A researcher needs to be familiar with the various statistical methods so as to be able to use the appropriate method in his research study. There are certain basic statistical methods, which can be classified into three groups as follows: (a) Descriptive statistics (b) Inferential statistics (c) Measures of central tendency and dispersion.
- According to Smith, descriptive statistics is the formulation of rules and procedures where data can be placed in a useful and significant order. The foundation of applicability of descriptive statistics is the need for complete data presentation.
- Statistical data can be organized into a frequency distribution which simply lists the value of the variable and frequency of its occurrence in a tabular form. A frequency distribution can be defined as the list of all the values obtained in the data and the corresponding frequency with which these values occur in the data.
- The number of groups and the size of class interval are more or less arbitrary in nature within the general guidelines established for constructing a frequency distribution.

NOTES

- In statistics, calculations are performed by arranging the large raw (ungrouped) data set into grouped data and are represented in tabular form called frequency distribution table. The data to be grouped must be homogenous and comparable. The frequency distribution table gives the size and the number of class intervals. The range of each class is defined by the class boundaries.
- The data collected first-hand for any statistical evaluation is considered as raw or ungrouped data as it is not meaningful and does not present a clear picture. It is then arranged in the ascending or the descending order in a tabular form called array.
- While the frequency distribution table tells us the number of units in each class interval, it does not tell us directly the total number of units that lie below or above the specified values of class intervals. This can be determined from a cumulative frequency distribution. When the interest of the investigator focusses on the number of items below a specified value, then this specified value is the upper limit of the class interval. It is known as less than cumulative frequency distribution.
- Stem and leaf display is another form of presentation of the data distribution. It allows us to condense data but still retain the individuality of the data. The idea is based on an analogy to plants. The leaves in the stem and leaf diagram are the last digit in each number of observed data. The first digit or digits, as the case may be, are the stem. All the values in the stem are listed in order in a column and a vertical line is drawn beside them and then all the corresponding leaf values are recorded for each stem in a row to the right of the vertical line.
- The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data. These important features are more simply presented in the form of graphs. Also, diagrams facilitate comparisons between two or more sets of data.
- Diagrams are more suitable to illustrate discrete data, while continuous data is better represented by graphs. The following are the diagrammatic and graphic representation methods that are commonly used.

1.6 KEY TERMS

- **Statistics:** It is the practice or science of collecting and analysing numerical data in large quantities, especially for the purpose of inferring proportions in a whole from those in a representative sample.
- **Economics:** It is a branch of knowledge concerned with the production, consumption, and transfer of wealth.

- **Gross National Product (GNP):** It is the total value of goods produced and services provided by a country during one year, equal to the gross domestic product plus the net income from foreign investments.
- **Frequency distribution:** It is a mathematical function showing the number of instances in which a variable takes each of its possible values.
- **Cumulative:** It refers to increasing or increased in quantity, degree, or force by successive additions.

1.7 ANSWERS TO 'CHECK YOUR PROGRESS'

1. According to Smith, descriptive statistics is the formulation of rules and procedures where data can be placed in a useful and significant order. The foundation of applicability of descriptive statistics is the need for complete data presentation.
2. Hypothesis testing attempts to validate or disprove preconceived ideas. In creating hypothesis, one thinks of a possible explanation for a remarked behaviour. The hypothesis dictates the data selected to be analysed for further interpretations.
3. The table must have a title and an identification number. The table title should be short and usually would not include any verbs or articles. It only refers to the population or parameter being studied. The title should be briefly yet clearly descriptive of the information provided. The numbering of tables is usually in a series and generally one makes use of roman numbers to identify them.
4. A frequency distribution can be defined as the list of all the values obtained in the data and the corresponding frequency with which these values occur in the data.
5. The data collected first-hand for any statistical evaluation is considered as raw or ungrouped data as it is not meaningful and does not present a clear picture.
6. Stem and leaf display is another form of presentation of the data distribution. It allows us to condense data but still retain the individuality of the data. The idea is based on an analogy to plants.
7. Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.
8. A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies and these plotted points are joined together by straight lines.

1.8 QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. 'Statistics are numerically expressed'. Explain this statement with examples.
2. What are the various functions of statistics?
3. How is a frequency distribution constructed?
4. Discuss Cumulative Frequency Curve (Ogive) with examples.

Long-Answer Questions

1. Discuss the limitations of statistics in detail.
2. Describe the areas in which statistics have been extensively and effectively used.
3. What are the three different types of graphic representation of statistics? Discuss in detail.

1.9 FURTHER READING

Garett, H. E. 1971. *Statistics in Psychology and Education, 6th Indian Edition*. Bombay: Vakils, Feffer and Simon.

Guilford, J. F. 1954. *Psychometric Methods, 2nd Edition*. New Delhi: Tata McGraw Hill.

Health, R. W. and N. M. Downie. 1970. *Basic Statistical Methods, 3rd Edition*. New York: Harper International.

McNemar, J. 1967. *Psychological Theory*. New York: McGraw Hill.

UNIT 2 MEASURES OF CENTRAL TENDENCY

NOTES

Structure

2.0 INTRODUCTION

In this unit, you will learn about the measures of central tendency and dispersion. There are several commonly used measures of central tendency, such as arithmetic mean, mode and median. These values are very useful not only in presenting the overall picture of the entire data but also for the purpose of making comparisons among two or more sets of data. In addition, you will learn about the harmonic mean. If *a*, *b*, *c* are in HP (harmonical progression), then *b* is called a Harmonic Mean between *a* and *c*, written as HM. Moreover, you will also learn about the measures of dispersion. A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency. Under this, you will study range, quartile deviation, average deviation and standard deviation. Coefficient of variation will also be discussed in this unit.

2.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Describe the interpretation and uses of measures of central tendency
- Discuss the computation of arithmetic mean, median and mode
- Explain the concept of measures of variability/dispersion and its significance in statistical analysis
- Discuss range, quartile deviation, average deviation and standard deviation

NOTES

2.2 MEASURES OF CENTRAL TENDENCY: CALCULATION, INTERPRETATION AND USE OF MEASURES OF CENTRAL TENDENCY

There are several commonly used measures of central tendency, such as arithmetic mean, mode and median. These values are very useful not only in presenting the overall picture of the entire data but also for the purpose of making comparisons among two or more sets of data.

As an example, questions like 'How hot is the month of June in Delhi?' can be answered, generally by a single figure of the average for that month. Similarly, suppose we want to find out if boys and girls at age 10 years differ in height for the purpose of making comparisons. Then, by taking the average height of boys of that age and average height of girls of the same age, we can compare and record the differences.

While arithmetic mean is the most commonly used measure of central location, mode and median are more suitable measures under certain set of conditions and for certain types of data. However, each measure of central tendency should meet the following requisites:

1. It should be easy to calculate and understand.
2. It should be rigidly defined. It should have only one interpretation so that the personal prejudice or bias of the investigator does not affect its usefulness.
3. It should be representative of the data. If it is calculated from a sample, then the sample should be random enough to be accurately representing the population.
4. It should have sampling stability. It should not be affected by sampling fluctuations. This means that if we pick 10 different groups of college students at random and compute the average of each group, then we should expect to get approximately the same value from each of these groups.
5. It should not be affected much by extreme values. If few very small or very large items are present in the data, they will unduly influence the value of the average by shifting it to one side or other, so that the average would not be really typical of the entire series. Hence, the average chosen should be such that it is not unduly affected by such extreme values.

2.2.1 Measures of Central Tendency: Arithmetic Mean, Median and Mode

If the progress scores of the students of a class are taken and they are arranged in a frequency distribution, we may sometime find that there are very few students who either score very high or very low. The marks of most of the student will lie somewhere between the highest and the lowest scores of the whole class. This tendency of a group about distribution is named as central tendency and typical score that lies in between the extremes and shared by most of the students is referred to as a measure of central tendency. Tate in 1955 defines the measures of central tendency as, *A sort of average or typical value of the items in the series and its function is to summarize the series in terms of this average value*.

The most common measures of central tendency are:

1. Arithmetic Mean or Mean
2. Median
3. Mode

Let us consider the three measures of central tendency.

(a) Arithmetic Mean: This is also commonly known as simply the mean. Even though average, in general, means any measure of central location, when we use the word average in our daily routine, we always mean the arithmetic average. The term is widely used by almost every one in daily communication. We speak of an individual being an average student or of average intelligence. We always talk about average family size or average family income or Grade Point Average (GPA) for students, and so on.

Calculating Arithmetic Mean (*M*): The simplest but most useful measure of central tendency is the arithmetic mean. It can be defined as the sum of all the values of the items in a series divided by the number of items. It is represented by the letter *M*.

Calculation of Mean in the Case of Ungrouped Data

Let $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9$ and X_{10} be the scores obtained by 10 students on an achievement test. Then the arithmetic mean or mean score of the group of these ten students can be calculated as:

$$M = X_1 + X_2 + X_3 + X_4 + X_5 + \ldots + X_{10}/10$$

The formula for calculating the mean of an ungrouped data is as follows:

$$M = \sum X/N$$

Where, $\sum X$ stands for the sum of sores or values of the items and N for the total number in a series or group.

Calculation of Mean in the Case of Grouped Data (Data in the form of Frequency Distribution)

General Method: In a frequency distribution where all the frequencies are greater than one, the mean is calculated by the formula:

$$\text{M} = \sum fX/N$$

Where, X represents the mid-point of the class interval, f its respective frequency and N the total of all frequencies.

Short-Cut Method: Mean for the grouped data can be computed easily with the help of following formula:

$$M = A + \sum fx'/N \times i$$

NOTES

Where,

A = Assumed mean.

i = Class interval.

f = Respective frequency of the mid-values of the class intervals.

N = Total Frequency.

$x' = X - A / i$

Combined Mean: If the arithmetic averages and the number of items in two or more related groups are known, the combined (or composite) mean of the entire group can be obtained by the following formula:

$$\overline{\overline{X}} = \left[\frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}\right]$$

The advantage of combined arithmetic mean is that, one can determine the overall mean of the combined data without having to going back to the original data.

For example:

We can find the combined mean for the data given below:

$n_1 = 10, x_1 = 2, n_2 = 15, x_2 = 3$

To obtain the mean:

$$\overline{\overline{X}} = \left[\frac{n_1\overline{x}_1 + n_2\overline{x}_2}{n_1 + n_2}\right]$$

$$= \left[\frac{10 \times 2 + 15 \times 3}{10 + 15}\right]$$

$$= \frac{20 + 45}{25}$$

$$= 2.6$$

For discussion purposes, let us assume a variable X which stands for some scores, such as the ages of students. Let the ages of 5 students be 19, 20, 22, 22 and 17 years. Then variable X would represent these ages as follows:

X: 19, 20, 22, 22, 17

Placing the Greek symbol σ(Sigma) before X would indicate a command that all values of X are to be added together. Thus:

$\sigma X = 19 + 20 + 22 + 22 + 17$

The mean is computed by adding all the data values and dividing it by the number of such values. The symbol used for sample average is $\overline{X}$ so that:

$$\overline{X} = \frac{19 + 20 + 22 + 22 + 17}{5}$$

In general, if there are *n* values in the sample, then:

$$\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$$

In other words,

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}, \; i = 1, 2, ..., \times 2\ n.$$

The above formula states, add up all the values of X_i where the value of *i* starts at 1 and ends at n with unit increments so that *i* = 1, 2, 3, ..., n.

If instead of taking a sample, we take the entire population in our calculations of the mean, then the symbol for the mean of the population is μ (mu) and the size of the population is *N*, so that:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}, \qquad i = 1,\ 2\ ...N.$$

If we have the data in grouped discrete form with frequencies, then the sample mean is given by:

$$\overline{X} = \frac{\Sigma f(X)}{\Sigma f}$$

Where, Σf = Summation of all frequencies' *n*.

$\Sigma f(X)$ = Summation of each value of *X* multiplied by its corresponding frequency (*f*).

Example 2.1: Let us take the ages of 10 students as follows:

19, 20, 22, 22, 17, 22, 20, 23, 17, 18

Solution: This data can be arranged in a frequency distribution as follows:

(*X*)	(*f*)	*f*(*X*)
17	2	34
18	1	18
19	1	19
20	2	40
22	3	66
23	1	23
	Total = 10	200

In the above case we have $\sum f = 10$ and $\sum f(X) = 200$, so that:

$$\overline{X} = \frac{\Sigma f(X)}{\Sigma f}$$

$$= 200/10 = 20$$

Characteristics of the Mean

The arithmetic mean has three interesting properties. These are:

1. The sum of the deviations of individual values of X from the mean will always add up to zero. This means that if we subtract all the individual values from their mean, then some values will be negative and some will be positive, but if all these differences are added together then the total sum will be zero. In other words, the positive deviations must balance the negative deviations. Or symbolically:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0, i = 1, 2, ..., n.$$

2. The second important characteristic of the mean is that it is very sensitive to extreme values. Since the computation of the mean is based upon inclusion of all values in the data, an extreme value in the data would shift the mean towards it, thus making the mean unrepresentative of the data.
3. The third property of the mean is that the sum of squares of the deviations about the mean is minimum. This means that if we take differences between individual values and the mean and square these differences individually and then add these squared differences, then the final figure will be less than the sum of the squared deviations around any other number other than the mean. Symbolically, it means that:

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \text{Minimum}, i = 1, 2, ..., n.$$

Advantages of Mean

The following are the various advantages of mean:

1. Its concept is familiar to most people and is intuitively clear.
2. Every data set has a mean, which is unique and describes the entire data to some degree. For instance, when we say that the average salary of a professor is ₹ 25,000 per month, it gives us a reasonable idea about the salaries of professors.
3. It is a measure that can be easily calculated.
4. It includes all values of the data set in its calculation.
5. Its value varies very little from sample to sample taken from the same population.
6. It is useful for performing statistical procedures, such as computing and comparing the means of several data sets.

Disadvantages of Mean

The following are the various disadvantages of mean:

1. It is affected by extreme values, and hence, not very reliable when the data set has extreme values especially when these extreme values are on one side

of the ordered data. Thus, a mean of such data is not truly a representative of such data. For instance, the average age of three persons of ages 4, 6 and 80 years gives us an average of 30.

2. It is tedious to compute for a large data set as every point in the data set is to be used in computations.
3. We are unable to compute the mean for a data set that has open-ended classes either at the high or at the low end of the scale.
4. The mean cannot be calculated for qualitative characteristics, such as beauty or intelligence, unless these can be converted into quantitative figures, such as intelligence into IQs.

(b) Median: The median is a measure of central tendency and it appears in the centre of an ordered data. It divides the list of ordered values in the data into two equal parts so that half of the data will have values less than the median and half will have values greater than the median.

If the total number of values is odd, then we simply take the middle value as the median. For instance, if there are 5 numbers arranged in order, such as 2, 3, 3, 5, 7, then 3 is the middle number and this will be the median. However, if the total number of values in the data is even, then we take the average of the middle two values. For instance, let there be 6 numbers in the ordered data, such as 2, 3, 3, 5, 7, 8, then the average of middle two numbers which are 3 and 5 would be the median, which is:

$$\text{Median} = \frac{(3+5)}{2} = 4$$

In general, the median is $\frac{n+1}{2}$th observation in the ordered data.

The median is a useful measure in the sense that it is not unduly affected by extreme values and is specially useful in open ended frequencies.

Calculating Median (M_d): If the items of a series are arranged in ascending or descending order of magnitude, the measure or value of the central item in the series is termed as median. The median of a distribution can thus be said as the point on the score scale below which half (or 50 per cent) of the scores fall. Thus, median is the score or the value of that central item which divides the series into two equal parts. Therefore, it should be understood that the central item itself is not the median. It is only the measure or value of the central item that is known as the median. For example, if we arrange in ascending or descending order the marks of 5 students, then the marks obtained by the third student from either side will be termed as median of the scores of the group of students under consideration.

Computation of Median for Ungrouped Data

The following two situations could arise:

1. **When N (no. of items in a series) is odd:** In this case where N is odd (not divisible by 2), the median can be computed by the following formula:

   M_d = The measure or value of the $(N + 1)/2$ th item.

2. **When N (no. of items in a series) is even:** In this case where N is even (divisible by 2), the median can be determined by the following formula:

   M_d = The value of the (N/2) th item + The value of [(N/2) + 1] th item/2

Calculation of Median for Grouped Data (In the Form of Frequency Distribution)

If the data is available in the form of a frequency distribution like below, then calculation of median first requires the location of median class.

Scores	*f*
65-69	1
60-64	3
55-59	4
50-54	7
45-49	9
40-44	11
35-39	8
30-34	4
25-29	2
20-24	1
	$N = 50$

Actually, median is the measure or score of the central item. Therefore, it is needed to locate the central item. It may be done through the formulae given earlier in case of ungrouped data for the odd and even values of N (total frequencies). Here, in the present distribution, N (= 50) is even. Therefore, median will fall somewhere between the score of 25th and 26th items in the given distribution. In the given frequency distribution table, if we add frequencies either above or below we may see that the class interval designated as 40-44 is to be labeled as the class where the score representing median will fall.

After estimating the median class, the median of the distribution may be interpolated with the help of following formula:

$M_d = L + [\ (N/2) - F\ /f] \times i$

Where,

L = Exact lower limit of the median class.

F = Total of all frequencies before in the median class.

f = Frequency of the median class.

i = Class interval.

N = Total of all the frequency.

By applying the above formula, we can compute the median of the given distribution in the following way:

$$M_d = 39.5 + (50/2) - 15 / 11 \times 5 = 39.5 + 10/11 \times 5$$
$$= 39.5 + 50/11 = 39.5 + 4.55 = 44.05$$

Advantages of Median

The following are the advantages of median:

1. Median is a positional average and hence the extreme values in the data set do not affect it as much as they do to the mean.
2. Median is easy to understand and can be calculated from any kind of data, even for grouped data with open-ended classes.
3. We can find the median even when our data set is qualitative and can be arranged in the ascending or the descending order, such as average beauty or average intelligence.
4. Similar to mean, median is also unique, meaning that there is only one median in a given set of data.
5. Median can be located visually when the data is in the form of ordered data.
6. The sum of absolute differences of all values in the data set from the median value is minimum, meaning that it is less than any other value of central tendency in the data set, which makes it more central in certain situations.

Disadvantages of Median

The following are the disadvantages of median:

1. The data must be arranged in order to find the median. This can be very time consuming for a large number of elements in the data set.
2. The value of the median is affected more by sampling variations. Different samples from the same population may give significantly different values of the median.
3. The calculation of median in case of grouped data is based on the assumption that the values of observations are evenly spaced over the entire class interval and this is usually not so.
4. Median is comparatively less stable than the mean, particularly for small samples, due to fluctuations in sampling.
5. Median is not suitable for further mathematical treatment. For instance, we cannot compute the median of the combined group from the median values of different groups.

(c) **Mode:** The mode is another form of average and can be defined as the most frequently occurring value in the data. The mode is not affected by extreme values in the data and can easily be obtained from an ordered set of data. It can be useful and more representative of the data under certain conditions and is the only measure of central tendency that can be used for qualitative data. For instance, when a researcher quotes the opinion of an average person, he is probably referring to the most frequently expressed opinion which is the modal opinion. In our example of ages of 10 students as:

19, 20, 22, 22, 17, 22, 20, 23, 17 and 18

The mode is 22, since it occurs more often than any other value in this data.

Calculating Mode (M_0): Mode is defined as the size of a variable which occurs most frequently. It is the point on the score sale that corresponds to the maximum frequency of the distribution. In any series, it is the value of the item which is most characteristics or common and is usually repeated the maximum number of times.

Computation of Mode for Ungrouped Data

Mode can easily be computed merely by looking at the data. All that one has to do is to find out the score which is repeated maximum number of times.

For example, suppose we have to find out the value of mode from the following scores of students:

25, 29, 24, 25, 27, 25, 28, 25, 29

Here, the score 25 is repeated maximum number of times and thus, value of the mode in this case is 25.

Computation of Mode for Grouped Data

When data is available in the form of frequency distribution, the mode is computed from the following formula:

Mode (M_0) = 3 $M_d - 2M$

Where, M_d is the median and M is the mean of the given distribution. The mean as well as the median of the distribution are first computed and then, with the help of the above formula, mode is computed.

Another Method for Grouped Data

Mode can be computed directly from the frequency distribution table without calculating mean and median. For this purpose, we can use the following formula:

$$M_0 = L + f_1 / (f_1 + f_{-1}) \times i$$

Where,

L = Lower limit of the model class (the class in which mode maybe supposed to lie).

i = Class interval.

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

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

Advantages of Mode

The following are the advantages of mode:

1. Similar to median, the mode is not affected by extreme values in the data.

NOTES

2. Its value can be obtained in open-ended distributions without ascertaining the class limits.
3. It can be easily used to describe qualitative phenomenon. For instance, if most people prefer a certain brand of tea then this will become the modal point.
4. Mode is easy to calculate and understand. In some cases it can be located simply by observation or inspection.

Disadvantages of Mode

The following are the disadvantages of mode:

1. Quite often, there is no modal value.
2. It can be bi-modal or multi-modal or it can have all modal values making its significance more difficult to measure.
3. If there is more than one modal value, the data is difficult to interpret.
4. A mode is not suitable for algebraic manipulations.
5. Since the mode is the value of maximum frequency in the data set, it cannot be rigidly defined if such frequency occurs at the beginning or at the end of the distribution.
6. It does not include all observations in the data set, and hence, less reliable in most of the situations.

2.2.2 Weighted Arithmetic Mean

In the computation of arithmetic mean we had given equal importance to each observation in the series. This equal importance may be misleading if the individual values constituting the series have different importance as in the following example:

The Raja Toy shop sells

Toy Cars at	₹ 3 each
Toy Locomotives at	₹ 5 each
Toy Aeroplanes at	₹ 7 each
Toy Double Decker at	₹ 9 each

What shall be the average price of the toys sold, if the shop sells 4 toys, one of each kind?

Mean Price, i.e., $\bar{x} = \frac{\sum x}{4} = ₹\frac{24}{4} = ₹\ 6$

In this case the importance of each observation (Price quotation) is equal in as much as one toy of each variety has been sold. In the above computation of the arithmetic mean this fact has been taken care of by including 'once only' the price of each toy.

But if the shop sells 100 toys: 50 cars, 25 locomotives, 15 aeroplanes and 10 double deckers, the importance of the four price quotations to the dealer is **not equal** as a source of earning revenue. In fact their respective importance is equal to the number of units of each toy sold, i.e.,

The importance of Toy Car	50
The importance of Locomotive	25

The importance of Aeroplane	15
The importance of Double Decker	10

It may be noted that 50, 25, 15, 10 are the quantities of the various classes of toys sold. It is for these quantities that the term 'weights' is used in statistical language. Weight is represented by symbol 'w', and Σw represents the sum of weights.

While determining the 'average price of toy sold' these weights are of great importance and are taken into account in the manner illustrated below:

$$\bar{x} = \frac{w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4}{w_1 + w_2 + w_3 + w_4} = \frac{\sum wx}{\sum w}$$

When w_1, w_2, w_3, w_4 are the respective weights of x_1, x_2, x_3, x_4 which in turn represent the price of four varieties of toys, viz., car, locomotive, aeroplane and double decker, respectively.

$$\bar{x} = \frac{(50\times3)+(25\times5)+(15\times7)+(10\times9)}{50+25+15+10}$$

$$= \frac{(150)+(125)+(105)+(90)}{100} = \frac{470}{100} = ₹\,4.70$$

The table below summarizes the steps taken in the computation of the weighted arithmetic mean.

$$\Sigma w = 100; \quad \Sigma wx = 470$$

$$\bar{x} = \frac{\sum wx}{\sum w} = \frac{470}{100} = 4.70$$

The weighted arithmetic mean is particularly useful where we have to compute the *mean of means.* If we are given two arithmetic means, one for each of two different series, in respect of the *same variable*, and are required to find the arithmetic mean of the combined series, the weighted arithmetic mean is the only suitable method of its determination.

Weighted Arithmetic Mean of Toys Sold by the Raja Toy Shop

Toys	*Price Per Toy* *₹x*	*Number Sold* *w*	*Price × Weight* *xw*
Car	3	50	150
Locomotive	5	25	125
Aeroplane	7	15	105
Double Decker	9	10	90
		Σw = 100	Σxw = 470

Example 2.2: The arithmetic mean of daily wages of two manufacturing concerns A Ltd. and B Ltd. is ₹ 5 and ₹ 7, respectively. Determine the average daily wages of both concerns if the number of workers employed were 2,000 and 4,000, respectively.

Solution: (*a*) Multiply each average (viz., 5 and 7) by the number of workers in the concern it represents.

(*b*) Add up the two products obtained in (*a*) above.

(*c*) Divide the total obtained in (*b*) by the total number of workers.

Weighted Mean of Mean Wages of A Ltd. and B Ltd.

Manufacturing Concern	*Mean Wages* *x*	*Workers Employed* *w*	*Mean Wages × Workers Employed* *wx*
A Ltd.	5	2,000	10,000
B Ltd.	7	4,000	28,000
		$\sum w = 6{,}000$	$\sum wx = 38{,}000$
		$\bar{x} = \frac{\sum wx}{\sum w}$ $= \frac{38{,}000}{6{,}000}$ = ₹ 6.33	

The above mentioned examples explain that 'Arithmetic Means and Percentage' are not original data. They are derived figures and their importance is relative to the original data from which they are obtained. This relative importance must be taken into account by weighting while averaging them (means and percentage).

2.2.3 Harmonic Mean

If *a*, *b*, *c* are in HP, then *b* is called a *Harmonic Mean* between *a* and *c*, written as HM.

Harmonical Progression

Non zero quantities whose reciprocals are in AP, or Arithmetic Progression are said to be in *Harmonical Progression*, written as HP.

Consider the following examples:

(*a*) $1, \frac{1}{3}, \frac{1}{5}, \frac{1}{7}, \ldots\ldots$

(*b*) $\frac{1}{2}, \frac{1}{5}, \frac{1}{8}, \frac{1}{11}, \ldots\ldots$

(*c*) $2, \frac{5}{2}, \frac{10}{3}, \ldots$

(*d*) $\frac{1}{a}, \frac{1}{a+b}, \frac{1}{a+2b}, \ldots\ldots \quad a, b > 0$

(*e*) $5, \frac{55}{9}, \frac{55}{7}, 11, \ldots\ldots$

It can be easily checked that in each case, the series obtained by taking reciprocal of each of the term is an AP.

To Insert *n* Harmonic Means between *a* and *b*

Let $H_1, H_2, H_3, \ldots, H_n$ be the required Harmonic Means. Then,

$a, H_1, H_2, \ldots, H_n, b$ are in HP

i.e., $\frac{1}{a}, \frac{1}{H_1}, \frac{1}{H_2}, \ldots, \frac{1}{H_n}, \frac{1}{b}$ are in AP

Then, $\frac{1}{b} = (n+2)\text{th term of an AP}$

$$= \frac{1}{a} + (n+1)d$$

Where d is the common difference of AP.

This gives, $d = \frac{a-b}{(n+1)ab}$

Now, $\frac{1}{H_1} = \frac{1}{a} + d = \frac{1}{a} + \frac{a-b}{(n+1)\,ab}$

$$= \frac{nb+b+a-b}{(n+1)\,ab} = \frac{a+nb}{(n+1)\,ab}$$

So, $\frac{1}{H_1} = \frac{a+nb}{(n+1)\,ab}$

$\Rightarrow$ $H_1 = \frac{(n+1)\,ab}{a+nb}$

Again, $\frac{1}{H_2} = \frac{1}{a} + 2d = \frac{1}{a} + \frac{2\,(a-b)}{(n+1)\,ab}$

$$= \frac{nb+b+2a-2b}{(n+1)\,ab} = \frac{2a-b+nb}{(n+1)\,ab}$$

$\Rightarrow$ $H_2 = \frac{(n+1)\,ab}{2a-b+nb}$

Similarly, $\frac{1}{H_3} = \frac{1}{a} + 3d = \frac{3a-2b+nb}{(n+1)\,ab}$

$\Rightarrow$ $H_3 = \frac{(n+1)\,ab}{3a-2b+nb}$ and so on,

$$\frac{1}{H_n} = \frac{1}{a} + nd = \frac{1}{a} + \frac{n\,(a-b)}{(n+1)\,ab}$$

$$= \frac{nb+b+na-nb}{(n+1)\,ab}$$

$$= \frac{na+b}{(n+1)\,ab} \Rightarrow H_n = \frac{(n+1)\,ab}{na+b}$$

Example 2.3: Find the 5th term of 2, $2\frac{1}{2}$, $3\frac{1}{3}$, ...

Solution: Let 5th term be x. Then, $\frac{1}{x}$ is 5th term of corresponding AP $\frac{1}{2}, \frac{2}{5}, \frac{3}{10}, \ldots\ldots$

Then, $\frac{1}{x} = \frac{1}{2} + 4\left(\frac{2}{5} - \frac{1}{2}\right) = \frac{1}{2} + 4\left(\frac{-1}{10}\right)$

$\Rightarrow$ $\frac{1}{x} = \frac{1}{2} - \frac{2}{5} = \frac{1}{10} \Rightarrow x = 10$

Example 2.4: Insert two harmonic means between $\frac{1}{2}$ and $\frac{4}{17}$.

NOTES

Solution: Let H_1, H_2 be two harmonic means between $\frac{1}{2}$ and $\frac{4}{17}$.

Thus, 2, $\frac{1}{H_1}, \frac{1}{H_2}, \frac{17}{4}$ are in AP. Let d be their common difference.

Then, $$\frac{17}{4} = 2 + 3d$$

$\Rightarrow$ $$3d = \frac{9}{4} \quad \Rightarrow \quad d = \frac{3}{4}$$

Thus, $$\frac{1}{H_1} = 2 + \frac{3}{4} = \frac{11}{4} \quad \Rightarrow \quad H_1 = \frac{4}{11}$$

$$\frac{1}{H_2} = 2 + 2 \times \frac{3}{4} = \frac{7}{2} \quad \Rightarrow \quad H_2 = \frac{2}{7}$$

Required harmonic means are $\frac{4}{11}, \frac{2}{7}$.

CHECK YOUR PROGRESS

1. Define median.
2. In a frequency list, where all the frequencies are greater than one, the mean is calculated by which formula?
3. Give any two disadvantages of mean.

2.3 MEASURES OF VARIABILITY/DISPERSION: WHEN AND WHERE TO USE THE VARIOUS MEASURES OF VARIABILITY

A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.

A measure of dispersion may be expressed in an 'absolute form' or in a 'relative form'. It is said to be in an absolute form when it states the actual amount by which the value of an item on an average deviates from a measure of central tendency. Absolute measures are expressed in concrete units, i.e., units in terms of which the data have been expressed, e.g., rupees, centimetres, kilograms, etc., and are used to describe frequency distribution.

A relative measure of dispersion computed is a quotient obtained by dividing the absolute measures by a quantity in respect to which absolute deviation has been computed. It is as such a pure number and is usually expressed in a percentage form. Relative measures are used for making comparisons between two or more distributions.

A measure of dispersion should possess all those characteristics which are considered essential for a measure of central tendency, which are as follows:

- It should be based on all observations.
- It should be readily comprehensible.

- It should be fairly easily calculated.
- It should be affected as little as possible by fluctuations of sampling.
- It should be amenable to algebraic treatment.

Types of Measures of Dispersion

There are four measures of dispersion which are given below:

(a) Range (*R*)

(b) Quartile Deviation (*QD*)

(c) Average Deviation (*AD*)

(d) Standard Deviation (*SD*)

Each of the above measures of dispersion tells us how the individual scores are scattered or spread throughout the distribution or the given data.

2.3.1 Quartile Deviation (QD)

There are many types of measures of dispersion, one of this is the semi-interquartile range, usually termed as 'Quartile Deviation' or QD. Quartiles are the points which divide the array into four equal parts. More precisely, Q_1 gives the value of the item 1/4th the way up the distribution and Q_3 the value of the item 3/4th the way up the distribution. Between Q_1 and Q_3 are included half the total number of items. The difference between Q_1 and Q_3 includes only the central items but excludes the extremes. Since under most circumstances, the central half of the series tends to be fairly typical of all the items, the interquartile range ($Q_3 - Q_1$) affords a convenient and often a good indicator of the absolute variability. The larger the interquartile range, the larger the variability.

Usually, one-half of the difference between Q_3 and Q_1 is used and it is given the name of quartile deviation or semi-interquartile range. The interquartile range is divided by 2 for the reason that half of the interquartile range will, in a normal distribution, be equal to the difference between the median and any quartile. This means that 50 per cent items of a normal distribution will lie within the interval defined by the median plus and minus the semi-interquartile range.

Symbolically,

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

Where, Q_1 and Q_3 represent the 1st and 3rd quartiles of dispersion under consideration. The value $Q_3 - Q_1$ is the difference or range between the 3rd and 1st quartiles and is the interquartile range.

For computing quartile deviation, this interquartile range is divided by 2 and, therefore, quartile deviation is also names as semi-interquartile range. In this way, for computing *Q,* the value of Q_1 and Q_3 are first determined by the method and then applying the above formula, we get the value of the quartile deviation.

Let us find quartile deviations for the weekly earnings of labour in the four workshops whose data is given in Table 2.1. The computations are as shown in Table 2.1.

Table 2.1 *Weekly Earnings of Labourers in Four Workshops of the Same Type*

Weekly Earnings ₹	*No. of Workers*			
	Workshop A	*Workshop B*	*Workshop C*	*Workshop D*
15–16	...	...	2	...
17–18	...	2	4	...
19–20	...	4	4	4
21–22	10	10	10	14
23–24	22	14	16	16
25–26	20	18	14	16
27–28	14	16	12	12
29–30	14	10	6	12
31–32	...	6	6	4
33–34	...	...	2	2
35–36	...	...	...	...
37–38	...	...	4	...
Total	80	80	80	80
Mean	25.5	25.5	25.5	25.5

The range is as follows:

Workshop	*Range*
A	9
B	15
C	23
D	15

As shown in Table 2.2, Q.D. of workshop *A* is ₹ 2.12 and median value in 25.3. This means that if the distribution is symmetrical, the number of workers whose wages vary between (25.3–2.1) = ₹ 23.2 and (25.3 + 2.1) = ₹ 27.4, shall be just half of the total cases. The other half of the workers will be more than ₹ 2.1 removed from the median wage. As this distribution is not symmetrical, the distance between Q_1 and the median Q_2 is not the same as between Q_3 and the median. Hence, the interval defined by median plus and minus semi inter-quartile range will not be exactly the same as given by the value of the two quartiles. Under such conditions the range between ₹ 23.2 and ₹ 27.4 will not include precisely 50 per cent of the workers.

If quartile deviation is to be used for comparing the variability of any two series, it is necessary to convert the absolute measure to a coefficient of quartile deviation. To do this the absolute measure is divided by the average size of the two quartiles.

Symbolically,

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Applying this to our illustration of four workshops in Table 2.1 the coefficients of Q.D. are as given in Table 2.2.

Table 2.2 *Calculation of Quartile Deviation*

		Workshop *A*	*Workshop* *B*	*Workshop* *C*	*Workshop* *D*
Location of Q_2	$\frac{N}{2}$	$\frac{80}{2}=40$	$\frac{80}{2}=40$	$\frac{80}{2}=40$	$\frac{80}{2}=40$
	Q_2	$24.5+\frac{40-30}{22}\times 2$ $= 24.5 + 0.9$ $= 25.4$	$24.5+\frac{40-30}{18}\times 2$ $= 24.5 + 1.1$ $= 25.61$	$24.5+\frac{40-30}{16}\times 2$ $= 24.5 + 0.75$ $= 25.25$	$24.5+\frac{40-30}{16}\times 2$ $= 24.5 + 0.75$ $= 25.25$
Location of Q_1	$\frac{N}{4}$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$	$\frac{80}{4}=20$
	Q_1	$22.5+\frac{20-10}{22}\times 2$ $= 22.5 + .91$ $= 23.41$	$22.5+\frac{20-16}{14}\times 2$ $= 22.5 + .57$ $= 23.07$	$20.5+\frac{20-10}{10}\times 2$ $= 20.5 + 2$ $= 22.5$	$22.5+\frac{20-18}{16}\times 2$ $= 22.5 + .25$ $= 22.75$
Location of Q_3	$\frac{3N}{4}$	$3\times\frac{80}{4}=60$	60	60	60
	Q_3	$26.5+\frac{60-52}{14}\times 2$ $= 26.5 + 1.14$ $= 27.64$	$26.5+\frac{60-48}{16}\times 2$ $= 26.5 + 1.5$ $= 28.0$	$26.5+\frac{60-50}{12}\times 2$ $= 26.5 + 1.67$ $= 28.17$	$26.5+\frac{60-50}{12}\times 2$ $= 26.5 + 1.67$ $= 28.17$
Quartile Deviation	$\frac{Q_3-Q_1}{2}$	$\frac{27.64-23.41}{2}$ $=\frac{4.23}{2}=$ ₹ 2.12	$\frac{28-23.07}{2}$ $=\frac{4.93}{2}=$ ₹ 2.46	$\frac{28.17-22.5}{2}$ $=\frac{5.67}{2}=$ ₹ 2.83	$\frac{28.17-22.75}{2}$ $=\frac{5.42}{2}=$ ₹. 2.71
Coefficient of Quartile Deviation =	$\frac{Q_3-Q_1}{Q_3+Q_1}$	$\frac{27.64-23.41}{27.64+23.41}$ $= 0.083$	$\frac{28-23.07}{28+23.07}$ $= 0.097$	$\frac{28.17-22.5}{28.17+22.5}$ $= 0.112$	$\frac{28.17-22.75}{28.17+22.75}$ $= 0.106$

Characteristics of Quartile Deviation

The following are the characteristics of quartile deviation:

(a) The size of the quartile deviation gives an indication about the uniformity or otherwise of the size of the items of a distribution. If the quartile deviation is small, it denotes large uniformity. Thus, a coefficient of quartile deviation may be used for comparing uniformity or variation in different distributions.

(b) Quartile deviation is not a measure of dispersion in the sense that it does not show the scatter around an average, but only a distance on scale. Consequently, quartile deviation is regarded as a measure of partition.

(c) It can be computed when the distribution has open-end classes.

Limitations of Quartile Deviation

Except for the fact that its computation is simple and it is easy to understand, a quartile deviation does not satisfy any other test of a good measure of variation.

2.3.2 Average Deviation (AD)

In this section you will study that a weakness of the measures of dispersion, based upon the range or a portion thereof, is that the precise size of most of the variants has no effect on the result. As an illustration, the quartile deviation will be the same whether the variates between Q_1 and Q_3 are concentrated just above Q_1 or they are spread uniformly from Q_1 to Q_3. This is an important defect from the viewpoint of measuring the divergence of the distribution from its typical value. The Average Deviation (AD) is employed to answer the objection.

Average Deviation (AD), also called mean deviation, of a frequency distribution is the mean of the absolute values of the deviation from some measure of central tendency. In other words, mean deviation is the arithmetic average of the variations (deviations) of the individual items of the series from a measure of their central tendency.

Garrett in 1971 defines Average Deviation (AD) as the mean of deviations of all the separate scores in the series taken from their mean (occasionally from the median or mode). It is the simplest measure of variability that takes into account the fluctuation or variation of all the items in a series.

Computation of Average Deviation from Ungrouped Data

In the case of ungrouped data, the average deviation is calculated by the formula

$$\text{AD} = \sum |x| / N$$

Where, $x = \text{X} - M =$ Deviation of the raw score from the mean of the series and $|x|$ signifies that in the deviation values we ignore the algebraic signs +ve or –ve.

Computation of Average Deviation from Grouped Data

From the grouped data, AD can be computed by the following formula:

$$\text{AD} = \sum |fx| / N$$

We can measure the deviations from any measure of central tendency, but the most commonly employed ones are the median and the mean. The median is preferred because it has the important property that the average deviation from it is the least.

Calculation of mean deviation then involves the following steps:

(a) Calculate the median (or the mean) M_d (or $\overline{X}$).

(b) Record the deviations $|\,d\,| = |\,x - M_d\,|$ of each of the items, ignoring the sign.

(c) Find the average value of deviations.

$$\text{Mean Deviation} = \frac{\sum |d|}{N}$$

Example 2.5 explains it better.

Example 2.5: Calculate the mean deviation from the following data giving marks obtained by 11 students in a class test.

14, 15, 23, 20, 10, 30, 19, 18, 16, 25, 12.

Solution:

$$\text{Median} = \text{Size of } \frac{11+1}{2} \text{ th item}$$

$$= \text{Size of 6th item} = 18.$$

Serial No.	*Marks*	$\mid x - \text{Median} \mid$ $\mid d \mid$
1	10	8
2	12	6
3	14	4
4	15	3
5	16	2
6	18	0
7	19	1
8	20	2
9	23	5
10	25	7
11	30	12
		$\sum \mid d \mid = 50$

$$\text{Mean deviation from median} = \frac{\sum |d|}{N}$$

$$= \frac{50}{11} = 4.5 \text{ marks}$$

For grouped data, it is easy to see that the mean deviation is given by:

$$\text{Mean deviation} = \frac{\sum f\,|d|}{\sum f}$$

Where,

$|d| = |x - \text{Median}|$ for grouped discrete data.

$|d| = |M - \text{Median}|$ for grouped continuous data with M as the mid-value of a particular group.

Examples 2.6 and 2.7 illustrate the use of this formula.

Example 2.6: Calculate the mean deviation from the following data:

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

Solution:

Size	*Frequency (f)*	*Cumulative Frequency*	*Deviations from Median (9) | d |*	*f | d |*
6	3	3	3	9
7	6	9	2	12
8	9	18	1	9
9	13	31	0	0
10	8	39	1	8
11	5	44	2	10
12	4	48	3	12
	48			60

$$\text{Median} = \text{The size of } \frac{48+1}{2} = 24.5\text{th item which is } 9.$$

Therefore, deviations d are calculated from 9, i.e., $|\,d\,| = |\,x - 9\,|$.

$$\text{Mean deviation} = \frac{\Sigma f\,|d|}{\Sigma f} = \frac{60}{48} = 1.25$$

Example 2.7: Calculate the mean deviation from the following data:

x	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
f	18	16	15	12	10	5	2	2

Solution:

This is a frequency distribution with continuous variable. Thus, deviations are calculated from mid-values.

x	*Mid-Value*	*(f)*	*Less than (c.f.)*	*Deviation from Median | d |*	*f| d |*
0–10	5	18	18	19	342
10–20	15	16	34	9	144
20–30	25	15	49	1	15
30–40	35	12	61	11	132
40–50	45	10	71	21	210
50–60	55	5	76	31	155
60–70	65	2	78	41	82
70–80	75	2	80	51	102
		80			1182

NOTES

$$\text{Median} = \text{The size of } \frac{80}{2} \text{ th item}$$

$$= 20 + \frac{6}{15} \times 10 = 24$$

and then, mean deviation $= \dfrac{\Sigma f\,|d|}{\Sigma f}$

$$= \frac{1182}{80} = 14.775.$$

Merits and Demerits of the Average (Mean) Deviation

Merits

The merits are as follows:

1. It is easy to understand.
2. As compared to standard deviation, its computation is simple.
3. As compared to standard deviation, it is less affected by extreme values.
4. Since it is based on all values in the distribution, it is better than range or quartile deviation.

Demerits

The demerits are as follows:

1. It lacks those algebraic properties which would facilitate its computation and establish its relation to other measures.
2. Due to this, it is not suitable for further mathematical processing.

Coefficient of Mean or Average Deviation

The coefficient or relative dispersion is found by dividing the mean deviations recorded. Thus,

$$\text{Coefficient of MD} = \frac{\text{Mean Deviation}}{\text{Mean}}$$

(when deviations were recorded from the mean)

$$= \frac{\text{Mean Deviation}}{\text{Median}}$$

(when deviations were recorded from the median)

Applying the above formula to Example 2.7.

$$\text{Coefficient of MD} = \frac{14.775}{24}$$

$$= 0.616$$

2.3.3 Standard Deviation (SD)

By far the most universally used and the most useful measure of dispersion is the Standard Deviation (SD) or root mean square deviation about the mean. We have seen that all the methods of measuring dispersion so far discussed are not universally adopted for want of adequacy and accuracy. The range is not satisfactory as its magnitude is determined by most extreme cases in the entire group. Further, the range is notable because it is dependent on the item whose size is largely a matter of chance. Mean deviation method is also an unsatisfactory measure of scatter, as it ignores the algebraic signs of deviation. We desire a measure of scatter which is free from these shortcomings. To some extent standard deviation is one such measure.

The calculation of standard deviation differs in the following respects from that of mean deviation. First, in calculating standard deviation, the deviations are squared. This is done so as to get rid of negative signs without committing algebraic violence. Further, the squaring of deviations provides added weight to the extreme items, a desirable feature for certain types of series.

Second, the deviations are always recorded from the arithmetic mean, because although the sum of deviations is the minimum from the median, the sum of squares of deviations is minimum when deviations are measured from the arithmetic average. The deviation from $\bar{x}$ is represented by σ.

Thus, standard deviation, σ (sigma) is defined as the square root of the mean of the squares of the deviations of individual items from their arithmetic mean.

Standard deviation of a set of scores is defined as the square root of the average of the squares of the deviations of each score from the mean. Symbolically, we can say that:

$$SD = \surd \sum (X - M)^2/2$$
$$= \surd \sum x^2/ N$$

Where,

X = Individual score.

M = Mean of the given set of scores.

N = Total number of the sores.

x =Derivation of each score from the mean.

Standard Deviation or SD is regarded as the most stable and reliable measure of variability as it employs the mean for its computation. It is often called *root mean square deviation* and is denoted by the Greek letter sigma (σ).

Computation of Standard Deviation (SD) from Ungrouped Data

Standard deviation can be computed from the ungrouped scores by the formula:

$$\sigma = \surd \sum x^2 / N$$

Computation of Standard Deviation (SD) from Grouped Data

Standard deviation in case of grouped data can be computed by the formula:

$$\sigma = \surd \sum fx^2 / N$$

Computation of Standard Deviation (SD) from Grouped Data by Short-Cut Method

Standard deviation from grouped data can be computed by the following formula:

$$\sigma = \surd \sum fx'^2 / N - (\sum fx'/N)^2$$

Example 2.8: Compute the standard deviation for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21.

Solution:

We first calculate the mean as $\bar{x} = \sum x/N = 176/11 = 16$, and then calculate the deviation as follows:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
11	–5	25
12	–4	16
13	–3	9
14	–2	4
15	–1	1
16	0	0
17	+1	1
18	+2	4
19	+3	9
20	+4	16
21	+5	25
176		110

Thus,

$$\sigma = \sqrt{\frac{110}{11}} = \sqrt{10} = 3.16$$

Example 2.9: Find the standard deviation of the data in the following distributions:

x	12	13	14	15	16	17	18	20
f	4	11	32	21	15	8	6	4

Solution:

Since for calculation of $\bar{x}$, we need $\sum fx$ and then for σ we need $\sum f(x - \bar{x})^2$, the calculations are conveniently made in the following format.

x	f	fx	$d = x - \bar{x}$	d^2	fd^2
12	4	48	–3	9	36
13	11	143	–2	4	44
14	32	448	–1	1	32
15	21	315	0	0	0
16	15	240	1	1	15
17	8	136	2	4	32
18	5	90	3	9	45
20	4	80	5	25	100
	$\sum f$=100	$\sum fx$=1500			$\sum fd^2$=304

Here, $\bar{x} = \sum fx / \sum f = 1500/100 = 15$

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

$$= \sqrt{\frac{304}{100}} = \sqrt{3.04} = 1.74$$

Example 2.10: Compute the standard deviation by the short-cut method for the following data:

11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Solution: Let us assume that $A = 15$.

	$x' = (x - 15)$	x'^2
11	–4	16
12	–3	9
13	–2	4
14	–1	1
15	0	0
16	1	1
17	2	4
18	3	9
19	4	16
20	5	25
21	6	36
$N = 11$	$\sum x' = 11$	$\sum x'^2 = 121$

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

$$= \sqrt{\frac{121}{11} - \left(\frac{11}{11}\right)^2}$$

$$= \sqrt{11 - 1}$$

$$= \sqrt{10}$$

$$= 3.16$$

Another Method

If we assume A as zero, then the deviation of each item from the assumed mean is the same as the value of item itself. Thus, 11 deviates from the assumed mean of zero by 11, 12 deviates by 12, and so on. As such, we work with deviations without having to compute them, and the formula takes the following shape:

NOTES

x	x^2
11	121
12	144
13	169
14	196
15	225
16	256
17	289
18	324
19	361
20	400
21	441
176	2,926

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

$$= \sqrt{\frac{2926}{11} - \left(\frac{176}{11}\right)^2} = \sqrt{266 - 256} = 3.16$$

Combining Standard Deviations of Two Distributions

If we were given two sets of data of N_1 and N_2 items with means $\bar{x}_1$ and $\bar{x}_2$ and standard deviations σ_1 and σ_2, respectively, we can obtain the mean and standard deviation $\bar{x}$ and σ of the combined distribution by the following formulae:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2}$$

and $$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2}{N_1 + N_2}}$$

Example 2.11: The mean and standard deviations of two distributions of 100 and 150 items are 50, 5 and 40, 6, respectively. Find the standard deviation of all taken together.

Solution:

Combined mean,

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2}{N_1 + N_2} = \frac{100 \times 50 + 150 \times 40}{100 + 150}$$

$$= 44$$

Combined standard deviation,

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2(\bar{x} - \bar{x}_2)^2}{N_1 + N_2}}$$

$$= \sqrt{\frac{100 \times (5)^2 + 150\,(6)^2 + 100\,(44 - 50)^2 + 150\,(44 - 40)^2}{100 + 150}}$$

$$= 7.46.$$

NOTES

Example 2.12: A distribution consists of three components with 200, 250, 300 items having mean 25, 10 and 15 and standard deviation 3, 4 and 5, respectively. Find the standard deviation of the combined distribution.

Solution:

In the usual notations, we are given here

$$N_1 = 200, N_2 = 250, N_3 = 300$$

$$\bar{x}_1 = 25, \bar{x}_2 = 10, \bar{x}_3 = 15$$

For the combination of three series the formula will be:

$$\bar{x} = \frac{N_1\bar{x}_1 + N_2\bar{x}_2 + N_3\bar{x}_3}{N_1 + N_2 + N_3}$$

$$= \frac{200 \times 25 + 250 \times 10 + 300 \times 15}{200 + 250 + 300}$$

$$= \frac{12000}{750} = 16$$

and,

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_3\sigma_3^2 + N_1(\bar{x} - \bar{x}_1)^2 + N_2\,(\bar{x} - \bar{x}_2)^2 + N_3\,(\bar{x} - \bar{x}_3)^2}{N_1 + N_2 + N_3}}$$

$$= \sqrt{\frac{200 \times 9 + 250 \times 16 + 300 \times 25 + 200 \times 81 + 250 \times 36 + 300 \times 1}{200 + 250 + 300}}$$

$$= \sqrt{51.73} = 7.19$$

2.3.4 Range (R)

Range (R) of a set of data is the difference between the largest and smallest values. It is the simplest measure of variability or dispersion and is calculated by subtracting the lowest score in the series from the highest. But it is very rough measure of the variability of series. It takes only extreme scores into consideration and ignores the variation of individual items.

Thus, the crudest measure of dispersion is the range of the distribution. The range of any series is the difference between the highest and the lowest values in the series. If the marks received in an examination taken by 248 students are arranged in ascending order, then the range will be equal to the difference between the highest and the lowest marks.

In a frequency distribution, the range is taken to be the difference between the lower limit of the class at the lower extreme of the distribution and the upper limit of the class at the upper extreme.

Consider the data on weekly earnings of worker on four workshops given in Table 2.1.

From these figure in Table 2.1, it is clear that the greater the range, the greater is the variation of the values in the group.

The range is a measure of absolute dispersion and as such cannot be usefully employed for comparing the variability of two distributions expressed in different units. The amount of dispersion measured, say, in pounds, is not comparable with dispersion measured in inches. Thus, the need of measuring relative dispersion arises.

An absolute measure can be converted into a relative measure if we divide it by some other value regarded as standard for the purpose. We may use the mean of the distribution or any other positional average as the standard.

For Table 2.1, the relative dispersion would be,

$$\text{Workshop } A = \frac{9}{25.5} \qquad \text{Workshop } C \quad = \frac{23}{25.5}$$

$$\text{Workshop } B = \frac{15}{25.5} \qquad \text{Workshop } D \quad = \frac{15}{25.5}$$

An alternate method of converting an absolute variation into a relative one would be to use the total of the extremes as the standard. This will be equal to dividing the difference of the extreme items by the total of the extreme items. Thus,

$$\text{Relative Dispersion} = \frac{\text{Difference of extreme items, i.e, Range}}{\text{Sum of extreme items}}$$

The relative dispersion of the series is called the coefficient or ratio of dispersion. In our example of weekly earnings of workers considered earlier, the coefficients would be,

$$\text{Workshop } A = \frac{9}{21+30} = \frac{9}{51} \qquad \text{Workshop } B = \frac{15}{17+32} = \frac{15}{49}$$

$$\text{Workshop } C = \frac{23}{15+38} = \frac{23}{53} \qquad \text{Workshop } D = \frac{15}{19+34} = \frac{15}{53}$$

Merits and Limitations of Range

Merits

Of the various characteristics that a good measure of dispersion should possess, the range has only two, which are as follows:

1. It is easy to understand.
2. Its computation is simple.

Limitations

Besides the aforesaid two qualities, the range does not satisfy the other test of a good measure and hence it is often termed as a crude measure of dispersion.

The following are the limitations that are inherent in the range as a concept of variability:

1. Since it is based upon two extreme cases in the entire distribution, the range may be considerably changed if either of the extreme cases happens to drop out, while the removal of any other case would not affect it at all.
2. It does not tell anything about the distribution of values in the series relative to a measure of central tendency.
3. It cannot be computed when distribution has open-end classes.
4. It does not take into account the entire data. These can be illustrated by the following illustration. Consider the data given in Table 2.3.

Table 2.3 *Distribution with the Same Number of Cases, but Different Variability*

Class	*No. of Students*		
	Section A	*Section B*	*Section C*
0–10	...	...	...
10–20	1	...	...
20–30	12	12	19
30–40	17	20	18
40–50	29	35	16
50–60	18	25	18
60–70	16	10	18
70–80	6	8	21
80–90	11	...	...
90–100	...	...	...
Total	110	110	110
Range	80	60	60

The table is designed to illustrate three distributions with the same number of cases but different variability. The removal of two extreme students from Section *A* would make its range equal to that of *B* or *C*.

The greater range of *A* is not a description of the entire group of 110 students, but of the two most extreme students only. Further, though sections *B* and *C* have the same range, the students in Section *B* cluster more closely around the central tendency of the group than they do in Section *C*. Thus, the range fails to reveal the greater homogeneity of *B* or the greater dispersion of *C*. Due to this defect, it is seldom used as a measure of dispersion.

Specific Uses of Range

In spite of the numerous limitations of the range as a measure of dispersion, it is the most appropriate under the following circumstances:

(a) In situations where the extremes involve some hazard for which preparation should be made, it may be more important to know the most extreme cases to be encountered than to know anything else about the distribution. For example, an explorer, would like to know the lowest and the highest temperatures on record in the region he is about to enter; or an engineer would like to know the maximum rainfall during 24 hours for the construction of a storm water drain.

(b) In the study of prices of securities, range has a special field of activity. Thus, to highlight fluctuations in the prices of shares or bullion, it is a common practice to indicate the range over which the prices have moved during a certain period of time. This information, besides being of use to the operators, gives an indication of the stability of the bullion market, or that of the investment climate.

(c) In statistical quality control, range is used as a measure of variation. We, for example, determine the range over which variations in quality are due to random causes, which is made the basis for the fixation of control limits.

2.3.5 Skewness

Skewness refers to lack of symmetry in a distribution. In a symmetrical distribution, the mean, median and mode coincide. Positive and negative skewness is shown in Figure 2.1.

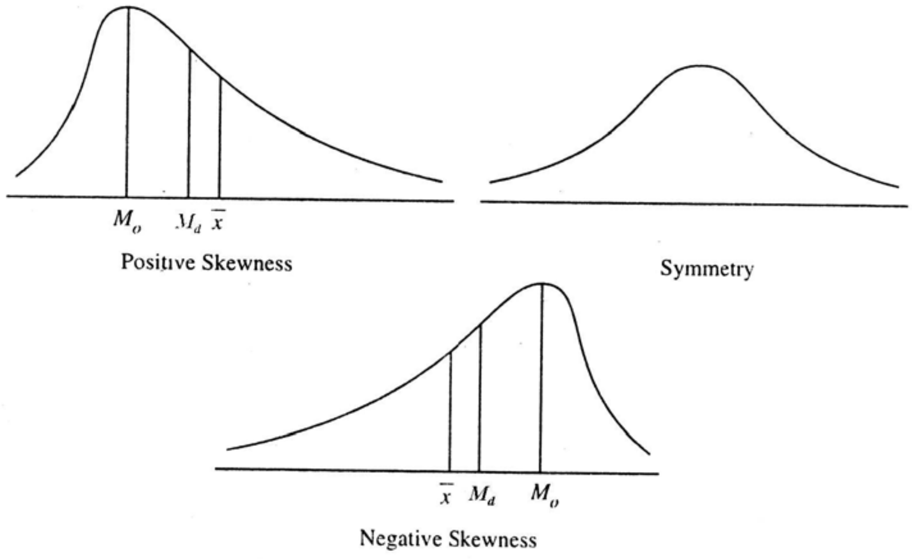


Fig. 2.1 Skewness

In a positively skewed distribution, the longer tail is on the right side and the mean is on the right of the median.

In a negatively skewed distribution, the longer tail is on the left and the mean is on the left of the median.

In a skewed distribution, the distance between the mean and the median is nearly one-third of that between the mean and the mode.

How to Check the Presence of Skewness in a Distribution

In the following cases skewness is present in the data:

(a) The graph is not symmetrical.

(b) The mean, median and mode do not coincide.

(c) The quartiles are not equidistant from the mean.

(d) The sum of positive and negative deviations from the median is not zero.

(e) Frequencies are not similarly distributed on either side of the mode.

Measure of Skewness

A measure of skewness gives a numerical expression and the direction of asymmetry in a distribution. It gives information about the shape of the distribution and the degree of variation on either side of the central value.

We consider some relative measures of skewness that are as follows:

(a) Pearson's Coefficient of Skewness

$$PSk = \frac{\bar{x} - M_O}{s} = \frac{3(\bar{x} - M_d)}{s}$$

It may have any value, but usually it lies between –1 and +1.

Illustration 1**:** If for a given data it is found that:

$\bar{x} = 10$, Mode = 8, s = 4, we have:

$$PSk = \frac{\bar{x} - M_O}{s} = \frac{10 - 8}{4} = 0.5$$

(***b*****)** ***Bowley's Coefficient of Skewness***

$$BSk = \frac{Q_3 - Q_1 - 2M_d}{Q_3 - Q_1}$$

Its value lies between –1 and +1.

Illustration 2: If for a given data $Q_1 = 2,\ Q_3 = 8,\ M_d = 5$

$$BSk = \frac{Q_3 + Q_1 - 2M_d}{Q_3 - Q_1} = \frac{8 + 2 - 5}{8 - 2} = 0.83$$

(***c*****)** ***Kelley's Coefficient of Skewness***

$$KSk = P_{50} - \frac{1}{2}(P_{10} + P_{90})$$

where P_{10}, P_{50} and P_{90} are the 10th, 50th and 90th percentiles of the data.

(***d*****)** ***Method of Moments***

If $\mu_2,\ \mu_3$ are moments about the mean we have the coefficient of skewness:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \mu_3^2 / \sigma^6$$

Sometimes, we define the coefficient of skewness as follows:

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\sigma^3}$$

NOTES

2.3.6 Kurtosis

Kurtosis is a measure of peakedness of a distribution. It shows the degree of convexity of a frequency curve.

If the normal curve is taken as the standard, symmetrical and bell-shaped curve then kurtosis gives a measure of departure from the normal convexity of a distribution. The normal curve is mesokurtic. It is of intermediate peakedness. The flat-topped curve, broader than the normal, is termed platykurtic. The slender, highly peaked curve is termed leptokurtic.

Measures of Kurtosis

(a) Moment Coefficient of Kurtosis : $\beta_2 = \frac{\mu_4}{\mu_2^2}$

Instead of β_2 we often use $\gamma_2 = \beta_2 - 3$ which is positive for a leptokurtic distribution, negative for a platykurtic distribution and zero for the normal distribution.

(b) Percentile Coefficient of Kurtosis $k = \frac{Q}{P_{90} - P_{10}}$, where $Q = \frac{1}{2}(Q_3 - Q_1)$ is the semi-interquartile range.

2.3.7 Comparison of Various Measures of Dispersion

The range is the easiest to calculate the measure of dispersion, but since it depends on extreme values, it is extremely sensitive to the size of the sample and to the sample variability. In fact, as the sample size increases the range increases dramatically, because the more the items one considers, the more likely it is that one item will turn up which is larger than the previous maximum or smaller than the previous minimum. So, it is, in general, impossible to interpret properly the significance of a given range unless the sample size is constant. It is for this reason that there appears to be only one valid application of the range, namely in statistical quality control where the same sample size is repeatedly used so that comparison of ranges is not distorted by differences in sample size.

The quartile deviations and other such positional measures of dispersions are also easy to calculate, but suffer from the disadvantage that they are not amenable to algebraic treatment. Similarly, the mean deviation is not suitable because we cannot obtain the mean deviation of a combined series from the deviations of component series. However, it is easy to interpret and easier to calculate than the standard deviation.

The standard deviation of a set of data, on the other hand, is one of the most important statistics describing it. It lends itself to rigorous algebraic treatment, is rigidly defined and is based on all observations. It is, therefore, quite insensitive to sample size (provided the size is 'large enough') and is least affected by sampling variations.

It is used extensively in testing of hypothesis about population parameters based on sampling statistics.

In fact, the standard deviations has such stable mathematical properties that it is used as a standard scale for measuring deviations from the mean. If we are told that the performance of an individual is 10 points better than the mean, it really does not tell us enough, for 10 points may or may not be a large enough difference to be of significance. However, if we know that the *s* for the score is only 4 points, so that on this scale, the performance is 1.5*s* better than the mean, the statement becomes meaningful. This indicates an extremely good performance. This sigma scale is a very commonly used scale for measuring and specifying deviations which immediately suggest the significance of the deviation.

The only disadvantages of the standard deviation lies in the amount of work involved in its calculation, and the large weight it attaches to extreme values because of the process of squaring involved in its calculations.

Solved Problems on Measures of Dispersion

Problem 1: Find out the range and its coefficient in the following series (individual series):

96, 180, 98, 75, 270, 80, 102, 100, 94.

Solution: Here, L = Largest value of the items = 270

and S = Smallest value of the items = 75

$\therefore$ Range (R) = (L – S) = (270 – 75) = 195

$$\text{and coefficient of range} = \frac{(L-S)}{(L+S)} = \frac{(270-75)}{(270+75)} = 0.56$$

Problem 2: Find out the range and its coefficient in the following (discrete) series:

Monthly Average (in ₹)	100	150	200	250	300	500
No. of Labourers	30	20	15	10	4	1

Solution: Here, L = Largest value of the items = 500

and S = Smallest value of the items = 100

$\therefore$ Range (R) = (L – S) = (500 – 100) = 400

$$\text{and coefficient of range} = \frac{(L-S)}{(L+S)} = \frac{(500-100)}{(500+100)} = 0.66$$

Problem 3: Find out the range and its coefficient in the following (continuous) series:

Size	10–20	20–30	30–40	40–50	50–60
Frequency	8	15	20	5	3

Solution: Here, L = Largest value of the items = 60

and S = Smallest value of the items = 10

$\therefore$ Range (R) = (L – S) = (60 – 10) = 50

and coefficient of range = $\frac{(L-S)}{(L+S)} = \frac{(60-10)}{(60+10)}$ = 0.714

Problem 4: Find the quartile deviation (or semi-interquartile range) and its coefficient from the following data:

Size	1	2	3	4	5	6	7	8	9	10
Frequency	3	2	5	7	9	5	8	10	2	1

Solution:

Size	*Frequency*	*Cumulative Frequency*
1	3	3
2	2	5
3	5	10
4	7	17
5	9	26
6	5	31
7	8	39
8	10	49
9	2	51
10	1	$52 = N = \Sigma f$

Now, lower quartile

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

$$= \text{Size of } \left(\frac{52+1}{4}\right) \text{th item or } 13\frac{1}{4} \text{ th item}$$

$$= \text{13th term} + \frac{1}{4} \text{ (difference between 13th and 14th items)}$$

$$= \left[4 + \frac{1}{4}(4-4)\right], \text{ since } T_{13} = T_{14} = 4$$

And upper quartile,

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

$$= \text{Size of } \frac{3}{4} \text{ (52 + 1)th item or } 39\frac{3}{4}\, th \text{ item}$$

$$= 39th \text{ item} + \frac{3}{4} \text{ (difference between 39th and 40th item)}$$

$$= \left[7 + \frac{3}{4}(8-7)\right] = 7.75$$

∴ Required quartile deviation = ½ $(Q_3 - Q_1)$ = ½ (7.75 – 4)
= 1.725

and coefficient of quartile deviation,

$$= \left(\frac{Q_3 - Q_1}{Q_3 + Q_1}\right) = \left(\frac{7.75 - 4}{7.75 + 4}\right) = 0.32$$

Problem 5: Find the quartile deviation and its coefficient from the following continuous series:

Weight (lb)	*No. of Persons*	*Weight (lb)*	*No. of Persons*
70–80	12	110–120	50
80–90	18	120–130	45
90–100	35	130–140	20
100–110	42	140–150	8

Solution: Here, we have:

Weight (lb)	*Frequency (No. of Persons)*	*Cumulative Frequency*
70–80	12	12
80– 90	18	30
90–100	35	65
100–110	42	107
110–120	50	157
120–130	45	202
130–140	20	222
140–150	8	230 = $N = \Sigma f$
Total	$N = \Sigma f = 230$	

Here $\frac{N}{4} = \frac{1}{4}(230) = 57.5$

∴ Q_1 = 57.5th or 58th item which lies in 90 – 100 group.

$$\therefore \quad Q_1 = \left\{L + \left(\frac{\frac{N}{4} - c.f}{f}\right) \times i\right\} = \left\{90 + \left(\frac{\frac{230}{4} - 30}{35}\right) \times 10\right\} = 97.85$$

Similarly,

$$\frac{3}{4}N = \frac{3}{4}(230) = 172.5$$

∴ Q_3 = 172.5th or 173rd item which lies in 120–30 group.

$$\therefore \quad Q_3 = \left\{L + \left(\frac{\frac{3N}{4} - c.f}{f}\right) \times i\right\} = \left\{120 + \left(\frac{\frac{3}{4}(230) - 157}{45}\right) \times 10\right\} = 123.22$$

Hence, the quartile deviation, Q = ½ $(Q_3 - Q_1)$
= ½ (123.22 – 97.85) = 12.685

Also, coefficient of quartile deviation,

$$= \left(\frac{Q_3 - Q_1}{Q_3 + Q_1}\right) = \left(\frac{123.22 - 97.85}{123.22 + 97.85}\right) = 0.114$$

NOTES

Problem 6: The following data gives the masses (in gm), to the nearest gramme (a metric unit of mass), of a sample of 20 eggs:

46, 51, 48, 62, 54, 51, 58, 60, 71, 75, 47, 73, 62, 65, 53, 57, 65, 72, 49, 51.

Calculate the mean deviation from the arithmetic mean of the masses of this sample.

Solution: Here, N = No. of items = 20.

Arithmetic mean,

$$\bar{X} = \frac{\Sigma X}{N} = \frac{(46+51+48+...+49+51)}{20}$$

$$= \frac{1170}{20} = 58.5 \text{ gm} = M$$

Writing in tabular form, we have:

X	$X - M$	$\lvert X - M \rvert$
46	– 12.5	12.5
51	– 7.5	7.5
48	– 10.5	10.5
62	+ 3.5	3.5
54	– 4.5	4.5
51	– 7.5	7.5
58	– 0.5	0.5
60	+ 1.5	1.5
71	+ 12.5	12.5
75	+ 16.5	16.5
47	– 11.5	11.5
73	+ 14.5	14.5
62	+ 3.5	3.5
65	+ 6.5	6.5
53	– 5.5	5.5
57	– 1.5	1.5
65	+ 6.5	6.5
72	+ 13.5	13.5
49	– 9.5	9.5
51	– 7.5	7.5
Total		157.0 = $\Sigma\lvert X - M \rvert$

The required mean deviation $= \frac{\Sigma \lvert X - M \rvert}{N}$

$$= \frac{157.0}{20} \text{gm} = 7.850 \text{ gm}$$

Problem 7: The monthly incomes of five labourers (in thousand of ₹) are given as 30, 40, 45, 50, 55. Find the deviation from the median.

Solution: N = No. of items = 5

The incomes were already in ascending order.

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

$$= \text{Size of 3rd item} = 45 \text{ ₹} = M$$

Writing in the tabular form, we have:

X	*(X – M) = (X – 45)*	*|X – M|*
30	– 15	15
40	– 5	5
45	0	0
50	+ 5	5
55	+ 10	10
Total	Σ|*X* – *M*|	= 35

$$\therefore \text{ The required mean deviation} = \frac{\Sigma|X-M|}{N} = \frac{35}{5} = 7$$

Problem 8: Calculate the mean deviation from the mean for the following data giving the neck circumference distribution of a typical group of students.

Mid-Value	30	31.5	33	34.5	36	37.5	39	40.5
No. of Students	4	19	30	63	66	29	18	1

Solution: Writing in tabular form, we have:

Mid-Value (X)	*No. of Students (f)*	*d = (X – A) = (X – 36)*	*fd = f(X – A)*
30	4	– 6.0	– 24.0
31.5	19	– 4.5	– 85.5
33	30	– 3.0	– 90.0
34.5	63	– 1.5	– 94.5
36	66	0	0.0
37.5	29	+ 1.5	43.5
39	18	+ 3.0	54.0
40.5	1	+ 4.5	4.5
Total	Σ*f* = 230 = *N*		Σ*fd* = – 192

(Assuming $A = 36$)

$$\text{Arithmetic mean} = \left\{A + \left(\frac{\Sigma fd}{N}\right)\right\}$$

$$= \left\{36 + \left(\frac{-192}{230}\right)\right\} = 35 \text{ cm} = M \text{ (say)}$$

Now, the calculation for mean deviation are shown in the following table:

x	*f*	$|X - M| = |X - 35|$	$f.|X - M|$
30	4	5	20
31.5	19	3.5	66.5
33	30	2	60
34.5	63	0.5	31.5
36	66	1	66
37.5	29	2.5	72.5
39	18	4	72
40.5	1	5.5	5.5
	$N = 230$ $(= \Sigma f)$		361 $(= \Sigma f|x - M|)$

$$\therefore \text{Required mean deviation} = \frac{\Sigma f \mid X - M \mid}{N} = \frac{361}{230} = 1.57$$

Problem 9: Calculate the mean deviation from the median of the following series:

Size	4	6	8	10	12	14	16
Frequency	2	4	5	3	2	1	4

Solution: $N = \Sigma f = 21$

$$\text{Median, } M = \text{Size of } \left(\frac{N+1}{2}\right) \text{th item} = \text{Size of 11th item} = 8$$

Size *(X)*	*Frequency* *(f)*	*Cumulative* *Frequency*	$|X – M|$ $= |X - 8|$	$f.|X – M|$
4	2	2	4	8
6	4	6	2	8
8	5	11	0	0
10	3	14	2	6
12	2	16	4	8
14	1	17	6	6
16	4	21	8	32
	$N = \Sigma f = 21$			$\Sigma f. |X - M| = 68$

$$\therefore \text{Required mean deviation} = \frac{\Sigma f. \mid X - M \mid}{N}$$

$$= \frac{68}{21} = 3.24$$

Problem 10: Calculate the mean deviation from mean for the following table:

Marks	0–10	10–20	20–30	30–40	40–50
No. of Students	5	8	15	16	6

Solution: Let *A* = 25 (Assumed mean)

Class	*Frequency* *(f)*	*Mid-Value* *(X)*	$d=\left(\frac{X-A}{i}\right)$ $\left(\frac{X-25}{10}\right)$	*fd*	*|X–M|* *=|X–27|*	*f.|X–M|*
0–10	5	5	– 2	– 10	22	110
10–20	8	15	– 1	– 8	12	96
20–30	15	25	0	0	2	30
30–40	16	35	1	16	8	128
40–50	6	45	2	12	18	108
	N = Σ*f* = 50			Σ*fd* = 10		Σ*f.|X – M|* = 472

$$\text{Arithmetic mean, } M=\left\{A+\left\{\frac{\Sigma fd}{N}\right\}\times i\right\}$$

$$=\left\{25+\left\{\frac{10}{50}\right\}\times 10\right\}$$

$$\therefore \text{ Required mean deviation} = \frac{\Sigma f.|X-M|}{N}$$

$$=\frac{472}{50}=9.44$$

(Also coefficient of mean deviation

$$=\frac{\text{Mean deviation}}{\text{Mean}}$$

$$=\frac{9.44}{0.27}=0.35)$$

Problem 11: Find the mean deviation from the mean of the following series:

Age Under	10	20	30	40	50	60	70	80
No. of Persons	15	30	53	75	100	110	115	125

Solution: Writing the given data in the tabular form, we have:

Class	0–10	10–20	20–30	30–40	40–50	50–60	60–70	70–80
Cumulative Frequency	15	30	53	75	100	110	115	125
Frequency	15	15 (30–15)	23 (53–30)	22 (75–53)	25 (100–75)	10 (110–100)	5 (115–110)	10 (125–115)

NOTES

The calculations are shown in the following table:

Class	*Frequency (f)*	*Mid-Value (X)*	$d = \left(\frac{X - A}{i}\right)$ $= \left(\frac{X - 45}{10}\right)$	*fd*	*|X–M|* *=|X–35.16|*	*f.|X–M|*
0–10	15	5	– 4	– 60	+ 30.16	452.40
10–20	15	15	– 3	– 45	+ 20.16	302.40
20–30	23	25	– 2	– 46	+ 10.16	233.68
30–40	22	35	– 1	– 22	+ 0.16	3.52
40–50	25	45	0	0	+ 9.84	246.00
50–60	10	55	+ 1	+ 10	+ 19.84	190.84
60–70	5	65	+ 2	+ 10	+ 29.84	149.20
70–80	10	75	+ 3	+ 30	+ 39.84	398.40
Total	$N = \Sigma f = 125$			$\Sigma fd =$ $= -123$		$\Sigma f\|X–M\|=$ $= 1976.44$

Let the assumed arithmetic mean, $A = 45$

Arithmetic (Actual) mean,

$$M = \left\{A + \left(\frac{\Sigma fd}{n}\right) \times i\right\}$$

$$= \left\{45 + \left(\frac{-123}{125}\right) \times 10\right\} = 35.16$$

$\therefore$ Required mean deviation $= \dfrac{\Sigma f.|X - M|}{N} = \dfrac{1976.44}{125} = 15.8$

Problem 12: Calculate the standard deviation and its coefficient from the marks 5, 7, 9, 11 obtained by four students A, B, C, D.

Solution:

Name of the Student	*Marks (X)*	$(X - \bar{X})$ $= (X - 8)$	$(X - \bar{X})^2$ $= (X - 8)^2$
A	5	– 3	9
B	7	– 1	1
C	9	+ 1	1
D	11	+ 3	9
$N = 4$	$\Sigma X = 32$	$\Sigma(X - \bar{X})^2$	$= 20$

Arithmetic mean, $\bar{X} = \dfrac{\Sigma X}{N} = \dfrac{32}{4} = 8$

Standard deviation, $\sigma = \sqrt{\dfrac{\Sigma(X - \bar{X})^2}{N}} = \sqrt{\dfrac{20}{4}} = \sqrt{5} = 2.23$

Also, coefficient of standard deviation,

$$= \left(\frac{\sigma}{\bar{X}}\right) = \frac{2.23}{8} = 0.28$$

Problem 13: Calculate the standard deviation for the following series:

X	0	10	20	30	40	50	60	70	80
F	150	140	100	80	80	70	30	14	0

Solution: Here, we have:

X	*f*	*fX*	$(X-\bar{X})$ $= (X - 23)$	$(X-\bar{X})^2$ $= (X - 23)^2$	$f(X-\bar{X})^2$ $f(X - 23)^2$
0	150	0	– 23	529	79350
10	140	1400	– 13	169	23660
20	100	2000	– 3	9	900
30	80	2400	7	49	3920
40	80	3200	17	289	23120
50	70	3500	27	729	51030
60	30	1800	37	1369	41070
70	14	980	47	2209	30926
80	0	0	57	3249	0
	$N = \Sigma f$ $= 664$	$\Sigma fX = 15{,}280$			$\Sigma f(X-\bar{X})^2$ $= 253976$

Here,

$$\bar{X} = \frac{\Sigma fX}{N}$$

$$= \left(\frac{15280}{664}\right) \approx 23$$

$$\therefore \sigma = \sqrt{\frac{\Sigma f(X-\bar{X})^2}{N}}$$

$$= \sqrt{\frac{253976}{664}} = 19.557$$

Problem 14: Calculate the standard deviation of the following:

Class	5–10	10–15	15–20	20–25	25–30	30–35	35–40	40–45
Frequency	6	5	15	10	5	4	3	2

NOTES

Solution: Here, we have:

Class	*Mid-Value* *(X)*	*Frequency* *(f)*	$\left(X-\bar{X}\right)$	$\left(X-\bar{X}\right)^2$	$f\left(X-\bar{X}\right)^2$
5–10	7.5	6	– 13.7	187.69	1126.14
10–15	12.5	5	– 8.7	75.69	378.45
15–20	17.5	15	– 3.7	13.69	205.35
20–25	22.5	10	1.3	1.69	16.90
25–30	27.5	5	6.3	39.69	198.45
30–35	32.5	4	11.3	127.69	510.76
35–40	37.5	3	16.3	265.69	797.07
40–45	42.5	2	21.3	453.69	907.38
Total		$N = \Sigma f = 50$			$\Sigma f\left(X-\bar{X}\right)^2$ = 4140.50

Here, $\bar{X} = \dfrac{\Sigma fX}{N} = \left\{\dfrac{7.5(6)+12.5(5)+...+42.5(2)}{50}\right\}$

$$\bar{X} = \frac{1060}{50} = 21.2$$

$$\sigma = \sqrt{\frac{\Sigma f\left(X-\bar{X}\right)^2}{N}} = \sqrt{\frac{4140.50}{50}} = 9.1$$

$$\text{Coefficient of S.D.} = \frac{\sigma}{\bar{X}} = \frac{9.1}{21.2} = 0.429$$

Problem 15: Calculate the standard deviation and its coefficient from the following table by short-cut method:

Class	20 – 25	25 – 30	30 – 35	35 – 40	40 – 45	45 – 50
Frequency	18	44	102	160	57	91

Solution: Let assumed mean $A = 32.5$

Mid-Value (X)	*Frequency (f)*	*(X –A) = d*	*d^2*	*fd*	*fd^2*
22.5	*18*	*– 10*	*100*	*– 180*	*1800*
27.5	44	– 5	25	– 220	1100
32.5	102	0	0	0	0
37.5	160	5	25	800	4000
42.5	57	10	100	570	5700
47.5	91	15	225	285	4275
Total	$N = 400$			Σfd = 1255	Σfd^2 = 16875

NOTES

$$\therefore \text{ Required S.D., } \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

$$= \sqrt{\frac{16875}{400} - \left(\frac{1255}{400}\right)^2} = 5.687$$

Also, arithmetic mean,

$$\bar{X} = \left\{A + \left(\frac{\Sigma fd}{N}\right)\right\} = \left\{32.5 + \left(\frac{1255}{400}\right)\right\} = 35.638$$

$$\text{Coefficient of S.D.} = \left(\frac{\sigma}{\bar{X}}\right) = \frac{5.687}{35.638} = 0.159$$

Problem 16: Calculate the S.D. and its coefficient from the following table by step deviation method:

Class	20–30	30–40	40–50	50–60	60–70	70–80	80–90
Frequency	3	61	132	153	140	51	2

Solution: Let us assume mean, *A* = 55, then we get:

Mid-Value	*Frequency*	*(X – A)*	$d = \left(\frac{X-A}{i}\right)$	*fd*	*fd*²
	(X)	*(f)*	*= (X – 55)*	*= (X – 55)/10*	
25	3	– 30	– 3	– 9	27
35	61	– 20	– 2	– 122	244
45	132	– 10	– 1	– 132	132
55	153	0	0	0	0
65	140	10	1	140	140
75	51	20	2	102	204
85	2	30	3	6	18
Total	*N* = Σ*f* = 542			Σ*fd* = – 15	Σ*fd*² = 765

$$\therefore \text{ Required S.D., } \sigma = i \times \sqrt{\left\{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2\right\}}$$

$$= 10 \times \sqrt{\left\{\frac{765}{542} - \left(\frac{-15}{542}\right)^2\right\}} = 11.84$$

Also, the arithmetic mean,

$$\bar{X} = \left\{A + \left(\frac{\Sigma fd}{N}\right) \times i\right\}$$

$$= \left\{55 + \left(\frac{-15}{542}\right) \times 10\right\} = 54.72$$

$$\text{Coefficient of S.D.} = \frac{\sigma}{\bar{X}} = \frac{11.84}{54.72} = 0.216$$

Problem 17: The following results were obtained on the basis of runs scored by two players *A* and *B* in 10 matches. Who is more consistent player?

	Player A	*Player B*
Average Runs	44.30	62.70
Standard Deviation	4.21	9.83

Solution: For Player *A*:

Coefficient of Variation $(\text{C.V}_A) = \left(\frac{\sigma}{\overline{X}}\right)\times 100 = \left(\frac{4.21}{44.30}\right)\times 100 = 9.503$

For Player *B*:

Coefficient of Variation $(\text{C.V}_B) = \left(\frac{\sigma}{\overline{X}}\right)\times 100 = \left(\frac{9.83}{62.70}\right)\times 100 = 15.678$

As $\text{C.V}_A < \text{C.V}_B$, Player *A* is more consistent than Player *B*.

Problem 18: The mean weight of 150 students is 60 kg. The mean weight of boys is 70 kg with a standard deviation of 10 kg. For the girls, the mean weight is 55 kg and the standard deviation is 15 kg. Find the no. of boys and the combined standard deviation.

Solution: (i) $\overline{X}_{12} = \left(\frac{N_1\overline{X}_1 + N_2\overline{X}_2}{N_1 + N_2}\right)$

Given $\overline{X}_{12} = 60, \overline{X}_1 = 70, \overline{X}_2 = 55, N_1 + N_2 = 150$

We have to determine the no. of boys.

Let $N_1 =$ No. of boys.

$N_2 =$ No. of girls $= (150 - N_1)$

On substitution, we have:

$$60 = \left\{\frac{N_1(70)+(150-N_1)55}{150}\right\}$$

Transposing, $N_1 = 100 \Rightarrow N_2 = 50$

(ii) Combined standard deviation:

$$\sigma_{12} = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + N_1d_1^2 + N_2d_2^2}{N_1 + N_2}}$$

$N_1 = 50, \sigma_1 = 10,$

$N_2 = 100, \sigma_2 = 15$

$d_1 = |\overline{X}_1 - \overline{X}_{12}| = |70-60| = 10$

$d_2 = |\overline{X}_2 - \overline{X}_{12}| = |55-60| = 5$

$$\therefore \qquad \sigma_{12} = \sqrt{\frac{50(10)^2+100(15)^2+50(10)^2+100(5)^2}{50+100}}$$

$$\sigma_{12} = 15.28$$

NOTES

Problem 19: Find the missing information from the following table:

	Group I	*Group II*	*Group III*	*Combined*
Number	50	?	90	200
Stanadrad Deviation	6	7	?	7.746
Mean	113	?	115	116

Solution: Let N_1, N_2, N_3 represent the no. of observations in 1st, 2nd and 3rd groups, respectively.

$$(N_1 + N_2 + N_3) = 200$$

$$\Rightarrow \quad (50 + N_2 + 90) = 200$$

$$\Rightarrow \qquad N_2 = 60$$

Combined mean:

$$\bar{X}_{123} = \left(\frac{N_1\bar{X}_1+N_2\bar{X}_2+N_3\bar{X}_3}{N_1+N_2+N_3}\right)$$

$$\Rightarrow \qquad 116 = \left\{\frac{50(113)+60\bar{X}_2+90(115)}{50+60+90}\right\}$$

$$\Rightarrow \qquad \bar{X}_2 = 120 \text{ (On transposing)}$$

Combined standard deviation:

$$\sigma_{123} = \sqrt{\frac{N_1\sigma_1^2+N_2\sigma_2^2+N_3\sigma_3^2+N_1d_1^2+N_2d_2^2+N_3d_3^2}{N_1+N_2+N_3}}$$

$\sigma_{123} = 7.746, \quad d_1 = \left|\bar{X}_1-\bar{X}_{123}\right| = |113-116| = 3$

$\sigma_1 = 6 \quad d_2 = \left|\bar{X}_2-\bar{X}_{123}\right| = |120-116| = 4$

$\sigma_2 = 7 \quad d_3 = \left|\bar{X}_3-\bar{X}_{123}\right| = |115-116| = 1$

$\sigma_3 = ?$

$$\text{Thus, } \sigma_{123} = \sqrt{\frac{50(6)^2+60(7)^2+90\sigma_3^2+50(3)^2+60(4)^2+90(1)^2}{50+60+90}}$$

$$= 7.746$$

$$\Rightarrow \qquad \sigma_3 = 8 \text{ (After transposing)}$$

2.3.8 Coefficient of Variation

The square of standard deviation, namely σ^2, is termed as variance and is more often specified than the standard deviation. Clearly, it has the same properties as standard deviation.

As is clear, the standard deviation σ or its square, the variance, cannot be very useful in comparing two series where either the units are different or the mean values are different. Thus, a σ of 5 on an examination where the mean score is 30 has an altogether different meaning than on an examination where the mean score is 90. Clearly, the variability in the second examination is much less. To take care of this problem, we define and use a coefficient of variation, *V*. Where,

$$V = \frac{\sigma}{\bar{x}} \times 100$$

This is expressed as percentage.

Example 2.13: The following are the scores of two batsmen A and B in a series of innings:

A	12	115	6	73	7	19	119	36	84	29
B	47	12	76	42	4	51	37	48	13	0

Who is the better run-getter? Who is more consistent?

Solution: In order to decide as to which of the two batsmen, *A* and *B*, is the better run-getter, we should find their batting averages. The one whose average is higher will be considered as a better batsman.

To determine the consistency in batting we should determine the coefficient of variation. The less this coefficient the more consistent will be the player.

	A			*B*	
Score *x*	*x*	x^2	*Scores* *x*	*x*	x^2
12	–38	1,444	47	14	196
115	+65	4,225	12	–21	441
6	–44	1,936	76	43	1,849
73	+23	529	42	9	81
7	–43	1,849	–4	– 29	841
19	–31	961	51	18	324
119	+69	4,761	37	4	16
36	–14	196	48	15	225
84	+34	1,156	13	–20	400
29	–21	441	0	–33	1,089
$\sum x = 500$		17,498	$\sum x = 330$		5,462

Batsman *A*:

$$\bar{x} = \frac{500}{10} = 50$$

Batsman *B*:

$$\bar{x} = \frac{330}{10} = 33$$

$$\sigma = \sqrt{\frac{17{,}498}{10}} = 41.83 \qquad \sigma = \sqrt{\frac{5{,}462}{10}} = 23.37$$

$$V = \frac{41.83 \times 100}{50} \qquad V = \frac{23.37}{33} \times 100$$

$$= 83.66 \text{ per cent} \qquad = 70.8 \text{ per cent}$$

A is a better batsman since his average is 50 as compared to 33 of *B*, but *B* is more consistent since the variation in his case is 70.8 as compared to 83.66 of *A*.

Example 2.14. The following table gives the age distribution of students admitted to a college in the years 1914 and 1918. Find which of the two groups is more variable in age.

Age	*Number of Students in*	
	1914	*1918*
15	–	1
16	1	6
17	3	34
18	8	22
19	12	35
20	14	20
21	13	7
22	5	19
23	2	3
24	3	–
25	1	–
26	–	–
27	1	–

Solution:

Age	*Assumed Mean–21* *1914*				*Assumed Mean–19* *1918*			
	f	x'	fx'	fx'^2	f	x'	fx	fx'^2
15	0	–6	0	0	1	–4	–4	16
16	1	–5	–5	25	6	–3	–18	54
17	3	–4	–12	48	34	–2	–68	136
18	8	–3	–24	72	22	–1	–22	22
19	12	–2	–24	48			–112	
20	14	–1	–14	14				
			–79		35	0	0	0
21	13	0	0	0	20	1	20	20
22	5	1	5	5	7	2	14	28
23	2	2	4	8	19	3	57	171
24	3	3	9	27	3	4	12	48
25	1	4	4	16	147		+103	495
26	0	5	0	0			–9	
27	1	6	6	36				
	63		+28	299				
			–51					

1914 Group:

$$\sigma = \sqrt{\frac{\sum fx'^2}{N} - \left[\frac{\sum (fx')}{N}\right]^2}$$

$$= \sqrt{\frac{299}{63} - \left(\frac{-51}{63}\right)^2}$$

$$= \sqrt{4.476 - 0.655} = \sqrt{4.091}$$

$$= 2.02.$$

$$\bar{x} = 21 + \left(\frac{-51}{63}\right) = 21 - 8 = 20.2$$

$$V = \frac{2.02}{20.2} \times 100$$

$$= \frac{202}{20.2} = 10$$

1918 Group:

$$\sigma = \sqrt{\frac{495}{147} - \left(\frac{-9}{147}\right)^2} = \sqrt{3.3673 - 0.0037}$$

$$= \sqrt{3.3636} = 1.834$$

$$\bar{x} = 19 + \left(\frac{-9}{147}\right)$$

$$= 19 - .06 = 18.94$$

$$V = \frac{1.834}{18.94} \times 100$$

$$= 9.68$$

The coefficient of variation of the 1914 group is 10 and that of the 1918 group 9.68. This means that the 1914 group is more variable, but only barely so.

Check Your Progress

4. Define measures of dispersion.
5. List any two characteristics which are considered essential for a measure of central tendency.
6. What are the four measures of dispersion?

2.4 SUMMARY

- There are several commonly used measures of central tendency, such as arithmetic mean, mode and median. These values are very useful not only in presenting the overall picture of the entire data but also for the purpose of making comparisons among two or more sets of data.

- Tate in 1955 defines the measures of central tendency as, 'A sort of average or typical value of the items in the series and its function is to summarize the series in terms of this average value.'
- The most common measures of central tendency are: (a) Arithmetic Mean or Mean (b) Median (c) Mode.
- The simplest but most useful measure of central tendency is the arithmetic mean. It can be defined as the sum of all the values of the items in a series divided by the number of items. It is represented by the letter M.
- The median is a measure of central tendency and it appears in the centre of an ordered data. It divides the list of ordered values in the data into two equal parts so that half of the data will have values less than the median and half will have values greater than the median.
- The mode is another form of average and can be defined as the most frequently occurring value in the data. The mode is not affected by extreme values in the data and can easily be obtained from an ordered set of data. It can be useful and more representative of the data under certain conditions and is the only measure of central tendency that can be used for qualitative data.
- The weighted arithmetic mean is particularly useful where we have to compute the mean of means. If we are given two arithmetic means, one for each of two different series, in respect of the same variable, and are required to find the arithmetic mean of the combined series, the weighted arithmetic mean is the only suitable method of its determination.
- A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency. A measure of dispersion may be expressed in an 'absolute form' or in a 'relative form'.
- A measure of dispersion should possess all those characteristics which are considered essential for a measure of central tendency. Some of them include: (a) It should be based on all observations (b) It should be readily comprehensible (c) It should be fairly easily calculated (d) It should be amenable to algebraic treatment
- The four measures of dispersion are: (i) Range (R) (ii) Quartile Deviation (QD) (iii) Average Deviation (AD) (iv) Standard Deviation (SD).
- One of the measures of dispersion is the semi-interquartile range, usually termed as 'Quartile Deviation' or QD. Quartiles are the points which divide the array into four equal parts.
- Garrett in 1971 defines Average Deviation (AD) as the mean of deviations of all the separate scores in the series taken from their mean (occasionally from the median or mode). It is the simplest measure of variability that takes into account the fluctuation or variation of all the items in a series.

- Merits of average (mean) deviation include: (a) Easy to understand (b) Computation is simple as compared to standard deviation (c) It is less affected by extreme values as compared to standard deviation (d) Since it is based on all values in the distribution, it is better than range or quartile deviation.
- Demerits of average (mean) deviation include: (a) It lacks those algebraic properties which would facilitate its computation and establish its relation to other measures (b) Due to this, it is not suitable for further mathematical processing
- By far the most universally used and the most useful measure of dispersion is the Standard Deviation (SD) or root mean square deviation about the mean. Standard Deviation or SD is regarded as the most stable and reliable measure of variability as it employs the mean for its computation. It is often called root mean square deviation and is denoted by the Greek letter sigma (s).
- Range (R) of a set of data is the difference between the largest and smallest values. It is the simplest measure of variability or dispersion and is calculated by subtracting the lowest score in the series from the highest. But it is very rough measure of the variability of series. It takes only extreme scores into consideration and ignores the variation of individual items.
- In a frequency distribution, the range is taken to be the difference between the lower limit of the class at the lower extreme of the distribution and the upper limit of the class at the upper extreme.
- Skewness refers to lack of symmetry in a distribution. In a symmetrical distribution, the mean, median and mode coincide.
- A measure of skewness gives a numerical expression and the direction of asymmetry in a distribution. It gives information about the shape of the distribution and the degree of variation on either side of the central value.
- Kurtosis is a measure of peakedness of a distribution. It shows the degree of convexity of a frequency curve.

2.5 KEY TERMS

- **Arithmetic mean:** It can be defined as the sum of all the values of the items in a series divided by the number of items.
- **Median:** It is a measure of central tendency and it appears in the centre of an ordered data.
- **Mode:** It is another form of average and can be defined as the most frequently occurring value in the data.
- **Measure of dispersion:** It may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency.
- **Skewness:** It refers to lack of symmetry in a distribution.
- **Kurtosis:** It is a measure of peakedness of a distribution. It shows the degree of convexity of a frequency curve.

2.6 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The median is a measure of central tendency and it appears in the centre of an ordered data. It divides the list of ordered values in the data into two equal parts so that half of the data will have values less than the median and half will have values greater than the median.
2. In a frequency distribution where all the frequencies are greater than one, the mean is calculated by the formula:

   M = $\sum fX/N$
3. Two disadvantages of mean are: (a) It is affected by extreme values, and hence, not very reliable when the data set has extreme values especially when these extreme values are on one side of the ordered data. Thus, a mean of such data is not truly a representative of such data. (b) It is tedious to compute for a large data set as every point in the data set is to be used in computations.
4. A measure of dispersion or simply dispersion may be defined as statistics signifying the extent of the scatteredness of items around a measure of central tendency. A measure of dispersion may be expressed in an 'absolute form' or in a 'relative form'.
5. Two characteristics which are considered essential for a measure of central tendency include: (a) It should be based on all observations (b) It should be readily comprehensible
6. The four measures of dispersion are: (i) Range (R) (ii) Quartile Deviation (QD) (iii) Average Deviation (AD) (iv) Standard Deviation (SD).

2.7 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the three interesting properties of arithmetic mean?
2. Write a short note on harmonical progression.
3. What are the various characteristics of quartile deviation?
4. What is average deviation? How is average deviation computed?
5. Discuss the specific uses of range.

Long-Answer Questions

1. Discuss standard deviation. Describe the computation of standard deviation from grouped and ungrouped data.
2. Define range. What are the merits and demerits of range?
3. How is skewness measured in a distribution?
4. Compare the various measures of dispersion.

2.8 FURTHER READING

Lindquist, E. C. 1951. *Education Measurement*. Washington D.C.: The American Council on Education.

Walker, H. M and J. Lev. 1953. *Statistical Inference*. New York: Henry Holt.

Health, R. W. and N. M. Downie. 1970. *Basic Statistical Methods, 3rd Edition*. New York: Harper International.

UNIT 3 CORRELATION AND REGRESSION

NOTES

Structure

3.0 INTRODUCTION

In this unit, you will learn about correlation analysis. This technique looks at the indirect relationships and establishes the variables which are most closely associated with a given data or mind-set. It is the process of finding how accurately the line fits using the observations. Correlation analysis can be referred to as the statistical tool used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. In fact, the word correlation refers to the relationship or the interdependence between two variables. There are various phenomena which have relation to each other. The theory by means of which quantitative connections between two sets of phenomena are determined is called the 'Theory of Correlation'. On the basis of the theory of correlation, you can study the comparative changes occurring in two related phenomena and their cause-effect relation can also be examined. Thus, correlation is concerned with the relationship between two related and quantifiable variables and can be positive or negative.

In this unit, you will also learn about regression analysis. It is the mathematical process of using observations to find the line of best fit through the data in order to make estimates and predictions about the behaviour of the variables. This technique is used to determine the statistical relationship between two or more variables and to make prediction of one variable on the basis of one or more other variables. While making use of the regression techniques for making predictions, it is always assumed that there is an actual relationship between the dependent and independent variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable. You will also learn about the scatter diagram, least squares method and standard error of the estimate. Standard error of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. The larger the standard error of estimate (SEe), the greater happens to be the dispersion, or scattering, of given observations around the regression line. But if the S.E. of estimate happens to be zero then the estimating equation is a 'perfect' estimator, i.e., cent per cent correct estimator of the dependent variable. You will be able to interpret coefficient of determination, i.e., r^2 using the coefficient of correlation.

3.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Discuss correlation and regression analysis
- Explain the calculation of correlation by the product moment method
- Describe coefficient of correlation by least square method and simple regression coefficients
- Discuss multiple regression and assumptions in regression analysis

3.2 CORRELATION AND REGRESSION: AN OVERVIEW

Correlation analysis is the statistical tool generally used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. This analysis is used quite frequently in conjunction with regression analysis to measure how well the regression line explains the variations of the dependent variable. In fact, the word correlation refers to the relationship or interdependence between two variables. There are various phenomenons which have relation to each other. For instance, when demand of a certain commodity increases, then its price goes up and when its demand decreases, its price comes down. Similarly, with age, the height of the children, with height the weight of the children, with money the supply and the general level of prices go up. Such sort of relationship can as well be noticed for several other phenomena. The theory by means of which quantitative connections between two sets of phenomena are determined is called the *Theory of Correlation.*

On the basis of the theory of correlation, one can study the comparative changes occurring in two related phenomena and their cause-effect relation can be examined. It should, however, be borne in mind that relationship like 'black cat causes bad luck', 'filled up pitchers result in good fortune' and similar other beliefs of the people cannot be explained by the theory of correlation, since they are all imaginary and are incapable of being justified mathematically. Thus, correlation is concerned with relationship between two related and quantifiable variables. If two quantities vary in sympathy, so that a movement (an increase or decrease) in one, tends to be accompanied by a movement in the same or opposite direction in the other and the greater the change in the one, the greater is the change in the other, the quantities are said to be correlated. This type of relationship is known as correlation or what is sometimes called, in statistics, as *covariation.*

For correlation, it is essential that the two phenomena should have cause-effect relationship. If such relationship does not exist then one should not talk of correlation. For example, if the height of the students as well as the height of the trees increases, then one should not call it a case of correlation because the two phenomena, viz., the height of students and the height of trees are not even casually related. But, the relationship between the price of a commodity and its demand, the price of a commodity and its supply, the rate of interest and savings, etc. are examples of correlation, since in all such cases the change in one phenomenon is explained by a change in other phenomenon.

It is appropriate here to mention that correlation in case of phenomena pertaining to natural sciences can be reduced to absolute mathematical term, e.g., heat always increases with light. But in phenomena pertaining to social sciences it is often difficult to establish any absolute relationship between two phenomena. Hence, in social sciences, we must take the fact of correlation being established if in a large number of cases, two variables always tend to move in the same or opposite direction.

Correlation can either be positive or it can be negative. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive, but, when the variations in the two variables take place in opposite direction, the correlation is termed as negative. This can be explained as under:

Changes in Independent Variable	*Changes in Dependent Variable*	*Nature of Correlation*
Increase (+)↑	Increase (+)↑	Positive (+)
Decrease (–)↓	Decrease (–)↓	Positive (+)
Increase (+)↑	Decrease (–)↓	Negative (–)
Decrease (–)↓	Increase (+)↑	Negative (–)

Statisticians have developed *two measures for describing the correlation* between two variables, viz., the coefficient of determination and the coefficient of correlation. We now explain, illustrate and interpret the said two coefficients concerning the relationship between two variables as under.

NOTES

3.2.1 Coefficient of Determination

The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say X and Y, one of which happens to be an independent variable and the other being dependent variable. This coefficient is based on the following two kinds of variations:

(a) The variation of the Y values around the fitted regression line viz., $\Sigma\left(Y-\hat{Y}\right)^2$, technically known as the unexplained variation

(b) The variation of the Y values around their own mean viz., $\Sigma\left(Y-\overline{Y}\right)^2$, technically known as the total variation

If we subtract the unexplained variation from the total variation, we obtain what is known as the explained variation, i.e., the variation explained by the line of regression. Thus, Explained Variation = (Total variation) – (Unexplained variation)

$$= \Sigma\left(Y-\overline{Y}\right)^2 - \Sigma\left(Y-\hat{Y}\right)^2$$

$$= \Sigma\left(\hat{Y}-\overline{Y}\right)^2$$

The Total and Explained as well as Unexplained variations can be shown as given in Figure 3.1.

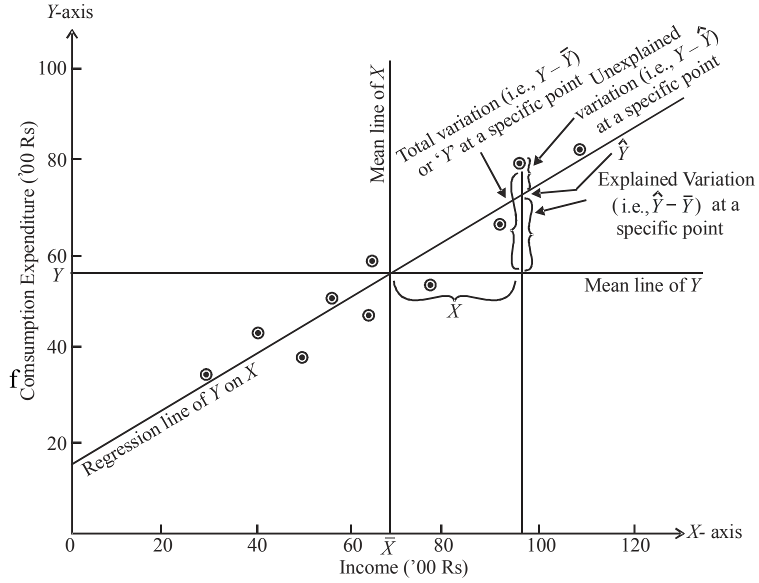


Fig. 3.1 *Diagram Showing Total, Explained and Unexplained Variations*

Coefficients of determination is that fraction of the total variation of *Y* which is explained by the regression line. In other words, coefficient of determination is the ratio of explained variation to total variation in the *Y* variable related to the *X* variable. Coefficient of determination algebraically can be stated as under:

$$r^2 = \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= \frac{\Sigma\left(\hat{Y}-\bar{Y}\right)^2}{\Sigma\left(Y-\bar{Y}\right)^2}$$

Alternatively, r^2 can also be stated as under:

$$r^2 = 1 - \frac{\text{Explained variation}}{\text{Total variation}}$$

$$= 1 - \frac{\Sigma\left(\hat{Y}-\bar{Y}\right)^2}{\Sigma\left(Y-\bar{Y}\right)^2}$$

Interpreting r^2

The coefficient of determination can have a value ranging from zero to one. The value of one can occur only if the unexplained variation is zero, which simply means that all the data points in the Scatter diagram fall exactly on the regression line. For a zero value to occur, $\Sigma(Y-\bar{Y})^2 = \Sigma(Y-\hat{Y})^2$, which simply means that *X* tells us nothing about *Y* and hence there is no regression relationship between *X* and *Y* variables. Values between 0 and 1 indicate the 'Goodness of fit' of the regression line to the sample data. The higher the value of r^2, the better the fit. In other words, the value of r^2 will lie somewhere between 0 and 1. If r^2 has a zero value then it indicates no correlation but if it has a value equal to 1 then it indicates that there is perfect correlation and as such the regression line is a perfect estimator. But in most of the cases, the value of r^2 will lie somewhere between these two extremes of 1 and 0. One should remember that r^2 close to 1 indicates a strong correlation between *X* and *Y* while an r^2 near zero means there is little correlation between these two variables. r^2 value can as well be interpreted by looking at the amount of the variation in *Y*, the dependant variable, that is explained by the regression line. Supposing, we get a value of $r^2 = 0.925$ then this would mean that the variations in independent variable (say *X*) would explain 92.5 per cent of the variation in the dependent variable (say *Y*). If r^2 is close to 1 then it indicates that the regression equation explains most of the variations in the dependent variable.

Example 3.1: Calculate the coefficient of determination (r^2) using data given below. Calculate and analyse the result.

Observations	1	2	3	4	5	6	7	8	9	10
Income (X) ('00 ₹)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (Y) ('00 ₹)	44	60	39	51	80	68	84	34	55	48

Solution: r^2 can be worked out as shown below:

Since, $r^2 = 1 - \dfrac{\text{Unexplained variation}}{\text{Total variation}} = 1 - \dfrac{\Sigma\left(Y - \hat{Y}\right)^2}{\Sigma\left(Y - \overline{Y}\right)^2}$

As, $\Sigma\left(Y - \overline{Y}\right)^2 = \Sigma Y^2 = \left(\Sigma Y^2 - n\overline{Y}^2\right)$, we can write,

$$r^2 = 1 - \frac{\Sigma\left(Y - \hat{Y}\right)^2}{\Sigma Y^2 - n\overline{Y}^2}$$

Calculating and putting the various values, we have the following equation:

$$r^2 = 1 - \frac{260.54}{34223 - 10\left(56.3\right)^2} = 1 - \frac{260.54}{2526.10} = 0.897$$

Analysis of result: The regression equation used to calculate the value of the coefficient of determination (r^2) from the sample data shows that, about 90 per cent of the variations in consumption expenditure can be explained. In other words, it means that the variations in income explain about 90 per cent of variations in consumption expenditure.

Observation	1	2	3	4	5	6	7	8	9	10
Income (X) ('00 ₹)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (Y) ('00 ₹)	44	60	39	51	80	68	84	34	55	48

3.2.2 Coefficient of Correlation: Product Moment Method

The coefficient of correlation, symbolically denoted by 'r', is another important measure to describe how well one variable is explained by another. It measures the degree of relationship between the two casually related variables. The value of this coefficient can never be more than +1 or less than –1. Thus, +1 and –1 are the limits of this coefficient. For a unit change in independent variable, if there happens to be a constant change in the dependent variable in the same direction, then the value of the coefficient will be +1 indicative of the perfect positive correlation; but if such a change occurs in the opposite direction, the value of the coefficient will be –1, indicating the perfect negative correlation. In practical life, the possibility of obtaining either a perfect positive or perfect negative correlation is very remote, particularly in respect of the phenomena concerning social sciences. If the coefficient of correlation has a zero value then it means that there exists no correlation between the variables under study.

There are several methods of finding the coefficient of correlation but the following ones are considered important:

(a) Coefficient of Correlation by the Method of Least Squares

(b) Coefficient of Correlation using Simple Regression Coefficients

(c) Coefficient of Correlation through Product Moment Method or Karl Pearson's Coefficient of Correlation

Whichever of these above mentioned three methods we adopt, we get the same value of *r*.

(a) Coefficient of Correlation by the Method of Least Squares

NOTES

Under this method, first of all, the estimating equation is obtained using least square method of simple regression analysis. The equation is worked out as,

$$\hat{Y} = a + bX_i$$

$$\text{Total variation} = \sum\left(Y - \overline{Y}\right)^2$$

$$\text{Unexplained variation} = \sum\left(Y - \hat{Y}\right)^2$$

$$\text{Explained variation} = \sum\left(\hat{Y} - \overline{Y}\right)^2$$

Then, by applying the following formulae, we can find the value of the coefficient of correlation:

$$r = \sqrt{r^2} = \sqrt{\frac{\text{Explained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\text{Unexplained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \overline{Y}\right)^2}}$$

This clearly shows that *the coefficient of correlation happens to be the squareroot of the coefficient of determination*.

Short-cut formula for finding the value of '*r*' by the method of least squares can be repeated and readily written as follows:

$$r = \sqrt{\frac{a\Sigma Y + b\Sigma XY - n\overline{Y}^2}{\Sigma Y^2 - n\overline{Y}^2}}$$

Where,

a = *Y*-intercept

b = Slope of the estimating equation

X = Values of the independent variable

Y = Values of dependent variable

$\overline{Y}$ = Mean of the observed values of *Y*

n = Number of items in the sample (i.e., pairs of observed data)

The plus (+) or the minus (–) sign of the coefficient of correlation worked out by the method of least squares, is related to the sign of '*b*' in the estimating equation viz., $\hat{Y} = a + bX_i$. If '*b*' has a minus sign, the sign of '*r*' will also be minus but if '*b*' has a plus sign, then the sign of '*r*' will also be plus. The value of '*r*' indicates the degree along with the direction of the relationship between the two variables *X* and *Y*.

NOTES

(b) Coefficient of Correlation using Simple Regression Coefficients

Under this method, the estimating equation of Y and the estimating equation of X is worked out using the method of least squares. From these estimating equations we find the regression coefficient of X on Y, i.e., the slope of the estimating equation of X (symbolically written as b_{XY}) and this happens to be equal to $r\frac{\sigma_X}{\sigma_Y}$ and similarly, we find the regression coefficient of Y on X, i.e., the slope of the estimating equation of Y (symbolically written as b_{YX}) and this happens to be equal to $r\frac{\sigma_Y}{\sigma_X}$. For finding '*r*', the square root of the product of these two regression coefficients are worked out as stated below:[1]

$$\begin{aligned} r &= \sqrt{b_{XY}.b_{YX}} \\ &= \sqrt{r\frac{\sigma_X}{\sigma_Y}.r\frac{\sigma_Y}{\sigma_X}} \\ &= \sqrt{r^2} \;= r \end{aligned}$$

As stated earlier, the sign of '*r*' will depend upon the sign of the regression coefficients. If they have minus sign, then '*r*' will take minus sign but the sign of '*r*' will be plus if regression coefficients have plus sign.

(c) Karl Pearson's Coefficient or Product Moment Method

Karl Pearson's method is the most widely used method of measuring the relationship between two variables. This coefficient is based on the following assumptions:

(*i*) There is a linear relationship between the two variables which means that straight line would be obtained if the observed data are plotted on a graph.

(*ii*) The two variables are casually related which means that one of the variables is independent and the other one is dependent.

(*iii*) A large number of independent causes are operating in both the variables so as to produce a normal distribution.

According to Karl Pearson, '*r*' can be worked out as under:

$$r = \frac{\sum XY}{n\sigma_X \sigma_Y}$$

Where, $X = (X - \bar{X})$

$Y = (Y - \bar{Y})$

σ_X = Standard deviation of X series and is equal to $\sqrt{\frac{\sum X^2}{n}}$

σ_Y = Standard deviation of Y series and is equal to $\sqrt{\frac{\sum Y^2}{n}}$

n = Number of pairs of X and Y observed.

NOTES

A short-cut formula, known as the Product Moment Formula, can be derived from the above stated formula as under:

$$r = \frac{\sum XY}{n\sigma_X \sigma_Y}$$

$$= \frac{\sum XY}{\sqrt{\frac{\sum X^2}{n} \cdot \frac{\sum Y^2}{n}}}$$

$$n = \frac{\sum XY}{\sqrt{\sum X^2 \sum Y^2}}$$

The above formulae are based on obtaining true means (viz. $\bar{X}$ and $\bar{Y}$) first and then doing all other calculations. This happens to be a tedious task, particularly if the true means are in fractions. To avoid difficult calculations, we make use of the assumed means in taking out deviations and doing the related calculations. In such a situation, we can use the following formula for finding the value of '*r*':[2]

(*i*) *In case of ungrouped data:*

$$r = \frac{\frac{\sum dX.dY}{n} - \left(\frac{\sum dX}{n} \cdot \frac{\sum dY}{n}\right)}{\sqrt{\frac{\sum dX^2}{n} - \left(\frac{\sum dX}{n}\right)^2} . \sqrt{\frac{\sum dY^2}{n} - . \left(\frac{\sum dY}{n}\right)^2}}$$

$$= \frac{\sum dX.dY - \left(\frac{\sum dX \times \sum dY}{n}\right)}{\sqrt{\sum dX^2 - \frac{(\sum dX)^2}{n}}\sqrt{\sum dY^2 - \frac{(\sum dY)^2}{n}}}$$

Where, $\sum dX = \sum(X - X_A)$ X_A = Assumed average of X

$\sum dY = \sum(Y - Y_A)$ Y_A = Assumed average of Y

$\sum dX^2 = \sum(X - X_A)^2$

$\sum dY^2 = \sum(Y - Y_A)^2$

$\sum dX \,.\, dY = \sum(X - X_A)\,(Y - Y_A)$

n = Number of pairs of observations of X and Y

(*ii*) *In case of grouped data:*

$$r = \frac{\frac{\sum fdX.dY}{n} - \left(\frac{\sum fdX}{n} \cdot \frac{\sum fdY}{n}\right)}{\sqrt{\frac{\sum fdX^2}{n} - \left(\frac{\sum fdX}{n}\right)^2} \sqrt{\frac{\sum fdY^2}{n} - \left(\frac{\sum fdY}{n}\right)^2}}$$

or
$$r = \frac{\sum fdX.dY - \left(\frac{\sum fdX.\sum fdY}{n}\right)}{\sqrt{\sum fdX^2 - \left(\frac{\sum fdX}{n}\right)^2}\sqrt{\sum fdY^2 - \left(\frac{\sum fdY}{n}\right)^2}}$$

Where, $\sum fdX.dY = 0\sum f\ (X - X_A)\ (Y - Y_A)$

$\sum fdX = \sum f(X - X_A)$

$\sum fdY = \sum f(Y - Y_A)$

$\sum fdY^2 = \sum f(Y - Y_A)^2$

$\sum fdX^2 = \sum f(X - X_A)^2$

n = Number of pairs of observations of X and Y.

3.2.3 Probable Error (P.E.) of the Coefficient of Correlation

Probable Error (P.E.) of r is very useful in interpreting the value of r and is worked out as under for Karl Pearson's coefficient of correlation:

$$\text{P.E.} = 0.6745\frac{1-r^2}{\sqrt{n}}$$

If r is less than its P.E., it is not at all significant. If r is more than P.E., there is correlation. *If r is more than 6 times its P.E. and greater than ± 0.5, then it is considered significant.*

Example 3.2:

From the following data, calculate 'r' between X and Y applying the following three methods:

(a) The method of least squares.

(b) The method based on regression coefficients.

(c) The product moment method of Karl Pearson.

Verify the obtained result of any one method with that of another.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

Solution:

Let us develop the following table for calculating the value of '*r*':

	X	Y	X^2	Y^2	XY
	1	9	1	81	9
	2	8	4	64	16
	3	10	9	100	30
	4	12	16	144	48
	5	11	25	121	55
	6	13	36	169	78
	7	14	49	196	98
	8	16	64	256	128
	9	15	81	225	135
n=9	$\sum X = 45$	$\sum Y = 108$	$\sum X^2 = 285$	$\sum Y^2 = 1356$	$\sum XY = 597$
$\therefore$	$\bar{X} = 5$;	$\bar{Y} = 12$			

(*a*) *Coefficient of correlation by the method of least squares is worked out as under*:

First of all find out the estimating equation,

$$\hat{Y} = a + bX_i$$

Where,
$$b = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$= \frac{597 - 9(5)(12)}{285 - 9(25)} = \frac{597 - 540}{285 - 225} = \frac{57}{60} = 0.95$$

and
$$a = \bar{Y} - b\bar{X}$$

$$= 12 - 0.95(5) = 12 - 4.75 = 7.25$$

Hence,
$$\hat{Y} = 7.25 + 0.95X_i$$

Now '*r*' can be worked out as under by the method of least squares,

$$r = \sqrt{1 - \frac{\text{Unexplained variation}}{\text{Total variation}}}$$

$$= \sqrt{1 - \frac{\Sigma\left(Y - \hat{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}} = \sqrt{\frac{\Sigma\left(\hat{Y} - \bar{Y}\right)^2}{\Sigma\left(Y - \bar{Y}\right)^2}}$$

$$= \sqrt{\frac{a\sum Y + b\sum XY - n\bar{Y}^2}{\sum Y^2 - n\bar{Y}^2}}$$

NOTES

This is as per short-cut formula,

$$r = \sqrt{\frac{7.25(108)+0.95(597)-9(12)^2}{1356-9(12)^2}}$$

$$= \sqrt{\frac{783+567.15-1296}{1356-1296}}$$

$$= \sqrt{\frac{54.15}{60}} = \sqrt{0.9025} = 0.95$$

(*b*) *Coefficient of correlation by the method based on regression coefficients is worked out as under*:

∵ Regression coefficients of Y on X,

i.e., $$b_{YX} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$= \frac{597-9\times5\times12}{285-9(5)^2} = \frac{597-540}{285-225} = \frac{57}{60}$$

Regression coefficient of X on Y,

i.e., $$b_{XY} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum Y^2 - n\overline{Y}^2}$$

$$= \frac{597-9\times5\times12}{1356-9(12)^2} = \frac{597-540}{1356-1296} = \frac{57}{60}$$

Hence, $$r = \sqrt{b_{YX}\,.\,b_{XY}}$$

$$= \sqrt{\frac{57}{60}\times\frac{57}{60}} = \frac{57}{60} = 0.95$$

(*c*) *Coefficient of correlation by the product moment method of Karl Pearson is worked out as under:*

$$r = \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum Y^2 - n\overline{Y}^2}}$$

$$= \frac{597-9(5)(12)}{\sqrt{285-9(5)^2}\sqrt{1356-9(12)^2}}$$

$$= \frac{597-540}{\sqrt{285-225}\sqrt{1356-1296}} = \frac{57}{\sqrt{60}\sqrt{60}} = \frac{57}{60} = 0.95$$

Hence, we get the value of $r = 0.95$. We get the same value applying the other two methods also. Therefore, whichever method we apply, the results will be the same.

3.2.4 Some Other Measures

Two other measures are often talked about along with the coefficients of determinations and that of correlation. These are as follows:

NOTES

(*a*) ***Coefficient of Nondetermination:*** Instead of using coefficient of determination, sometimes coefficient of nondetermination is used. Coefficient of non-determination (denoted by k^2) is the ratio of unexplained variation to total variation in the *Y* variable related to the *X* variable. Algebrically, we can write it as follows:

$$k^2 = \frac{\text{Unexplained variation}}{\text{Total variation}} = \frac{\Sigma\left(Y-\hat{Y}\right)^2}{\Sigma\left(Y-\overline{Y}\right)^2}$$

Concerning the data of Example 3.1 of this unit, coefficient of nondetermination will be calculated as follows:

$$k^2 = \frac{260.54}{2526.10} = 0.103$$

The value of k^2 shows that about 10 per cent of the variation in consumption expenditure remains unexplained by the regression equation we had worked out, viz., $\hat{Y} = 14.000 + 0.616X_i$. In simple terms, this means that variable other than *X* is responsible for 10 per cent of the variations in the dependent variable *Y* in the given case.

Coefficient of nondetermination can as well be worked out as under:

$$k^2 = 1 - r^2$$

Accordingly, for Example 3.1, it will be equal to 1–0.897 = 0.103

Note: *Always remember that* $r^2 + k^2 = 1$.

(*b*) ***Coefficient of Alienation:*** Based on k^2, we can work out one more measure namely the Coefficient of alienation, symbolically written as '*k*'. Thus,

Coefficient of alienation, i.e., 'k' $= \sqrt{k^2}$

Unlike $r + k^2 = 1$, the sum of '*r*' and '*k*' will not be equal to 1 unless one of the two coefficients is 1 and in this case the remaining coefficients must be zero. In all other cases, '*r*' + '*k*' > 1. Coefficient of alienation is not a popular measure from practical point of view and is used very rarely.

Spearman's Rank Correlation

If observations on two variables are given in the form of ranks and not as numerical values, it is possible to compute what is known as rank correlation between the two series.

The rank correlation, written as ρ, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2-1)}$$

NOTES

Here, n is the number of observations and D_i, the positive difference between ranks associated with the individuals i.

Like r, the rank correlation lies between –1 and +1.

Example 3.3: The ranks given by two judges to 10 individuals are as follows:

	Rank given by			
Individual	*Judge I*	*Judge II*	*D*	D^2
	x	y	$= x - y$	
1	1	7	6	36
2	2	5	3	9
3	7	8	1	1
4	9	10	1	1
5	8	9	1	1
6	6	4	2	4
7	4	1	3	9
8	3	6	3	9
9	10	3	7	49
10	5	2	3	9
				$\Sigma D^2 = 128$

Solution: The rank correlation is given by,

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 128}{10^3 - 10} = 1 - 0.776 = 0.224$$

The value of $\rho = 0.224$ shows that the agreement between the judges is not high.

Example 3.4: Consider Example 3.3 and compute r and compare.

Solution: The simple coefficient of correlation r for the previous data is calculated as follows:

x	y	x^2	y^2	xy
1	7	1	49	7
2	5	4	25	10
7	8	49	64	56
9	10	81	100	90
8	9	64	81	72
6	4	36	16	24
4	1	16	1	4
3	6	9	36	18
10	3	100	9	30
5	2	25	4	10
$\Sigma x = 55$	$\Sigma y = 55$	$\Sigma x^2 = 385$	$\Sigma y^2 = 385$	$\Sigma xy = 321$

$$r = \frac{321 - 10 \times \frac{55}{10} \times \frac{55}{10}}{\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}\sqrt{385 - 10 \times \left(\frac{55}{10}\right)^2}} = \frac{18.5}{\sqrt{82.5 \times 82.5}} = \frac{18.5}{82.5} = 0.224$$

This shows that the Spearman ρ for any two sets of ranks is the same as the Pearson *r* for the set of ranks. But it is much easier to compute ρ.

Often, the ranks are not given. Instead, the numerical values of observations are given. In such a case, we must attach the ranks to these values to calculate ρ.

Example 3.5: On the basis of given table define correlation and calculate rank.

Marks in Maths	*Marks in Stats.*	*Rank in Maths*	*Rank in Stats.*	*D*	D^2
45	60	4	2	2	4
47	61	3	1	2	4
60	58	1	3	2	4
38	48	5	4	1	1
50	46	2	5	3	9
				ΣD^2 = 22	

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 22}{125 - 5} = -0.1$$

Solution: This shows a negative, though small, correlation between the ranks.

If two or more observations have the same value, their ranks are equal and obtained by calculating the means of the various ranks.

If in this data, marks in maths are 45 for each of the first two students, the rank of each would be $\frac{3+4}{2}$ = 3.5. Similarly, if the marks of each of the last two students in statistics are 48, their ranks would be $\frac{4+5}{2}$ = 4.5

The problem takes the following shape:

Marks in Maths	*Marks in Stats*	*Rank* *x*	*y*	*D*	D^2
45	60	3.5	2	1.5	2.25
45	61	3.5	1	2.5	6.25
60	58	1	3	2	4.00
38	48	5	4.5	0.5	0.25
50	48	2	4.5	2.5	6.25
				ΣD^2	= 19

$$\rho = 1 - \frac{6\Sigma D^2}{n^3 - n} = 1 - \frac{6 \times 19}{120} = +0.05$$

An elaborate formula which can be used in case of equal ranks is:

$$\rho = 1 - \frac{6}{n^3 - n}\left[\Sigma D^2 + \frac{1}{12}\Sigma(m^3 - m)\right].$$

Here, $\frac{1}{12}\Sigma(m^3 - m)$ is to be added to ΣD^2 for each group of equal ranks, *m* being the number of equal ranks each time.

For the given data, we have:

(a) For series *x*, the number of equal ranks $m = 2$.

(b) For series *y*, also, $m = 2$; so that,

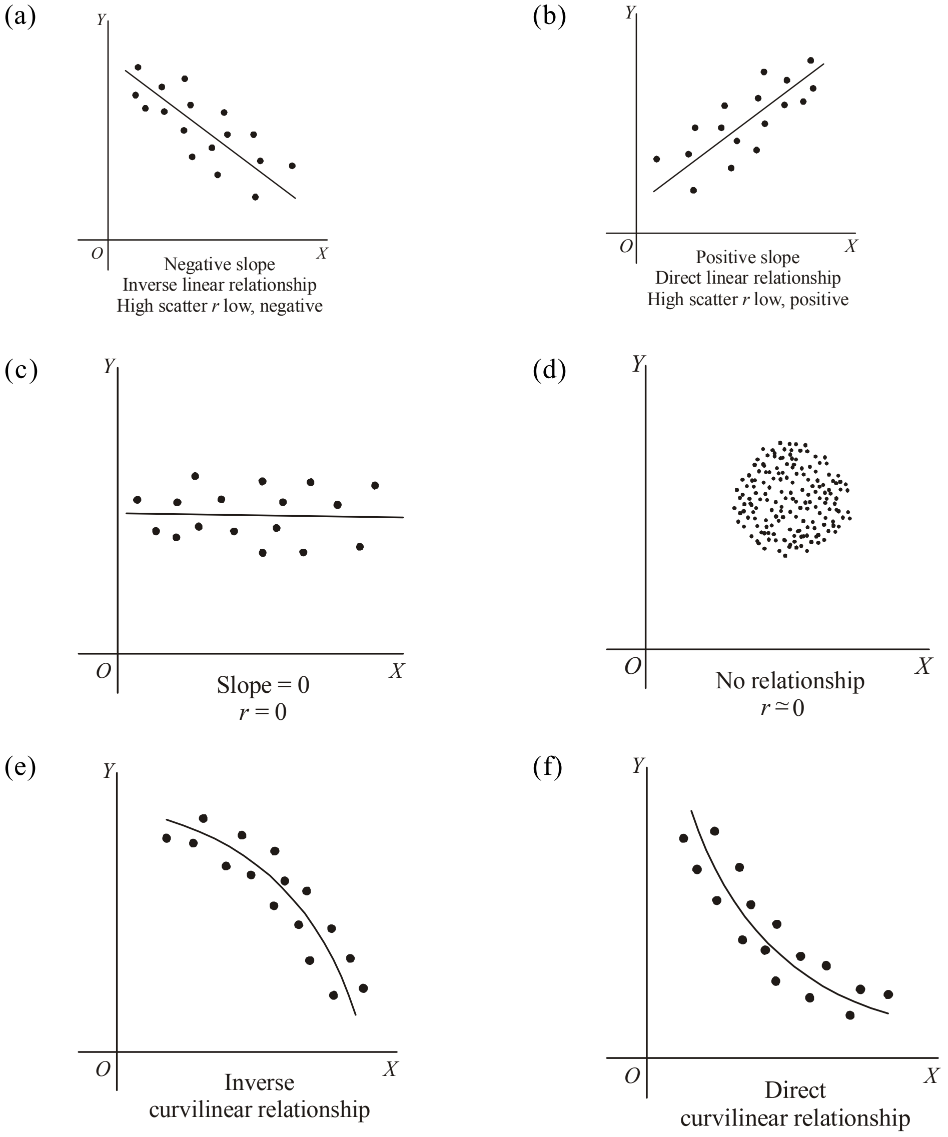
$$\rho = 1 - \frac{6}{5^3 - 5}\left[21 + \frac{1}{12}(2^3 - 2) + \frac{1}{12}(2^3 - 2)\right]$$

$$= 1 - \frac{6}{120}\left[21 + \frac{6}{12} + \frac{6}{12}\right]$$

$$= 1 - \frac{6 \times 22}{120} = -0.1$$

Example 3.6: Show by means of diagrams various cases of scatter expressing correlation between *x, y.*

Solution:



(a) Negative slope
Inverse linear relationship
High scatter *r* low, negative

(b) Positive slope
Direct linear relationship
High scatter *r* low, positive

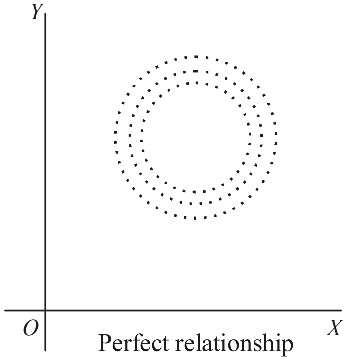
(c) Slope = 0
$r = 0$

(d) No relationship
$r \simeq 0$

(e) Inverse
curvilinear relationship

(f) Direct
curvilinear relationship

(g)



Perfect relationship
But, $r = 0$ because of
non-linear relation

Correlation analysis helps us in determining the degree to which two or more variables are related to each other.

When there are only two variables, we can determine the degree to which one variable is linearly related to the other. Regression analysis helps in determining the pattern of relationship between one or more independent variables and a dependent variable. This is done by an equation estimated with the help of data.

CHECK YOUR PROGRESS

1. What is Karl Pearson's Method of Coefficient?
2. Define Coefficient of Nondetermination.

3.3 MULTIPLE REGRESSION

The term 'regression' was first used in 1877 by Sir Francis Galton who made a study that showed that the height of children born to tall parents will tend to move back or 'regress' towards the mean height of the population. He designated the word regression as the name of the process of predicting one variable from the another variable. He coined the term multiple regression to describe the process by which several variables are used to predict another. Thus, when there is a well-established relationship between variables, it is possible to make use of this relationship in making estimates and to forecast the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s). A banker, for example, could predict deposits on the basis of per capita income in the trading area of the bank. A marketing manager may plan his advertising expenditures on the basis of the expected effect on total sales revenue of a change in the level of advertising expenditure. Similarly, a hospital superintendent could project his need for beds on the basis of total population. Such predictions may be made by using regression analysis. An investigator may employ regression analysis to test his theory having the cause and effect relationship. All this explains that regression analysis is an extremely useful tool specially in problems of business and industry involving predictions.

Assumptions in Regression Analysis

While making use of the regression techniques for making predictions, it is always assumed that:

- There is an actual relationship between the dependent and independent variables.
- The values of the dependent variable are random but the values of the independent variable are fixed quantities without error and are chosen by the experimentor.
- There is clear indication of direction of the relationship. This means that dependent variable is a function of independent variable. (For example, when we say that advertising has an effect on sales, then we are saying that sales has an effect on advertising).
- The conditions (that existed when the relationship between the dependent and independent variable was estimated by the regression) are the same when the regression model is being used. In other words, it simply means that the relationship has not changed since the regression equation was computed.
- The analysis is to be used to predict values within the range (and not for values outside the range) for which it is valid.

3.3.1 Linear Regression Analysis

In case of simple linear regression analysis, a single variable is used to predict another variable on the assumption of linear relationship (i.e., relationship of the type defined by $Y = a + bX$) between the given variables. The variable to be predicted is called the dependent variable and the variable on which the prediction is based is called the independent variable.

Simple linear regression model^3 (or the Regression Line) is stated as,

$$Y_i = a + bX_i + e_i$$

Where, Y_i is the dependent variable

X_i is the independent variable

e_i is unpredictable random element (usually called as residual or error term)

(a) a represent the Y-intercept, i.e., the intercept specifies the value of the dependent variable when the independent variable has a value of zero. (But this term has practical meaning only if a zero value for the independent variable is possible).

(b) b is a constant, indicating the slope of the regression line. Slope of the line indicates the amount of change in the value of the dependent variable for a unit change in the independent variable.

If the two constants (viz., a and b) are known, the accuracy of our prediction of Y (denoted by $\hat{Y}$ and read as Y-hat) depends on the magnitude of the values of e_i. If in the model, all the e_i tend to have very large values then the estimates will not be very good but if these values are relatively small, then the predicted values ($\hat{Y}$) will tend to be close to the true values (Y_i).

Estimating the intercept and slope of the regression model (or estimating the regression equation)

The two constants or the parameters viz., '*a*' and '*b*' in the regression model for the entire population or universe are generally unknown and as such are estimated from sample information. The following are the two methods used for estimation:

(a) Scatter diagram method

(b) Least squares method

1. Scatter Diagram Method

This method makes use of the Scatter diagram also known as Dot diagram. *Scatter diagram*[4] is a diagram representing two series with the known variable, i.e., independent variable plotted on the *X*-axis and the variable to be estimated, i.e., dependent variable to be plotted on the *Y*-axis on a graph paper (see Figure 3.2) to get the following information:

Income *X* *(Hundreds of Rupees)*	*Consumption Expenditure* *Y* *(Hundreds of Rupees)*
41	44
65	60
50	39
57	51
96	80
94	68
110	84
30	34
79	55
65	48

The scatter diagram by itself is not sufficient for predicting values of the dependent variable. Some formal expression of the relationship between the two variables is necessary for predictive purposes. For the purpose, one may simply take a ruler and draw a straight line through the points in the scatter diagram and this way can determine the intercept and the slope of the said line and then the line can be defined as $\hat{Y} = a + bX_i$, with the help of which we can predict *Y* for a given value of *X.* But there are shortcomings in this approach. For example, if five different persons draw such a straight line in the same scatter diagram, it is possible that there may be five different estimates of *a* and *b,* specially when the dots are more dispersed in the diagram. Hence, the estimates cannot be worked out only through this approach. A more systematic and statistical method is required to estimate the constants of the predictive equation. The least squares method is used to draw the best fit line.

NOTES

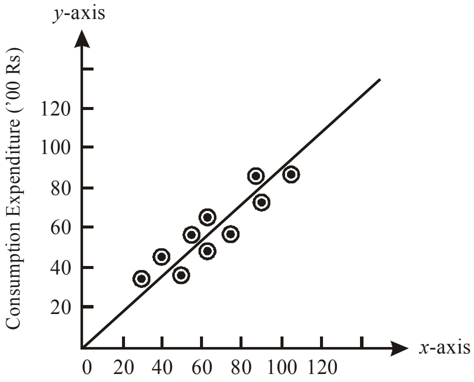


Fig. 3.2 Scatter Diagram

2. Least Square Method

Least square method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.

The meaning of the least squares criterion can be easily understood through reference to Figure 3.3, where the earlier figure in scatter diagram has been reproduced along with a line which represents the least squares line fit to the data.

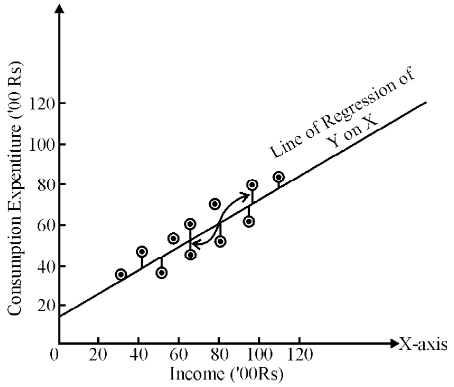


Fig. 3.3 *Scatter Diagram, Regression Line and Short Vertical Lines Representing 'e'*

In Figure 3.3, the vertical deviations of the individual points from the line are shown as the short vertical lines joining the points to the least squares line. These deviations will be denoted by the symbol '*e*'. The value of '*e*' varies from one point to another. In some cases it is positive, while in others it is negative. If the line drawn happens to be least squares line, then the values of $\sum e_i$ is the least possible. It is so, because of this feature the method is known as Least Squares Method.

NOTES

Why we insist on minimizing the sum of squared deviations is a question that needs explanation. If we denote the deviations from the actual value Y to the estimated value $\hat{Y}$ as $(Y - \hat{Y})$ or e_i, it is logical that we want the $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, to be as small as possible. However, mere examining $\Sigma(Y - \hat{Y})$ or $\sum_{i=1}^{n} e_i$, is inappropriate, since any e_i can be positive or negative. Large positive values and large negative values could cancel one another. But large values of e_i regardless of their sign, indicate a poor prediction. Even if we ignore the signs while working out $\sum_{i=1}^{n} |\, e_i \,|$, the difficulties may continue. Hence, the standard procedure is to eliminate the effect of signs by squaring each observation. Squaring each term accomplishes two purposes viz., (*i*) it magnifies (or penalizes) the larger errors, and (*ii*) it cancels the effect of the positive and negative values (since a negative error when squared becomes positive). The choice of minimizing the squared sum of errors rather than the sum of the absolute values implies that there are many small errors rather than a few large errors. Hence, in obtaining the regression line, we follow the approach that the sum of the squared deviations be minimum and on this basis work out the values of its constants viz., '*a*' and '*b*' also known as the intercept and the slope of the line. This is done with the help of the following two normal equations:[5]

$$\Sigma Y = na + b\Sigma X$$

$$\Sigma XY = a\Sigma X + \mathrm{b}\Sigma X^2$$

In the above two equations, '*a*' and '*b*' are unknowns and all other values viz., $\sum X$, $\sum Y$, $\sum \mathrm{X}^2$, $\sum \mathrm{XY}$, are the sum of the products and cross products to be calculated from the sample data, and '*n*' means the number of observations in the sample.

The following examples explain the Least squares method.

Example 3.7: Fit a regression line $\hat{Y} = a + bX_i$ by the method of least squares to the given sample information.

Observations	1	2	3	4	5	6	7	8	9	10
Income (*X*) ('00 ₹)	41	65	50	57	96	94	110	30	79	65
Consumption Expenditure (*Y*) ('00 ₹)	44	60	39	51	80	68	84	34	55	48

Solution: We are to fit a regression line $\hat{Y} = a + bX_i$ to the given data by the method of Least squares. Accordingly, we work out the '*a*' and '*b*' values with the help of the normal equations as stated above and also for the purpose, work out $\sum X$, $\sum Y$, $\sum XY$, $\sum X^2$ values from the given sample information table on Summations for Regression Equation.

Summations for Regression Equation

Observations	*Income* *X* *('00 ₹)*	*Consumption* *Expenditure* *Y* *('00 ₹)*	*XY*	X^2	Y^2
1	41	44	1804	1681	1936
2	65	60	3900	4225	3600
3	50	39	1950	2500	1521
4	57	51	2907	3249	2601
5	96	80	7680	9216	6400
6	94	68	6392	8836	4624
7	110	84	9240	12100	7056
8	30	34	1020	900	1156
9	79	55	4345	6241	3025
10	65	48	3120	4225	2304
$n = 10$	$\sum X = 687$	$\sum Y = 563$	$\sum XY = 42358$	$\sum X^2 = 53173$	$\sum Y^2 = 34223$

Putting the values in the required normal equations we have,

$$563 = 10a + 687b$$

$$42358 = 687a + 53173b$$

Solving these two equations for *a* and *b* we obtain,

$$a = 14.000 \quad \text{and} \quad b = 0.616$$

Hence, the equation for the required regression line is,

$$\hat{Y} = a + bX_i$$

or,

$$\hat{Y} = 14.000 + 0.616X_i$$

This equation is known as the regression equation of *Y* on *X* from which *Y* values can be estimated for given values of *X* variable.[6]

3.3.2 Checking the Accuracy of Equation: Regression Line in Prediction

After finding the regression line as stated above, one can check its accuracy also. The method to be used for the purpose follows from the mathematical property of a line fitted by the method of least squares viz., the individual positive and negative errors must sum to zero. In other words, using the estimating equation, one must find out whether the term $\sum\left(Y - \hat{Y}\right)$ is zero and if this is so, then one can reasonably be sure that he has not committed any mistake in determining the estimating equation.

The problem of prediction

When we talk about prediction or estimation, we usually imply that if the relationship $Y_i = a + bX_i + e_i$ exists, then the regression equation, $\hat{Y} = a + bX_i$ provides a base for making estimates of the value for *Y* which will be associated with particular values of *X*. In Example 3.7, we worked out the regression equation for the income and consumption data as,

$$\hat{Y} = 14.000 + 0.616X_i$$

On the basis of this equation we can make a *point estimate* of *Y* for any given value of *X*. Suppose we wish to estimate the consumption expenditure of individuals with income of ₹ 10,000. We substitute $X = 100$ for the same in our equation and get an estimate of consumption expenditure as follows:

$$\hat{Y} = 14.000 + 0.616(100) = 75.60$$

Thus, the regression relationship indicates that individuals with ₹ 10,000 of income may be expected to spend approximately ₹ 7,560 on consumption. But this is only an expected or an estimated value and it is possible that actual consumption expenditure of same individual with that income may deviate from this amount and if so, then our estimate will be an error, the likelihood of which will be high if the estimate is applied to any one individual. The *interval estimate* method is considered better and it states an interval in which the expected consumption expenditure may fall. Remember that the wider the interval, the greater the level of confidence we can have, but the width of the interval (or what is technically known as the precision of the estimate) is associated with a specified level of confidence and is dependent on the variability (consumption expenditure in our case) found in the sample. This variability is measured by the standard deviation of the error term, '*e*', and is popularly known as the standard error of the estimate.

3.3.3 Standard Error of the Estimate

Standard error of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. Like the standard deviation, the Standard Error (S.E.) of $\hat{Y}$ measures the variability or scatter of the observed values of Y around the regression line. Standard Error of Estimate (S.E. of $\hat{Y}$) is worked out as under:

$$\text{S.E. of } \hat{Y} \text{ (or } S_e) = \sqrt{\frac{\sum (Y - \hat{Y})^2}{n-2}} = \sqrt{\frac{\sum e^2}{n-2}}$$

where, S.E. of $\hat{Y}$ (or S_e) = Standard error of the estimate

Y = Observed value of Y

$\hat{Y}$ = Estimated value of Y

e = The error term = $(Y - \hat{Y})$

n = Number of observations in the sample

Note: In the above formula, $n - 2$ is used instead of n because of the fact that two degrees of freedom are lost in basing the estimate on the variability of the sample observations about the line with two constants viz., '*a*' and '*b*' whose position is determined by those same sample observations.

The square of the S_e, also known as the variance of the error term, is the basic measure of reliability. The larger the variance, the more significant are the magnitudes of the *e*'s and the less reliable is the regression analysis in predicting the data.

Interpreting the standard error of estimate and finding the confidence limits for the estimate in large and small samples

The larger the S.E. of estimate (SE_e), the greater happens to be the dispersion, or scattering, of given observations around the regression line. But if the S.E. of estimate happens to be zero then the estimating equation is a 'perfect' estimator (i.e., cent per cent correct estimator) of the dependent variable.

In case of large samples, i.e., where $n > 30$ in a sample, it is assumed that the observed points are normally distributed around the regression line and we may find,

68 per cent of all points within $\hat{Y} \pm 1$ SE$_e$ limits

95.5 per cent of all points within $\hat{Y} \pm 2$ SE$_e$ limits

99.7 per cent of all points within $\hat{Y} \pm 3$ SE$_e$ limits

This can be stated as,

(*a*) The observed values of Y are normally distributed around each estimated value of $\hat{Y}$ and;

(*b*) The variance of the distributions around each possible value of $\hat{Y}$ is the same.

In case of small samples, i.e., where $n \leq 30$ in a sample the 't' distribution is used for finding the two limits more appropriately.

This is done as follows:

$$\text{Upper limit} = \hat{Y} + \text{'}t\text{'}\ (\text{SE}_e)$$

$$\text{Lower limit} = \hat{Y} - \text{'}t\text{'}\ (\text{SE}_e)$$

Where, $\hat{Y}$ = The estimated value of Y for a given value of X.

SE_e = The standard error of estimate.

't' = Table value of 't' for given degrees of freedom for a specified confidence level.

Some other details concerning simple regression

Sometimes the estimating equation of Y also known as the Regression equation of Y on X, is written as follows:

$$\left(\hat{Y} - \overline{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right)$$

or,

$$\hat{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X_i - \overline{X}\right) + \overline{Y}$$

Where, r = Coefficient of simple correlation between X and Y

σ_Y = Standard deviation of Y

σ_X = Standard deviation of X

$\overline{X}$ = Mean of X

$\overline{Y}$ = Mean of Y

$\hat{Y}$ = Value of Y to be estimated

X_i = Any given value of X for which Y is to be estimated.

This is based on the formula we have used, i.e., $\hat{Y} = a + bX_i$. The coefficient of X_i is defined as,

$$\text{Coefficient of } X_i = b = r\frac{\sigma_Y}{\sigma_X}$$

(Also known as regression coefficient of Y on X or slope of the regression line of Y on X) or b_{YX}.

$$= \frac{\sum XY - n\overline{X}\overline{Y} \times \sqrt{\sum Y^2 - n\overline{Y}^2}}{\sqrt{\sum Y^2 - n\overline{Y^2}}\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum X^2 - n\overline{X}^2}}$$

$$= \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

and $$a = -r\frac{\sigma_Y}{\sigma_X}\overline{X} + \overline{Y}$$

$$= \overline{Y} - b\overline{X} \qquad \left(\text{since } b = r\frac{\sigma_Y}{\sigma_X}\right)$$

Similarly, the estimating equation of X, also known as the regression equation of X on Y, can be stated as:

$$\left(\hat{X} - \overline{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right)$$

or $$\hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right) + \overline{X}$$

and the regression coefficient of X on Y (or b_{XY}) $= r\dfrac{\sigma_X}{\sigma_Y} = \dfrac{\sum XY - n\overline{X}\overline{Y}}{\sum Y^2 - n\overline{Y}^2}$

If we are given the two regression equations as stated above, along with the values of 'a' and 'b' constants to solve the same for finding the value of X and Y, then the values of X and Y so obtained, are the mean value of X (i.e., $\overline{X}$) and the mean value of Y (i.e., $\overline{Y}$).

If we are given the two regression coefficients (viz., b_{XY} and b_{YX}), then we can work out the value of coefficient of correlation by just taking the square root of the product of the regression coefficients as shown below:

$$r = \sqrt{b_{YX}.b_{XY}}$$

$$= \sqrt{r\frac{\sigma_Y}{\sigma_X}.r\frac{\sigma_X}{\sigma_Y}}$$

$$= \sqrt{r.r} = r$$

The (±) sign of *r* will be determined on the basis of the sign of the regression coefficients given. If regression coefficients have minus sign then *r* will be taken with minus (–) sign and if regression coefficients have plus sign then *r* will be taken with plus (+) sign. (Remember that both regression coefficients will necessarily have the same sign whether it is minus or plus for their sign is governed by the sign of coefficient of correlation.)

Example 3.8: Given is the following information:

	$\bar{X}$	$\bar{Y}$
Mean	39.5	47.5
Standard Deviation	10.8	17.8

Simple correlation coefficient between X and Y is = + 0.42

Find the estimating equation of Y and X.

Solution:

Estimating equation of Y can be worked out as,

$$\because \quad \left(\hat{Y}-\bar{Y}\right) = r\frac{\sigma_Y}{\sigma_X}\left(X_i-\bar{X}\right)$$

$$\text{or} \quad \hat{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X_i-\bar{X}\right)+\bar{Y}$$

$$= 0.42\frac{17.8}{10.8}\left(X_i-39.5\right)+47.5$$

$$= 0.69X_i-27.25+47.5$$

$$= 0.69X_i+20.25$$

Similarly, the estimating equation of X can be worked out as under:

$$\because \quad \left(\hat{X}-\bar{X}\right) = r\frac{\sigma_X}{\sigma_Y}\left(Y_i-\bar{Y}\right)$$

$$\text{or} \quad \hat{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y_i-\bar{Y}\right)+\bar{X}$$

$$\text{or} \quad = 0.42\frac{10.8}{17.8}\left(Y_i-47.5\right)+39.5$$

$$= 0.26Y_i - 12.35 + 39.5$$

$$= 0.26Y_i + 27.15$$

Example 3.9: Given is the following data:

Variance of $X = 9$

Regression equations:

$$4X - 5Y + 33 = 0$$

$$20X - 9Y - 107 = 0$$

Find: (*a*) Mean values of X and Y.

(*b*) Coefficient of Correlation between X and Y.

(*c*) Standard deviation of Y.

Solution:

(*a*) For finding the mean values of *X* and *Y*, we solve the two given regression equations for the values of *X* and *Y* as follows:

$$4X - 5Y + 33 = 0 \qquad (1)$$

$$20X - 9Y - 107 = 0 \qquad (2)$$

If we multiply equation (1) by 5, we have the following equations:

$$20X \; - \; 25Y \; = \; -165 \qquad (3)$$

$$20X \; - \; 9Y \; = \; 107 \qquad (2)$$

$$\underline{- \quad + \qquad -}$$

$$-16Y \; = -272$$

Subtracting equation (2) from (3):

or $\quad Y = 17$

Putting this value of *Y* in equation (1) we have,

$$4X = -33 + 5(17)$$

or $\quad X = \dfrac{-33+85}{4} = \dfrac{52}{4} = 13$

Hence, $\quad \overline{X} = 13 \quad$ and $\quad \overline{Y} = 17$

(*b*) For finding the coefficient of correlation, first of all we presume one of the two given regression equations as the estimating equation of *X.* Let equation $4X - 5Y + 33 = 0$ be the estimating equation of *X,* then we have,

$$\hat{X} = \frac{5Y_i}{4} - \frac{33}{4}$$

and

From this we can write $b_{XY} = \dfrac{5}{4}$

The other given equation is then taken as the estimating equation of *Y* and can be written as,

$$\hat{Y} = \frac{20X_i}{9} - \frac{107}{9}$$

and from this we can write $b_{YX} = \dfrac{20}{9}$

If the above equations are correct then *r* must be equal to,

$$r = \sqrt{5/4 \times 20/9} = \sqrt{25/9} = 5/3 = 1.6$$

which is an impossible equation, since *r* can in no case be greater than 1. Hence, we change our supposition about the estimating equations and by reversing it, we re-write the estimating equations as under:

$$\hat{X} = \frac{9Y_i}{20} + \frac{107}{20}$$

and $$\hat{Y} = \frac{4X_i}{5} + \frac{33}{5}$$

NOTES

Hence, $$\begin{aligned} r &= \sqrt{9/20 \times 4/5} \\ &= \sqrt{9/25} \\ &= 3/5 \\ &= 0.6 \end{aligned}$$

Since, regression coefficients have plus signs, we take $r = + 0.6$

(c) Standard deviation of Y can be calculated as follows:

$\because$ Variance of $X = 9$

$\therefore$ Standard deviation of $X = 3$

$$\because \qquad b_{YX} = r\frac{\sigma_Y}{\sigma_X} \quad = \quad \frac{4}{5} = 0.6\frac{\sigma_Y}{3} = 0.2\sigma_Y$$

Hence, $\sigma_Y = 4$

Alternatively, we can work it out as under:

$$\because \qquad b_{XY} = r\frac{\sigma_X}{\sigma_Y} \quad = \quad \frac{9}{20} = 0.6\frac{\sigma_Y}{3} = \frac{1.8}{\sigma_Y}$$

Hence, $\sigma_Y = 4$.

3.3.4 Predicting an Estimate, and its Preciseness

Regression analysis helps predict the most likely value of the dependent variable corresponding to a given, known, value of the independent variable. This is achieved through the predictive function of the *regression equation.* Substituting any known X value in and solving it yields a Y_e value, which serves as the most likely estimate of Y corresponding to that X value.

Accordingly, a Y_e value so estimated may also be read as the expected value of Y against a given value of X. *For example,* for a given value of $X = 10$, $Y_e = 2.823 + 0.252\,(10) = 5.343$ is the expected value based on the linear relationship.

Alternatively, the estimated or expected Y_e value for a given value of X can also be found by using either of the regression equations

$$\left.\begin{aligned} Y_e &= \bar{Y} + b\left(X - \bar{X}\right) \\ \text{or} \quad Y_e - \bar{Y} &= b\left(X - \bar{X}\right), \end{aligned}\right| \qquad (3.1)$$

The use of Eq. (3.1) do not require knowing the value of a. Instead, the value of b and mean $\bar{X}$ and $\bar{Y}$ based on the available data suffice.

Equation 3.1 is obtained as under:

Since

$Y_e = a + bX,$

substituting $(\overline{Y} - b\overline{X})$ for *a* gives

$$Y_e = (\overline{Y} - b\overline{X}) + bX$$

or

$$Y_e = \overline{Y} + b(X - \overline{X})$$

or

$$Y_e - \overline{Y} = b(X - \overline{X}).$$

Given the estimating characteristic of (Eq. 3.1) as used in the manner demonstrated above, it is important to carefully note the following:

(a) An estimated Y_e value against a given value of *X* obtained by using (Eq. 3.1) will be the best approximation as long as the basic conditions affecting the relationship between *X* and *Y,* do not change.

(b) An estimated Y_e value may not be exactly the same as it is actually observed. The difference between an estimated Y_e value and the corresponding observed *Y* value will depend on the extent of scatter of various points around the line of best fit.

(c) The closer the sample paired points (*Y X*) scattered around the line of best fit, the smaller the difference between the estimated Y_e and observed *Y* values, and vice versa. On the whole, the lesser the scatter of the various points around (and, consequently, the lesser the vertical distance by which these deviate from) the line of best fit, the more likely it is that an estimated Y_e value is close to the corresponding observed *Y* value.

(d) The estimated Y_e values will be the same as the observed *Y* values only when all the points on the scatter diagram form a straight line. If this were to be so, the rice yield based on the use of a given quantity of fertilizer could be estimated with 100 per cent accuracy. Since some of the points must lie above and some below the straight line, perfect prediction is practically non-existent in the case of most business and economic variables.

It is evident that the estimated values of one variable based on the known values of the other variable are bound to be different from those that are actually observed. The former are thus always in error compared to the latter. The smaller this error, the more precise and reliable the estimate, and vice versa. That is, the preciseness of an estimate can be known by measuring the magnitude of the said errors. The magnitude of error in an estimate is thus called the *error of estimate.*

3.3.5 Multiple and Partial Correlation

This section describes the measure of simple linear correlation in the following ways:

a. In Terms of Coefficient of Multiple Determination

Just as there was measure of simple linear correlation in the two-variable case, a measure of linear multiple correlation (or just multiple correlation) in the three-variable case is defined in terms of the coefficient of multiple determination. *Denoted as* $R_{1.23}^2$, it uses $S_{1.23}^2$ in its computation, the same way as R^2 uses $S_{Y \cdot X}^2$ in the two-variable case.

The coefficient of multiple determination $R_{1.23}^2$ holds the same significance as that of the coefficient of determination R^2 (or $R_{1.2}^2$) does in the two-variable case. Here, the former represents the proportion of total variations in the dependent variable X_1 which is explained by the variations in the two independent variables X_2 and X_3.

Thus, the coefficient of multiple correlation, *denoted as* $r_{1.23,}$ in the case of three variables is defined as

$$r_{1.23} = \sqrt{R_{1.23}^2} = \sqrt{1 - \frac{S_{1.23}^2}{S_1^2}}. \qquad (3.2)$$

For the illustration under consideration,

$$r_{1.23} = \sqrt{1 - \frac{S_{1.23}^2}{S_1^2}} = \sqrt{1 - \frac{0.32513}{2.2801}} = 0.9260.$$

The subscripts used in the coefficient of determination $R_{1.23}^2$ (or $r_{1.23}$) are the same as for $S_{1.23}$. Use of 1 before the decimal point refers to the dependent variable X_1 and that of 2 and 3 after the decimal point to the two independent variables X_2 and X_3. The subscripts used thus make it evident that $R_{1.23}^2$ (or $r_{1.23}^2$) is a measure of the proportion of total variations in the dependent variable X_1 explained by the combined influence of the variations in the two independent variables X_2 and X_3.

When either X_2 or X_3 is the dependent variable in place of X_1, the remaining two being the independent variables, the corresponding multiple correlation coefficients are

$$r_{2.13} = \sqrt{R_{2.13}^2} = \sqrt{1 - \frac{S_{2.13}^2}{S_2^2}}.$$

and $$r_{3.12} = \sqrt{R_{3.12}^2} = \sqrt{1 - \frac{S_{3.12}^2}{S_3^2}}. \qquad (3.3)$$

respectively.

we have

$$r_{2.13} = \sqrt{1 - \frac{1.8512}{30.69}} = 0.9694$$

and

$$r_{3.12} = \sqrt{1 - \frac{0.6309}{7.29}} = 0.9558.$$

NOTES

b. In Terms of Simple Linear Correlation Coefficients

In terms of the simple linear correlation coefficients r_{12}, r_{13} and r_{23}, the multiple correlation coefficient corresponding to Eq. 3.2 is given by

$$r_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}}. \qquad (3.4)$$

Substituting the values,

$$r_{1.23} = \sqrt{\frac{0.86 + 0.79 - (2)(0.93)(0.89)(0.96)}{1 - 0.91}} = 0.9260.$$

It is the same as obtained by using Eq. (3.2).

The other two equations corresponding to Eq. (3.3) are:

$$r_{2.13} = \sqrt{\frac{r_{12}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{13}^2}}$$

and $$r_{3.12} = \sqrt{\frac{r_{13}^2 + r_{23}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{12}^2}}. \qquad (3.5)$$

Substituting the required values in Eq. (3.5),

$$r_{2.13} = \sqrt{\frac{0.86 + 0.91 - (2)(0.93)(0.89)(0.96)}{1 - 0.79}} = 0.9694$$

and

$$r_{3.12} = \sqrt{\frac{0.79 + 0.91 - (2)(0.93)(0.89)(0.96)}{1 - 0.86}} = 0.9558,$$

which are the same as obtained by using Eqs. (3.3).

It may be noted that multiple correlation coefficient ranges from 0 to 1. The closer it tends to 1, the better is the linear relationship between the variables involved. When $r_{1.23} = 1$, it means the multiple correlation is perfect. A zero value of $r_{1.23}$ means there is no linear relationship between the variables, though a non-linear relationship might exist.

Partial Correlation

NOTES

We know that the two partial regression coefficients ($b_{12.3}$ and $b_{13.2}$) serve as measures of the net influence of the two independent variables X_2 and X_3 on the values of the dependent variable X_1. In the same way, the partial correlation coefficients, *denoted as* $r_{12.3}$ and $\mathrm{r}_{13.2}$, are measures of the *net correlation* between the dependent variable and one of the two independent variables (holding the other independent variable constant, or ignoring its effect on the dependent one).

Thus, the subscript to the right of the decimal point indicates the variable which is held constant or whose relationship with the dependent variable has been ignored, while those to the left of the decimal point indicate the two variables between which net correlation is computed. Accordingly, $r_{12.3}$ represents the *partial correlation* between X_1 and X_2, holding X_3 constant. Likewise, $r_{13.2}$ represents partial correlation between X_1 and X_3, holding X_2 constant.

To obtain a measure of, say, $r_{13.2}$ (or the coefficient of partial determination $R^2_{13.2}$), note the following:

(i) $S^2_{1.23}$ represents the variance in X_1 not explained by the two independent variables X_2 and X_3. That is, it is a measure of variance in X_1 that remains unexplained even after adding the second independent variable X_3.

(ii) $S^2_{1.23}$ is similar to $S^2_{1.2}$ for the two-variable case (X_1 and X_2, X_1 being dependent), where the latter represents the variance in X_1 not explained by the single independent variable X_2.

(iii) The difference between $S^2_{1.2}$ and $S^2_{1.23}$ is the reduction in the unexplained variance in X_1 obtained by adding the second independent variable X_3 in the estimating equation.

(iv) It is this reduction in unexplained variance which is used as a measure of net correlation between X_1 and X_3, when the influence of X_2 both on X_1 and X_3 has been taken into account.

Thus, the coefficient of partial determination between X_1 and X_3 (eliminating the effect of X_2) is computed as

$$R^2{}_{13.2} = 1 - \frac{S^2_{1.23}}{S^2_{1.2}},$$

so that the coefficient of partial correlation is

$$r_{13.2} = \sqrt{1 - \frac{S^2_{1.23}}{S^2_{1.2}}}. \qquad (3.6)$$

It may be recalled that $S^2_{1.2}$ is the square of standard error of estimate $S_{1.2}$, as against S^2_1, as the variance of X_1.

Substituting the values,

$$r_{13.2} = \sqrt{1 - \frac{0.32537}{0.3267}} = \sqrt{1 - 0.99593} = 0.0638.$$

Similarly,

$$R^2_{12.3} = 1 - \frac{S^2_{1.23}}{S^2_{1.3}} \text{ and } r_{12.3} = \sqrt{1 - \frac{S^2_{1.23}}{S^2_{1.3}}}. \tag{3.7}$$

It may be noted that $r_{12.3}$ and $r_{13\ 2}$ take the same signs as the corresponding partial regression coefficients $b_{12.3}$ and $b_{13.2}$.

The results of Eqs. (3.6) and (3.7) can also be obtained in terms of the simple correlation coefficients as

$$r_{13.2} = \frac{r_{13} - r_{12}r_{23}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{23}^2\right)}}$$

and $$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{\left(1 - r_{13}^2\right)\left(1 - r_{23}^2\right)}}. \tag{3.8}$$

Similarly, $$r_{23.1} = \frac{r_{23} - r_{12}r_{23}}{\sqrt{\left(1 - r_{12}^2\right)\left(1 - r_{13}^2\right)}}$$

Substituting the values of simple correlation coefficients in the first of the two Eqs. (3.8), we have

$$r_{13.2} = \frac{0.8911 - (0.9256)(0.9554)}{\sqrt{(1 - 0.8568)(1 - 0.9128)}} = 0.068,$$

which is approximately the same as obtained from Eq. (3.6).

Likewise,

$$r_{12.3} = \frac{0.9256 - (0.8911)(0.9554)}{\sqrt{(1 - 0.7941)(1 - 0.9128)}} = 0.5537.$$

For the three-variable case, an interesting result connecting the multiple and partial correlation coefficients lies in

$$\left(1 - R^2_{1.23}\right) = \left(1 - r^2_{12}\right)\left(1 - r^2_{13.2}\right)$$

or

$$R^2_{1.23} = 1 - \left(1 - r^2_{12}\right)\left(1 - r^2_{13.2}\right).$$

CHECK YOUR PROGRESS

3. What is the least square method of fitting a line?
4. What is standard error of estimate?

3.4 SUMMARY

- Correlation analysis looks at the indirect relationships and establishes the variables which are most closely associated with a given data or mindset. It is the process of finding how accurately the line fits using the observations. Correlation analysis can be referred to as the statistical tool used to describe the degree to which one variable is related to another. The relationship, if any, is usually assumed to be a linear one. In fact, the word correlation refers to the relationship or the interdependence between two variables.
- The theory by means of which quantitative connections between two sets of phenomena are determined is called the 'Theory of Correlation'. On the basis of the theory of correlation, you can study the comparative changes occurring in two related phenomena and their cause-effect relation can also be examined. Thus, correlation is concerned with relationship between two related and quantifiable variables and can be positive or negative.
- On the basis of the theory of correlation, one can study the comparative changes occurring in two related phenomena and their cause-effect relation can be examined. It should, however, be borne in mind that relationship like 'black cat causes bad luck', 'filled up pitchers result in good fortune' and similar other beliefs of the people cannot be explained by the theory of correlation, since they are all imaginary and are incapable of being justified mathematically.
- Correlation can either be positive or it can be negative. Whether correlation is positive or negative would depend upon the direction in which the variables are moving. If both variables are changing in the same direction, then correlation is said to be positive, but, when the variations in the two variables take place in opposite direction, the correlation is termed as negative.
- The coefficient of determination (symbolically indicated as r^2, though some people would prefer to put it as R^2) is a measure of the degree of linear association or correlation between two variables, say X and Y, one of which happens to be independent variable and the other being dependent variable.
- Coefficients of determination is that fraction of the total variation of Y which is explained by the regression line. In other words, coefficient of determination is the ratio of explained variation to total variation in the Y variable related to the X variable.

NOTES

- The coefficient of determination can have a value ranging from zero to one. The value of one can occur only if the unexplained variation is zero, which simply means that all the data points in the Scatter diagram fall exactly on the regression line.
- The coefficient of correlation, symbolically denoted by 'r', is another important measure to describe how well one variable is explained by another. It measures the degree of relationship between the two casually related variables. The value of this coefficient can never be more than +1 or less than –1. Thus, +1 and –1 are the limits of this coefficient.
- There are several methods of finding the coefficient of correlation. Some of the important methods include: a) Coefficient of Correlation by the Method of Least Squares b) Coefficient of Correlation using Simple Regression Coefficients c) Coefficient of Correlation through Product Moment Method or Karl Pearson's Coefficient of Correlation
- Probable Error (P.E.) of r is very useful in interpreting the value of r and is worked out as under for Karl Pearson's coefficient of correlation:

$$\text{P.E.} = 0.6745\frac{1-r^2}{\sqrt{n}}$$

- Two other measures are often talked about along with the coefficients of determinations and that of correlation. These are as follows: (a) Coefficient of Non-determination (b) Coefficient of Alienation.
- If observations on two variables are given in the form of ranks and not as numerical values, it is possible to compute what is known as rank correlation between the two series. The rank correlation, written as r, is a descriptive index of agreement between ranks over individuals. It is the same as the ordinary coefficient of correlation computed on ranks, but its formula is simpler.

$$\rho = 1 - \frac{6\Sigma D_i^2}{n(n^2-1)}$$

- Sir Francis Galton coined the term multiple regression to describe the process by which several variables are used to predict another. Thus, when there is a well-established relationship between variables, it is possible to make use of this relationship in making estimates and to forecast the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s).
- Scatter Diagram Method makes use of the Scatter diagram also known as Dot diagram. Scatter diagram is a diagram representing two series with the known variable, i.e., independent variable plotted on the X-axis and the variable to be estimated, i.e., dependent variable to be plotted on the Y-axis on a graph paper.
- Least square method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the

squared vertical deviations from the fitted line. In other words, the line to be fitted will pass through the points of the scatter diagram in such a way that the sum of the squares of the vertical deviations of these points from the line will be a minimum.

NOTES

3.5 KEY TERMS

- **Correlation analysis:** It is a statistical tool, used to describe the degree to which one variable is related to another.
- **Coefficient of determination:** It is a measure of the degree of linear association or correlation between two variables, one of which must be an independent variable and the other, a dependent variable.
- **Coefficient of correlation:** It measures the degree of relationship between the two casually related variables.
- **Regression analysis:** It is a relationship used for making estimates and forecasts about the value of one variable (the unknown or the dependent variable) on the basis of the other variable/s (the known or the independent variable/s).
- **Scatter diagram:** It is a diagram used to represent two series with the known variables, i.e., independent variable plotted on the X-axis and the variable to be estimated, i.e., dependent variable to be plotted on the Y-axis on a graph paper for the given information.
- **Standard error of the estimate**: It is a measure developed by statisticians for measuring the reliability of the estimating equation.

3.6 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Karl Pearson's method is the most widely used method of measuring the relationship between two variables. The coefficient is based on various assumptions. There is a linear relationship between the two variables which means that straight line would be obtained if the observed data are plotted on a graph.
2. Coefficient of non-determination (denoted by k^2) is the ratio of unexplained variation to total variation in the Y variable related to the X variable. Algebraically, we can write it as follows:

$$k^2 = \frac{\text{Unexplained variation}}{\text{Total variation}} = \frac{\Sigma\left(Y-\hat{Y}\right)^2}{\Sigma\left(Y-\overline{Y}\right)^2}$$

3. Least square method of fitting a line (the line of best fit or the regression line) through the scatter diagram is a method which minimizes the sum of the squared vertical deviations from the fitted line.

4. Standard error of estimate is a measure developed by the statisticians for measuring the reliability of the estimating equation. Like the standard deviation, the Standard Error (S.E.) of measures the variability or scatter of the observed values of Y around the regression line.

3.7 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is the importance of correlation analysis?
2. How will you determine the coefficient of determination?
3. What is the relationship between coefficient of nondetermination and coefficient of alienation?
4. List the basic precautions and limitations of regression and correlation analyses.
5. What is Spearman's rank correlation?
6. Define regression analysis.
7. How will you predict the value of dependent variable?
8. Differentiate between scatter diagram and least square method.
9. Can the accuracy of estimated equation be checked? Explain.
10. How is the standard error of estimate calculated?

Long-Answer Questions

1. Explain the method to calculate the coefficient of correlation using simple regression coefficient.
2. Describe Karl Pearson's method of measuring coefficient of correlation.
3. Calculate correlation coefficient and the two regression lines for the following information:

		Ages of Wives (in years)				
		10–20	*20–30*	*30–40*	*40–50*	*Total*
Ages of	10–20	20	26	—	—	46
Husbands	20–30	8	14	37	—	59
(in	30–40	—	4	18	3	25
years)	40–50	—	—	4	6	10
	Total	28	44	59	9	140

4. Two random variables have the regression with equations,

   $3X + 2Y - 26 = 0$

   $6X + Y - 31 = 0$

   Find the mean value of X as well as of Y and the correlation coefficient between X and Y. If the variance of X is 25, find σ_Y from the data given above.

5. (*i*) Give one example of a pair of variables which would have,
   (a) An increasing relationship
   (b) No relationship
   (c) A decreasing relationship

   (*ii*) Suppose that the general relationship between height in inches (X) and weight in kg (Y) is $Y' = 10 + 2.2\,(X)$. Consider that weights of persons of a given height are normally distributed with a dispersion measurable by $\sigma_e = 10$ kg.
   (a) What would be the expected weight for a person whose height is 65 inches?
   (b) If a person whose height is 65 inches should weigh 161 kg., what value of e does this represent?
   (c) What reasons might account for the value of e for the person in case (*b*)?
   (d) What would be the probability that someone whose height is 70 inches would weigh between 124 and 184 kg?

6. Calculate correlation coefficient from the following results:

$$\begin{aligned} n &= 10;\ \Sigma X = 140;\ \Sigma Y = 150 \\ \Sigma(X-10)^2 &= 180;\ (Y-15)^2 = 215 \\ \Sigma(X-10)\,(Y-15)) &= 60 \end{aligned}$$

7. Examine the following statements and state whether each one of the statements is true or false, assigning reasons to your answer.
   (a) If the value of the coefficient of correlation is 0.9 then this indicates that 90 per cent of the variation in dependent variable has been explained by variation in the independent variable.
   (b) It would not be possible for a regression relationship to be significant if the value of r^2 was less than 0.50.
   (c) If there is found a high significant relationship between the two variables X and Y, then this constitutes definite proof that there is a casual relationship between these two variables.
   (d) Negative value of the 'b' coefficient in a regression relationship indicates a weaker relationship between the variables involved than would a positive value for the 'b' coefficient in a regression relationship.
   (e) If the value for the 'b' coefficient in an estimating equation is less than 0.5, then the relationship will not be a significant one.
   (f) $r^2 + k^2$ is always equal to one. From this it can also be inferred that $r + k$ is equal to one

$$\begin{pmatrix} r = \text{coefficient of correlation; } r^2 = \text{coefficient of determination} \\ k = \text{coefficient of alienation; } k^2 = \text{coefficient of non-determination} \end{pmatrix}$$

3.8 FURTHER READING

Chance, William A. 1969. *Statistical Methods for Decision Making*. Illinois: Richard D Irwin.

Chandan, J.S. 1998. *Statistics for Business and Economics*. New Delhi: Vikas Publishing House.

Chandan, J.S., Jagjit Singh and K.K. Khanna. 1995. *Business Statistics*, Second Edition. New Delhi: Vikas Publishing House.

Elhance, D.N. 2006. *Fundamental of Statistics*. Allahabad: Kitab Mahal.

Freud, J.E., and F.J. William. 1997. *Elementary Business Statistics – The Modern Approach*. Third Edition. New Jersey: Prentice-Hall International.

Goon, A.M., M.K. Gupta, and B. Das Gupta. 1983. *Fundamentals of Statistics*. Vols. I & II, Kolkata: The World Press Pvt. Ltd.

References

1. Remember the short-cut formulae to workout b_{XY} and b_{YX}:

$$b_{XY} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum Y^2 - n\overline{Y}^2}$$

and $$b_{YX} = \frac{\sum XY - n\overline{X}\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

2. In case we take assumed mean to be zero for X variable as for Y variable then our formula will be as under:

$$r = \frac{\frac{\sum XY}{n} - \left(\frac{\sum X}{n}\right)\left(\frac{\sum Y}{n}\right)}{\sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}\sqrt{\frac{\sum Y^2}{n} - \left(\frac{\sum Y}{n}\right)^2}}$$

or $$r = \frac{\frac{\sum XY}{n} - \overline{X}\overline{Y}}{\sqrt{\frac{\sum X^2}{n} - \overline{X}^2}\sqrt{\frac{\sum Y^2}{n} - \overline{Y}^2}}$$

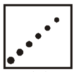
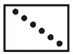
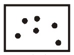
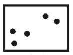
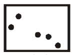
$$r = \frac{\sum XY - n\overline{X}\overline{Y}}{\sqrt{\sum X^2 - n\overline{X}^2}\sqrt{\sum Y^2 - n\overline{Y}^2}}$$

3. Usually the estimate of Y denoted by $\hat{Y}$ is written as,

$$\hat{Y} = a + bX_i$$

on the assumption that the random disturbance to the system averages out or has an expected value of zero (i.e., $e = 0$) for any single observation. This regression model is known as the Regression line of Y on X from which the value of Y can be estimated for the given value of X.

NOTES

4. 
(1)

(2)

(3)

(4)

(5)

Five possible forms which the Scatter diagram may assume has been depicted in the above five diagrams. *First* diagram is indicative of perfect positive relationship, *Second* shows perfect negative relationship, *Third* shows no relationship, *Fourth* shows positive relationship and *Fifth* shows negative relationship between the two variables under consideration.

5. If we proceed centering each variable, i.e., setting its origin at its mean, then the two equations will be as under:

$$\sum Y = na + b\sum X$$

$$\sum XY = a\sum X + b\sum X^2$$

But since $\sum Y$ and $\sum X$ will be zero, the first equation and the first term of the second equation will disappear and we shall simply have the following equations:

$$\sum XY = b\sum X^2$$

$$b = \sum XY/\sum X^2$$

The value of '*a*' can then be worked out as:

$$a = Y\_ - bX\_$$

6. It should be pointed out that the equation used to estimate the *Y* variable values from values of *X* should not be used to estimate the values of *X* variable from given values of *Y* variable. Another regression equation (known as the regression equation of *X* on *Y* of the type $X = a + bY$) that reverses the two value should be used if it is desired to estimate *X* from value of *Y*.

UNIT 4 NORMAL DISTRIBUTION

NOTES

Structure

4.0 INTRODUCTION

In this unit, you will learn about the significant characteristics and applications of normal distribution. Normal distribution is the most common type of distribution. Typically, a normal distribution is a very important statistical data distribution pattern occurring in many natural phenomena, such as height, blood pressure, lengths of objects produced by machines. Certain data, when graphed as a histogram (data on the horizontal axis, amount of data on the vertical axis), creates a bell-shaped curve known as a normal curve or normal distribution.

Thus, the normal distribution is a very important class of statistical distribution. All normal distributions are symmetric and have bell shaped density curves with a single peak. Specifically, in any normal distribution, two quantities have to be specified, the mean *m* where the peak of the density occurs and the standard deviation *s* which indicates the spread or girth of the bell curve.

You will also learn about Z-test and *t*-test for independent and dependent group in this unit. A Z-test is any statistical test for which the distribution of the test state can be approximated by normal distribution under the null hypothesis. Typically, it is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large. Like normal distribution, *t* distribution is also symmetrical but happens to be flatter than normal distribution. The *t*-test can be used to compare a sample mean to an accepted value (of a population mean) or it can be used to compare the means of two sample sets. For applying *t*-test in the context of small samples, the *t* value is calculated first of all and then the calculated value is compared with the table value of *t* at certain level of significance for given degrees of freedom. The *t*-test is used when two conditions are fulfilled, such as when the sample size is less than 30, i.e., when $n \leq 30$ and when the population standard deviation (σ_p) must be unknown.

4.1 UNIT OBJECTIVES

After going through this unit, you will be able to:

- Explain the significance and characteristics of normal distribution
- Discuss how to measure the area under the normal curve
- Assess the significance of Z-test
- Explain the steps of Z-test for independent and dependent group
- Discuss the importance of *t*-test
- Discuss sampling distribution of means
- Explain the two-tailed and one-tailed tests of significance

4.2 MEANING, SIGNIFICANCE AND CHARACTERISTICS OF NORMAL CURVE

Among all the probability distributions, the normal probability distribution is by far the most important and frequently used continuous probability distribution. This is so because this distribution fits well in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter (i.e., between the sample results and the population from which the sample is drawn). The name Karl Gauss, 18th century mathematician–astronomer, is associated with this distribution and in honour of his contribution, this distribution is often known as the Gaussian distribution.

Normal distribution can be theoretically derived as the limiting form of many discrete distributions. For instance, if in the binomial expansion of $(p + q)^n$, the value of '*n*'is infinity and $p = q = \frac{1}{2}$, then a perfectly smooth symmetrical curve would be obtained. Even if the values of *p* and *q* are not equal but if the value of the exponent '*n*' happens to be extremely large, we get a curve of normal probability smooth and symmetrical. Such curves are called normal probability curves (or at times known as normal curves of error) and such curves represent the normal distributions.

The probability function in case of normal probability distribution is given as:

$$f(x) = \frac{1}{\sigma.\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Where, μ = The mean of the distribution.

σ^2 = Variance of the distribution.

The normal distribution is thus, defined by two parameters viz., μ and σ^2. This distribution can be represented graphically as in Figure 4.1.

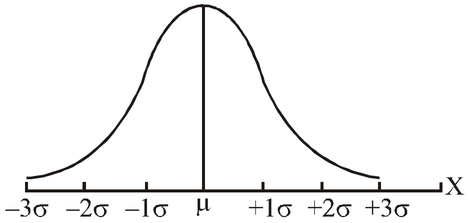


Fig. 4.1 *Curve Representing Normal Distribution*

Characteristics of Normal Distribution

The characteristics of the normal distribution or that of a normal curve are as given below:

1. It is a symmetric distribution.
2. The mean μ defines where the peak of the curve occurs. In other words, the ordinate at the mean is the highest ordinate. The height of the ordinate at a distance of one standard deviation from the mean is 60.653 per cent of the height of the mean ordinate and similarly the height of other ordinates at various standard deviations (σ_s) from mean happens to be a fixed relationship with the height of the mean ordinate (refer Figure 4.2).
3. The curve is asymptotic to the base line which means that it continues to approach but never touches the horizontal axis.
4. The variance (σ^2) defines the spread of the curve.
5. Area enclosed between mean ordinate and an ordinate at a distance of one standard deviation from the mean is always 34.134 per cent of the total area of the curve. It means that the area enclosed between two ordinates at one sigma (SD) distance from the mean on either side would always be 68.268 per cent of the total area. This can be shown as follows:

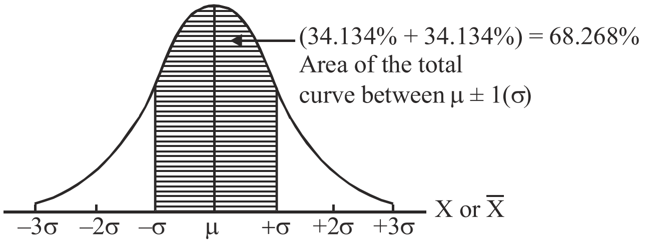


Fig. 4.2 *Area of the Total Curve between* μ ± 1 *(σ)*

Similarly, the other area relationships are as follows:

Between		*Area Covered to Total Area of the Normal Curve*
μ ± 1	S.D.	68.27 per cent
μ ± 2	S.D.	95.45 per cent
μ ± 3	S.D.	99.73 per cent
μ ± 1.96	S.D.	95 per cent
μ ± 2.578	S.D.	99 per cent
μ ± 0.6745	S.D.	50 per cent

6. The normal distribution has only one mode since the curve has a single peak. In other words, it is always a unimodal distribution.
7. The maximum ordinate divides the graph of normal curve into two equal parts.
8. In addition to all the above stated characteristics the curve has the following properties:

   (a) $\mu = \overline{x}$

   (b) $\mu_2 = \sigma^2 =$ Variance

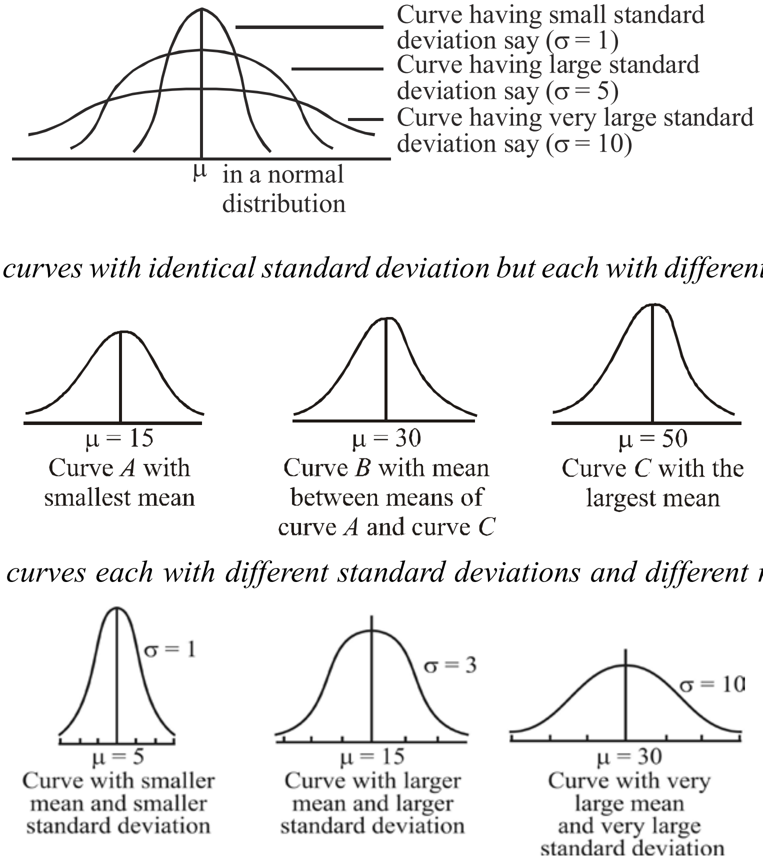
   (c) $\mu_4 = 3\sigma^4$

   (d) Moment Coefficient of Kurtosis = 3

Family of Normal Distributions

We can have several normal probability distributions but each particular normal distribution is being defined by its two parameters viz., the mean (μ) and the standard deviation (σ). There is, thus, not a single normal curve but rather a family of normal curves (refer Figure 4.3). We can exhibit some of these as under:

Normal curves with identical means but different standard deviations:



Curve having small standard deviation say (σ = 1)
Curve having large standard deviation say (σ = 5)
Curve having very large standard deviation say (σ = 10)
μ in a normal distribution

Normal curves with identical standard deviation but each with different means:

μ = 15
Curve *A* with smallest mean
μ = 30
Curve *B* with mean between means of curve *A* and curve *C*
μ = 50
Curve *C* with the largest mean

Normal curves each with different standard deviations and different means:

σ = 1
μ = 5
Curve with smaller mean and smaller standard deviation
σ = 3
μ = 15
Curve with larger mean and larger standard deviation
σ = 10
μ = 30
Curve with very large mean and very large standard deviation

Fig. 4.3 *Family of Curves*

How to Measure the Area under the Normal Curve

NOTES

We have stated above some of the area relationships involving certain intervals of standard deviations (plus and minus) from the means that are true in case of a normal curve. But what should be done in all other cases? We can make use of the statistical tables constructed by mathematicians for the purpose. Using these tables we can find the area (or probability, taking the entire area of the curve as equal to 1) that the normally distributed random variable will lie within certain distances from the mean. These distances are defined in terms of standard deviations. While using the tables showing the area under the normal curve we talk in terms of standard variate (symbolically Z) which really means standard deviations without units of measurement and this 'Z' is worked out as under:

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

Where, Z = The standard variate (or number of standard deviations from X to the mean of the distribution).

X = Value of the random variable under consideration.

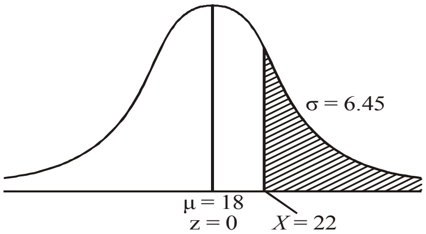
μ = Mean of the distribution of the random variable.

σ = Standard deviation of the distribution.

The table showing the area under the normal curve (often termed as the standard normal probability distribution table) is organized in terms of standard variate (or Z) values. It gives the values for only half the area under the normal curve, beginning with $Z = 0$ at the mean. Since the normal distribution is perfectly symmetrical the values true for one half of the curve are also true for the other half. We now illustrate the use of such a table for working out certain problems.

Example 4.1: A banker claims that the life of a regular savings account opened with his bank averages 18 months with a standard deviation of 6.45 months. Answer the following: (a) What is the probability that there will still be money in 22 months in a savings account opened with the said bank by a depositor? (b) What is the probability that the account will have been closed before two years?

Solution: (a) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown in figure given below:

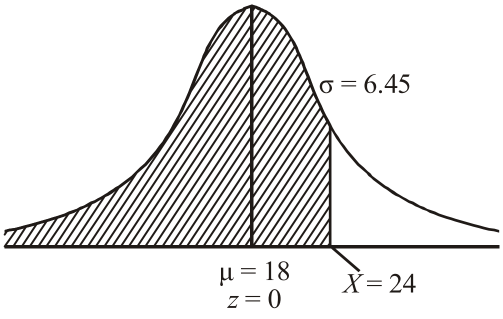


Let us calculate Z as under:

$$Z = \frac{X - \mu}{\sigma} = \frac{22 - 18}{6.45} = 0.62$$

The value from the table showing the area under the normal curve for $Z = 0.62$ is 0.2324. This means that the area of the curve between $\mu = 18$ and $X = 22$ is 0.2324. Hence, the area of the shaded portion of the curve is (0.5) – (0.2324) = 0.2676 since the area of the entire right hand portion of the curve always happens to be 0.5. Thus, the probability that there will still be money in 22 months in a savings account is 0.2676.

(b) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown in figure given below:



For the purpose we calculate,

$$Z = \frac{24 - 18}{6.45} = 0.93$$

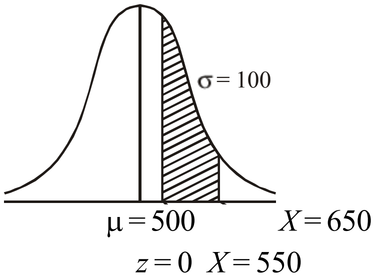
The value from the concerning table, when $Z = 0.93$, is 0.3238 which refers to the area of the curve between $\mu = 18$ and $X = 24$. The area of the entire left hand portion of the curve is 0.5 as usual.

Hence, the area of the shaded portion is (0.5) + (0.3238) = 0.8238 which is the required probability that the account will have been closed before two years, i.e., before 24 months.

Example 4.2: Regarding a certain normal distribution concerning the income of the individuals we are given that mean=500 rupees and standard deviation =100 rupees. Find the probability that an individual selected at random will belong to income group,

(a) ₹ 550 to ₹ 650 (b) ₹ 420 to ₹570

Solution: (a) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:



For finding the area of the curve between $X = 550$ to 650, let us do the following calculations:

$$Z = \frac{550 - 500}{100} = \frac{50}{100} = 0.50$$

Corresponding to which the area between $\mu = 500$ and $X = 550$ in the curve as per table is equal to 0.1915 and,

$$Z = \frac{650 - 500}{100} = \frac{150}{100} = 1.5$$

Corresponding to which, the area between $\mu = 500$ and $X = 650$ in the curve, as per table, is equal to 0.4332.

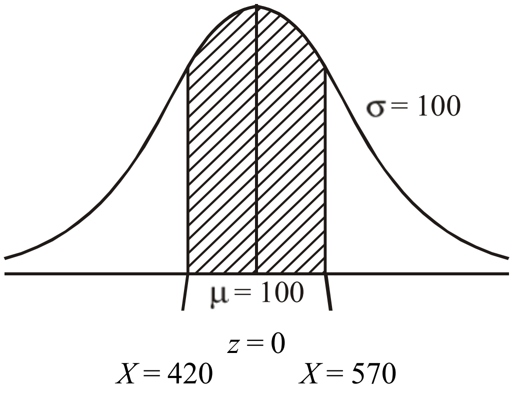
Hence, the area of the curve that lies between $X = 550$ and $X = 650$ is,

$$(0.4332) - (0.1915) = 0.2417$$

This is the required probability that an individual selected at random will belong to income group of ₹ 550 to ₹ 650.

(b) For finding the required probability we are interested in the area of the portion of the normal curve as shaded and shown below:

To find the area of the shaded portion we make the following calculations:



$$Z = \frac{570 - 500}{100} = 0.70$$

Corresponding to which the area between $\mu = 500$ and $X = 570$ in the curve as per table is equal to 0.2580.

and $\quad Z = \frac{420 - 500}{100} = -0.80$

Corresponding to which the area between $\mu = 500$ and $X = 420$ in the curve as per table is equal to 0.2881.

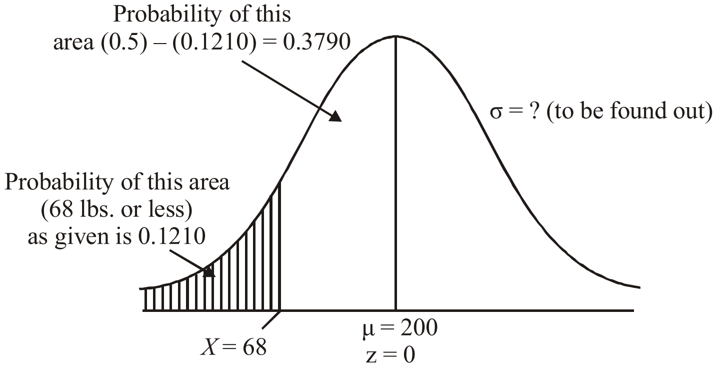
Hence, the required area in the curve between $X = 420$ and $X = 570$ is,

$$(0.2580) + (0.2881) = 0.5461$$

This is the required probability that an individual selected at random will belong to income group of ₹ 420 to ₹ 570.

Example 4.3: A certain company manufactures $1\frac{1''}{2}$ all-purpose rope using imported hemp. The manager of the company knows that the average load-bearing capacity of the rope is 200 lbs. Assuming that normal distribution applies, find the standard deviation of load-bearing capacity for the $1\frac{1''}{2}$ rope if it is given that the rope has a 0.1210 probability of breaking with 68 lbs. or less pull.

Solution: Given information can be depicted in a normal curve as shown below:


Probability of this area (0.5) – (0.1210) = 0.3790
σ = ? (to be found out)
Probability of this area (68 lbs. or less) as given is 0.1210
X = 68
μ = 200
z = 0

If the probability of the area falling within $\mu = 200$ and $X = 68$ is 0.3790 as stated above, the corresponding value of Z as per the table showing the area of the normal curve is – 1.17 (minus sign indicates that we are in the left portion of the curve).

Now to find σ, we can write,

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

or $$-1.17 = \frac{68 - 200}{\sigma}$$

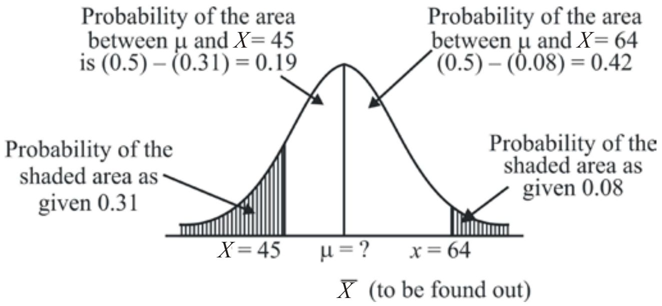
or $$-1.17\sigma = -132$$

or $$\sigma = 112.8 \text{ lbs. approx.}$$

Thus, the required standard deviation is 112.8 lbs. approximately.

Example 4.4: In a normal distribution, 31 per cent items are below 45 and 8 per cent are above 64. Find the $\overline{X}$ and σ of this distribution.

Solution: We can depict the given information in a normal curve as shown below:


Probability of the area between μ and X= 45 is (0.5) – (0.31) = 0.19
Probability of the area between μ and X= 64 (0.5) – (0.08) = 0.42
Probability of the shaded area as given 0.31
Probability of the shaded area as given 0.08
X= 45
μ = ?
x = 64
X̄ (to be found out)

If the probability of the area falling within μ and X = 45 is 0.19 as stated above, the corresponding value of Z from the table showing the area of the normal curve is –0.50. Since, we are in the left portion of the curve, we can express this as under,

$$-0.50 = \frac{45 - \mu}{\sigma} \qquad (1)$$

Similarly, if the probability of the area falling within μ and X = 64 is 0.42, as stated above, the corresponding value of Z from the area table is, +1.41. Since, we are in the right portion of the curve we can express this as under,

$$1.41 = \frac{64 - \mu}{\sigma} \qquad (2)$$

If we solve Equations (1) and (2) above to obtain the value of μ or $\bar{X}$, we have,

$$-0.5\,\sigma = 45 - \mu \qquad (3)$$

$$1.41\,\sigma = 64 - \mu \qquad (4)$$

By subtracting Equation (4) from Equation (3) we have,

$$-1.91\,\sigma = -19$$

$$\therefore \quad \sigma = 10$$

Putting σ = 10 in Equation (3) we have,

$$-5 = 45 - \mu$$

$$\therefore \quad \mu = 50$$

Hence, $\bar{X}$ (or μ)=50 and σ =10 for the concerning normal distribution.

Applications of Normal Distribution

The following are the applications of normal distribution:

1. **Random processes:** Many naturally occurring random processes tend to have a distribution that is approximately normal. Examples can be found in any field, these include: SAT test scores of college bound students and body temperature of a healthy adult.

2. **Approximation of binomial distribution:** When $np>5$ and $n(1-p)>5$, the normal distribution provides a good approximation of the binomial distribution. Distributions that are based on multiple observations, for example the Binomial distribution, approach the normal distribution when n gets large. The value $n>30$ is usually considered large.
3. **Standardization:** It is used where it is usually hypothesized that the theoretical distribution of a certain variable is normal, whereas the measurement of such variable may not give a normal distribution.

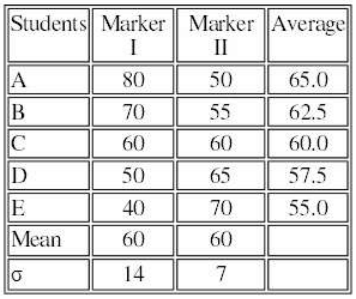
   For example, in the introductory classes of Statistics there are 200 students and it has been assumed that the performance of all the students in the examination should be normally distributed. In addition, for giving reasonable distribution of marks, the mean should be 55 and the standard deviation should be 10. After the examinations being over, the lecturer marked all the papers, and the mean and standard deviation of the raw scores given by the lecturer are 50 and 6, respectively. For converting the raw score to standardize score, the following steps were taken:

   (a) The standard score is obtained by $Z= (X\text{-}50)/6$.

   (b) Then the converted (standardized) $= 10(Z) + 55$.

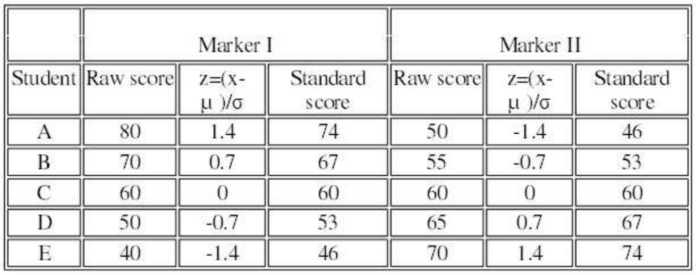
   Hence, a raw score of 56 will be converted into 65.
4. **Composite scores:** When more than one measure is used to measure a variable, the distribution of each measure usually differs from each other. In order to obtain an unbiased measure using several different measurements, each sub-measure is standardized before added together.

   For example, if the marks are awarded according to the average of the marks given by the Marker I and Marker II, then clearly the final grades are greatly affected by the Marker I than by Marker II as Marker I is awarded the marks with higher standard deviation as shown in table below:



Students	Marker I	Marker II	Average
A	80	50	65.0
B	70	55	62.5
C	60	60	60.0
D	50	65	57.5
E	40	70	55.0
Mean	60	60	
σ	14	7	

For computing the composite score, the standardized scores of Marker I and Marker II should be averaged. If the ideal average score (μ) and standard deviation (σ) is taken to be 60 and 10, respectively, then the Z scores is converted into the standard score for each marker. The following table shows the resulted average standardized score 60 for every student.



	Marker I			Marker II		
Student	Raw score	z=(x-μ)/σ	Standard score	Raw score	z=(x-μ)/σ	Standard score
A	80	1.4	74	50	-1.4	46
B	70	0.7	67	55	-0.7	53
C	60	0	60	60	0	60
D	50	-0.7	53	65	0.7	67
E	40	-1.4	46	70	1.4	74

5. **Probability distribution:** The probability distribution of $\bar{X}$ for large *n* is the normal distribution. The Central Limit Theorem states that if the observations are independent for one population which has a mean (μ) and standard deviation (σ) then for large *n* (*n*>30) $\bar{X}$ has a normal distribution with the same mean and a standard deviation of σ / √*n*.

4.2.1 Normal Probability Curve and its Uses

The Normal Probability Curve (NPC), simply known as normal curve, is a symmetrical bell-shaped curve. This curve is based upon the law of probability and discovered by French mathematician Abraham Demoivre (1667–1754) in the 18th century. In this curve, the mean, median and mode lie at the middle point of the distribution. The total area of the curve represents the total number of cases and the middle point represents the mean, median and mode. The base line is divided into six sigma units (σ units). The scores more than the mean come on the + σ side and the scores less than the mean come on the – σ side. The mean point (middle point) is marked as zero (0). All the scores are expected to lie between – 3σ to + 3σ.

Characteristics of Normal Probability Curve

The NPC or Normal Probability Curve has several features which are essential to understand for its use. The major characteristics are limited. They are as follows:

- It is a bell shaped curve.
- The measures of central tendency are equal, i.e, mean, mode and median concentrate on one point.
- The height of the curve is 0.3989.
- It is an asymptotic curve. The ends of the curve approach but never touch the *X*-axis at the extremes because of the possibility of locating in the population, in cases where scores are still higher than our highest score or lower than our lowest score. Therefore theoretically, it extends from minus infinity to plus infinity as illustrated in Figure 4.4. Here, *M* is the mean or expectation (location of the peak) and σ is the standard deviation.

NOTES

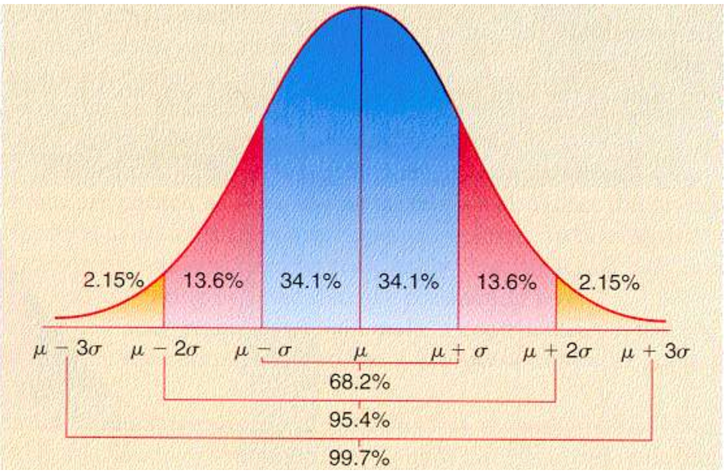


Fig. 4.4 *Normal Curve Showing Areas at Different Distances from the Mean*

- It has 50 per cent frequency above and 50 per cent below the mean. The mean is zero and it is always reference point.
- Standard deviation of a normal curve is always 1.
- The points of inflection of the curve occur at points –1 unit above and below mean.
- The distribution of frequency per cent has the definite limits.
- There is a definite relation between quartile deviation and standard deviation in a normal distribution curve.
- It is a mathematical curve and is an open-ended curve.

Some limits are as follows:

- The middle 68 per cent frequency is between –1 and +1.
- The middle 95 per cent frequency is between –1.96 and + 1.96.
- The middle 99 per cent frequency is between –2.58 and + 2.58.

The total area under the normal curve is arbitrarily taken as 10,000. Every score should be converted into standard score (Z score) by using the following formula:

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

The area in proportion should be converted into a percentage at the time of reading the table. From the table, we can see the areas from mean to σ and also we can read the value of σ scores from the mean for the corresponding fractional area.

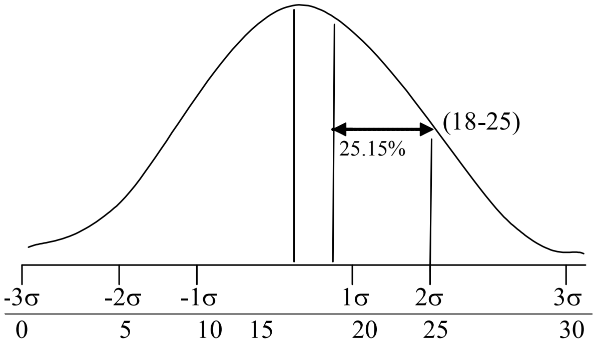
4.2.2 Uses of Normal Probability Curve: Computing Percentiles and Percentile Ranks

The uses of normal probability curve are discussed in this section.

NOTES

NPC is used to Determine the Percentage of Cases within Given Limits

Example 4.5: Given a distribution of scores with a mean of 15 and a standard deviation of 5, what percentage of cases lie between 18 and 25 refer to figure given below to calculate the answer.



Solution: Both the raw scores (18 and 25) are to be converted into *Z* scores.

$$\text{Z score of 18} = \frac{X-M}{\sigma} = \frac{18-15}{5}$$

$$= \frac{3}{5}$$

$$= 0.6\sigma$$

$$\text{Z score of 25} = \frac{X-M}{\sigma} = \frac{25-15}{5}$$

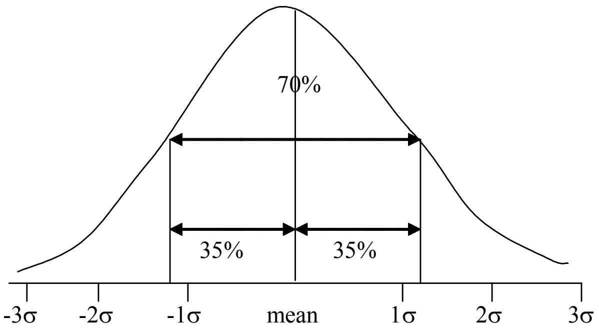
$$= \frac{10}{5}$$

According to the table of area of a normal probability curve, the total percentage of cases lie between the mean and 0.6σ is 22.57. The percentage of cases lie between the mean and 2σ is 47.72. So, the total percentage of cases that fall between the scores 18 and 25 is $47.72 - 22.57 = 25.15$.

NPC is used to Determine the Limit which Includes a Given Percentage of Cases

Example 4.6: Given a distribution of scores with a mean of 12 and an σ of 6, what limits will include the middle 70 per cent of the cases? Refer to figure given to calculate the answer.

NOTES



Mean = 12, σ = 6

Solution: The middle 70 per cent of the cases in a normal distribution signifies that 35 per cent cases above the mean and also 35 per cent cases below the mean. According to the table of area under NPC, 35 per cent of cases fall between the mean and 1.04 σ. So the middle 70 per cent of the cases will lie between –1.04σ to + 1.04σ.

The value of 1 σ = 6

So 1.04 σ = 6 × 1.04 = 6.24

The value of mean = 12

So the lowest limit for the middle 70 per cent cases of the distribution is:

12 – 6.24 = 5.76.

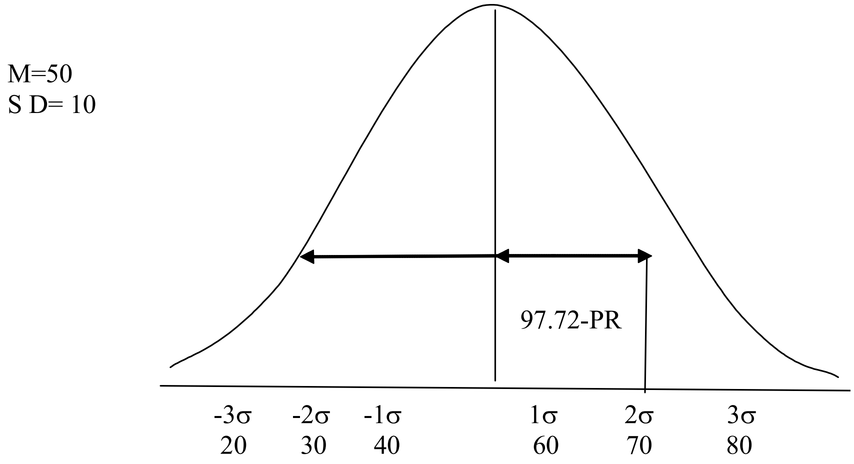
The highest limit for the middle 70 per cent cases of the distribution is:

12 + 6.24 = 18.24.

Thus, the middle 70 per cent cases lie in between 5.76 and 18.24.

NPC is used to Determine the Percentile Rank of a Student in his Class

Example 4.7: The score of a student in a class test is 70. The mean for the whole class is 50 and the σ is 10. Find the percentile rank of the student in the class. Refer the figure given below to find the answer.



Solution: The Z score for the score 70 is $\frac{70-50}{10} = 2\sigma$

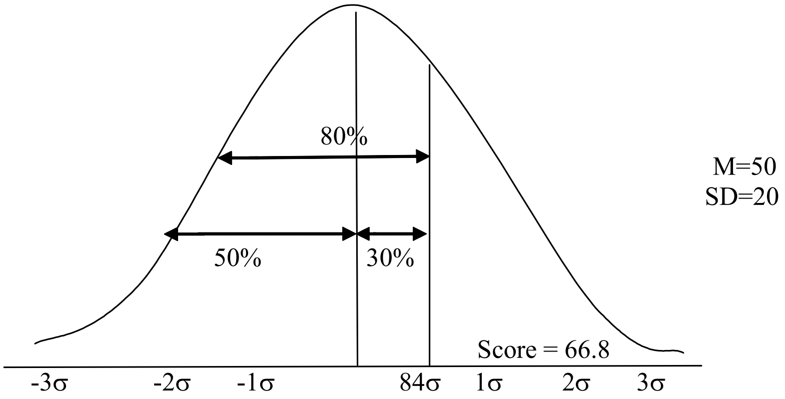
As per the table of area under the NPC, the area of the curve that lies between mean and 2σ is 47.72 per cent. The total percentage of cases below 70 is:

50 + 47.72 = 97.72 per cent or 98 per cent.

Thus, the percentile rank of the student is 98.

NPC is used to Find out the Percentile Value of a Student whose Percentile Rank is Known

Example 4.8: The percentile rank of a student in a class test is 80. The mean of the class in the test is 50 and the σ is 20. Calculate the student's score in the class test. Figure given below illustrates the case.



Solution: The student has scored 30 per cent scores above the mean. According to the table of area under NPC, 30 per cent cases from the mean is 0.84σ.

1σ = 20.

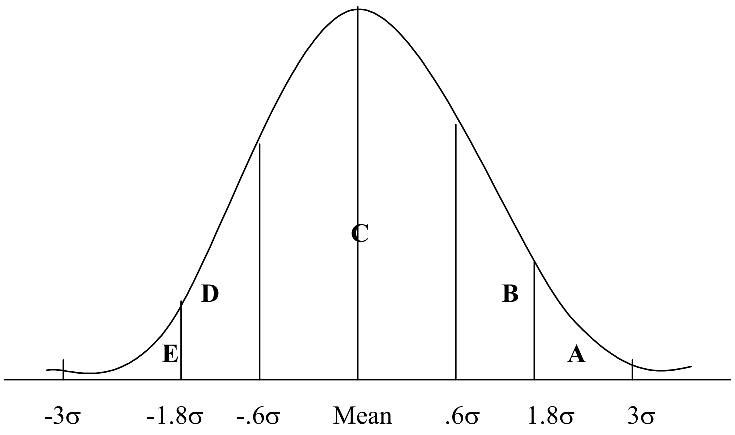
0.84σ = 20 × .84 = 16.8

Thus, the percentile value of the student is 50 + 16.8 = 66.8.

NPC is used to Divide a Group into Sub-Groups According to their Capacity

Example 4.9: Suppose there is a group of 100 students in a Commerce class. We want to divide them into five small groups A, B, C, D and E according to their ability, the range of ability being equal in each sub-group. Find out how many students should be placed in each category.

NOTES



Solution: The total area under NPC is –3σ to + 3σ, that is 6σ. This 6σ should be divided into five parts, so 6σ ÷ 5 = 1.2σ.

According to the table of area under NPC:

3.5 per cent of the cases lie between 1.8 σ to 3σ (Group A, the high scorers). 23.8 per cent of the cases lie between .6σ to 1.8σ (23.8 per cent of the cases for B and also 23.8 per cent of the cases for D), the middle 45 per cent of the cases lie – 0.6σ to + 0.6σ (Group C), and the lowest 3.5 per cent of the cases lie between –3σ to – 1.8σ (Group E)

In category 'A' the number of students = 3.5 per cent = 3 or 4 students.

In category 'B' the number of students = 23.8 per cent = 24 students.

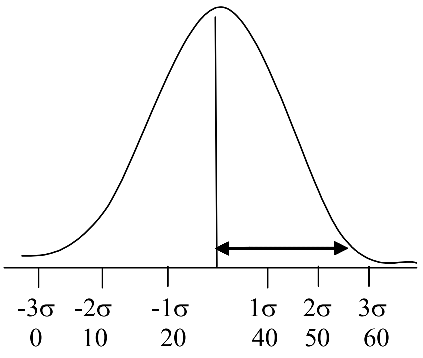
In category 'C' the number of students = 45 per cent = 45 students.

In category 'D' the number of students = 23.8 per cent = 24 students.

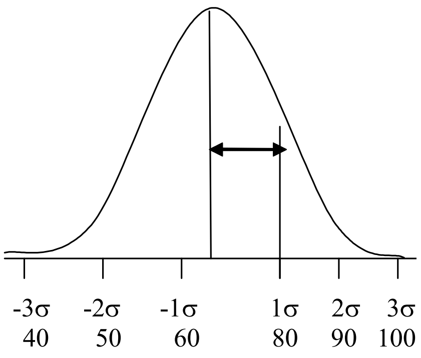
In category 'E' the number of students = 3.5 per cent = 3 or 4 students.

NPC is used to Compare the Scores of Students in Two Different Tests

Example 4.10: Suppose, a student scored 60 marks in English test and 80 marks in statistics test. The mean and SD for the English test is 30 and 10 respectively, whereas for the statistics test the mean is 70 and SD is 10. Find out, in which subject the student performed better. Refer to Figure given below.



M = 30
SD = 10
English Test



Mean = 70
SD = 10
Statistics Test

NOTES

Solution: In case of the English test:

Raw score = 60

Mean = 30

SD =10

So Z score for the English test = $\frac{X-M}{\sigma} = \frac{60-30}{10} = \frac{30}{10} = 3\sigma$

In case of statistics test raw score = 80

Mean = 70

SD = 10

So Z Score for the statistics test = $\frac{X-M}{\sigma} = \frac{80-70}{10} = \frac{10}{10} = 1\sigma$

So, the student has done better in the English than the statistics on.

NPC is Used to Determine the Relative Difficulty Level of Test Items

Example 4.11: In a standardized test of psychology, question numbers A, B, C and D were solved by the students, 45 per cent, 38 per cent, 30 per cent and 15 per cent respectively. Assuming the normality, find out the relative difficulty level of the questions. Also explain the difficulty levels of questions. Table given below displays the information in tabular form.

Solution:

Determining the Difficulty Level of Test Items

Question Number	Percentage of Successful Students	Percentage of Unsuccessful Students	Percentage distance of Mean of Unsuccessful Students	Difficulty Level
A	45	55	55-50=5	0.13σ
B	38	62	62-50=12	0.31σ
C	30	70	70-50=20	0.52σ
D	15	85	85-50=35	1.04σ

As we know that in an NPC, 50 – 50 cases lie both the sides of mean. The mean of NPC is that point which is shown as 0. In an NPC, the explanation of difficulty level is done on the basis of **σ** — distance. Therefore, if a question is at the positive side of the NPC and **σ** has more distance from the mean, the question of a test will be much difficult. The relative difficulty value of the test items has been shown:

The question

A to B is 0.18σ is more difficult ($0.31\sigma - 0.13\sigma = 0.18\sigma$)

A to C is 0.39σ is more difficult ($0.52\sigma - 0.13\sigma = 0.39\sigma$)

A to D is 0.91σ is more difficult ($1.04\sigma - 0.13\sigma = 0.91\sigma$)

B to C is 0.21σ is more difficult (0.52σ –0.31σ = 0.21σ)

B to D is 0.73σ is more difficult (1.04σ –0.31σ = 0.73σ)

C to D is 0.52σ is more difficult (1.04σ –0.152σ = 0.52σ)

NOTES

CHECK YOUR PROGRESS

1. List any three characteristics of normal distribution.
2. What is a Normal Probability Curve?

4.3 NEED, IMPORTANCE AND SIGNIFICANCE OF THE DIFFERENCE BETWEEN MEANS AND OTHER STATISTICS

A statistical hypothesis is a statement about a population parameter. The statement is tentative in the sense that it carries some belief or involves an assumption that may or may not be found valid on verification. The act of verification consists of testing the validity of belief(s) or examining the veracity of the assumption(s), which is indeed what we do in testing hypotheses. When this is done on the basis of sample evidence, testing is called statistical testing, and the hypothesis tested is known as statistical hypothesis.

4.3.1 Null Hypothesis

A statistical hypothesis stated with a view to testing its validity is in fact a null hypothesis and is known as such. It is always the null hypothesis, *denoted as* H_0, that is tested on the basis of sample information.

As the sample information may or may not be found consistent with the stated hypothesis, it is important to note the following:

(a) If the information is found inconsistent with H_0, the null hypothesis is rejected and we conclude that it is false. On the contrary, if the sample information is found consistent with H_0, it is accepted even though we do not conclude that it is true.

(b) The reason for accepting H_0 and yet not concluding that it is true, is that the sample information that justifies acceptance of H_0 is not sufficient to conclude that it is indeed true. When sample information supports H_0, it can at best be considered adequate to conclude that H_0 is not false.

The interpretation in b) above has a direct bearing on how the null hypothesis is to be stated. This requires that the hypothesis is stated with a view to rejecting it. For, in the event of the sample information being inconsistent with the stated hypothesis, it is safer to conclude that the hypothesis is false and be rejected. This precisely is the reason why the hypothesis to be tested is called the null hypothesis.

And, it is in this sense that testing essentially means testing a null hypothesis, which is stated specifically with a view to be rejected.

Alternative Hypothesis

Rejection of H_0 implies that it is rejected in favour of some other hypothesis being accepted. A hypothesis that comes to be accepted at the cost of H_0 is called the alternative hypothesis. *Denoted as* H_1 it may be stated in different ways depending on the nature of the problem statement.

If the hypothesis to be tested relates to any parameter Θ whose value is predetermined or otherwise specified as, say, Θ = 60 units, the null hypothesis is stated as

$H_0 : \Theta = 60$ units.

The form in which H_1 is to be stated depends on our what the present value of Θ is expected to be.

The alternative hypothesis H_1 may thus be stated in keeping with either of the following two situations:

1. In a problem situation where our interest is limited to knowing whether the value of Θ is the same as before or has changed, the alternative hypothesis is stated as

   $H_1: \Theta \neq 60$ units.

   The value of Θ = 60 units in H_0 known as the hypothesised value. It is *denoted as* Θ_0, so that H_0 may be stated as

   $H_0 : \Theta \neq \Theta_0$

   and the alternative hypothesis as

   $H_1 : \Theta \neq \Theta_0$,

   where $\Theta_0 = 60$ units.

2. In the same or a different problem situation, our interest may be to know if the value of Θ has increased or decreased compared to the hypothesised value Θ_0. Where Θ is expected to have increased, the null hypothesis

   $H_0 : \Theta = \Theta_0$

   is tested against the alternative hypothesis

   $H_1 : \Theta > \Theta_0$.

   On the contrary, where Θ is expected to have decreased the null hypothesis H_0 is tested against the alternative hypothesis

   $H_1 : \Theta < \Theta_0$.

Thus, in effect, there are three different ways of stating H_1. In each case, statistical testing means testing H_0 against H_1. It essentially involves a two-choice testing in as much as it results either in the rejection of H_0 in favour of H_1, or in the acceptance of H_0 against H_1.

The various points that emerge from the above discussion may thus be summarised and restated as follows:

- A statistical hypothesis, often called a null hypothesis H_0, is tested against an alternative hypothesis H_1. The latter can be stated in different ways depending on the problem situation as to what the parameter is expected to be at a given point of time.
- A statistical hypothesis is stated with reference to a population parameter, never in relation to the corresponding sample statistic. The appropriate sample statistic serves merely as a means, by providing a point estimate, to decide whether a statistical hypothesis is to be rejected or accepted.
- A statistical hypothesis is always stated in the present tense in a manner that some general state of affairs currently exists. Use of future tense in stating a hypothesis is not admissible, as in that case the null hypothesis will involve a future state of affairs about something that may not exist.

4.3.2 Z-test

The Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by means of a normal distribution. Typically, it is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large. The test statistic is assumed to have a normal distribution and parameters, such as standard deviation should be known in order for an accurate Z-test to be performed. Because of the Central Limit Theorem (CLT), many test statistics are approximately normally distributed for large samples.

For each significance level, the Z-test has a single critical value. Consequently, several statistical tests can be easily performed as approximation of Z-tests if the sample size is large or the population variance is known. But if the population variance is unknown and has to be estimated from the sample itself and the sample size is not large ($n < 30$) then the Student *t*-test may be more appropriate.

Formula for the *Z*-Score for a Single Value

Z-score can be calculated using the following formula:

$$Z = \frac{x - \mu}{\sigma}$$

Where,

x = Sample score.

μ = Population mean.

σ = Population standard deviation.

The Z-score is the number of standard deviations which are away from the mean and can be used to compare different scores when the mean and standard deviation of the population are known. These comparisons are based on the assumptions that the distribution of the population is normal and that the two scores are drawn from tests that measure the same construct(s).

Example 4.12: Ram has taken tests a few weeks ago and got scores of 630 in Verbal and 700 in Quantitative. The means and standard deviations of a set of test takers are as follows:

	Verbal	**Quantitative**
Mean	469	591
Standard Deviation	119	148

Calculate the Z-score of Ram for both the subject.

Solution: The Z-score of Ram for both the subject is calculated as follows:

Here,

Mean of verbal test = 469

Standard deviation of verbal test = 119

Sample score of verbal test = 630

Z-score for verbal test = 630 – 469 / 119

= 161 / 119

= 1.35

Again,

Mean of quantitative test = 591

Standard deviation of quantitative test = 148

Sample score of quantitative test = 700

Z-score for quantitative test = 700 – 591 / 148

= 109 / 148

= 0.73

Example 4.13: Find the Z-score corresponding to a raw score of 132 from a normal distribution with mean 100 and standard deviation 15.

Solution: The Z-score is calculated as follows:

Here,

Raw score = 132

Mean = 100

Standard deviation =15

$$\text{Z-score} = \frac{132-100}{15} = 2.133$$

Example 4.14: A Z-score of 1.7 was found from an observation coming from a normal distribution with mean 14 and standard deviation 3. Find the raw score.

Solution: The raw score is obtained as follows:

Here,

Z-score = 1.7

Mean = 14

Standard deviation = 3

Let the raw score be x.

We know:

$$\text{Z-Score}=\frac{x-\text{Mean}}{\text{Standard Deviation}}$$

$\Rightarrow$ Z-Score × Standard Deviation $= x -$ Mean

$\Rightarrow$ $1.7 \times 3 = x - 14$

$\Rightarrow$ $5.1 = x - 14$

$\Rightarrow$ $5.1 + 14 = x$

$\Rightarrow$ $x = 19.1$

The raw score is 19.1.

Formula for Sample Mean using One Sample Z-Test

One sample Z-test can be calculated from the following formula:

$$z = \frac{(\bar{x}-\mu)}{\frac{\sigma}{\sqrt{n}}}$$

Where,

$\bar{x}$ = Sample mean.

μ = Specific value to be tested.

σ = Population standard deviation.

n = Size of the sample.

Example 4.15: A test is conducted for H_0 where $\mu = 20$, with $\sigma = 4$. A sample of size 36 has $\bar{x} = 21.4$. Find the value of one sample Z-test.

Solution: The value of one sample Z-test is calculated as follows:

Here,

$\bar{x} = 21.4$

$\mu = 20$

$\sigma = 4$

$n = 36$

$Z = 21.4 - 20 / (4/\sqrt{}36)$

$= 1.4 / (4 / 6)$

$= 1.4 / (2/3)$

$= 3 \times 1.4 / 2$

$= 4.2 / 2$

$= 2.1$

The value of one sample Z-test is 2.1.

Z-Test for Independent and Dependent Group

A practical example is given below using the hypothesis testing to illustrate the Z-test for differences in proportions between two independent groups. The research question for our example will be: Do a greater proportion of high school students in the population smoke cigarettes in urban rather than rural cities? This is a directional question. A step-by-step outline will be followed to test this research question.

Step 1: The directional research question in a statistical hypothesis format should be stated.

H_0: $P_1 \leq P_2$ (or $P_1 - P_2 \leq 0$)

H_A: $P_1 > P_2$ (or $P_1 - P_2 > 0$)

Here, 'H' stands for hypothesis with subscripts '0' for the null statistical hypothesis and 'A' for the alternative statistical hypothesis. The alternative statistical hypothesis is stated to reflect the research question. In this example, the alternative statistical hypothesis indicates the directional nature of the research question. Also, P_1 is the population proportion of high school students who smokes in the urban city whereas P_2 is the population proportion of high school students who smoke in the rural city.

Step 2: The criteria for rejecting the null hypothesis and accepting the alternative hypothesis should be determined.

Given $\alpha = 0.1$ level of significance, we select the corresponding Z value which is the closest to1 per cent of the area under the normal curve. If our computer Z-test statistics is greater than Z-value, we would reject the null hypothesis. This establishes our region of rejection, R, or the probability area under the normal curve where differences in sample proportions are unlikely to occur by random chance.

Step 3: The sample data of the collected Z-test statistics should be computed.

A random sample of 20 percent of all high school students from both urban and rural city was selected. In urban city, 20,000 high school students were sampled with 25 per cent smoking cigarettes (n_1 = 5,000). In the rural city, 1,000 high school students were sampled with 15 per cent smoking cigarette (n_2=150). The proportions were:

P_1 = 25 (25 per cent of the boys in the sample of high school students smoking cigarettes).

P_2 = 15 (15 per cent of the girls in the sample of high school students smoking cigarettes).

The standard deviation of the sampling distribution of the difference in independent sample proportions is called the standard error of the difference between independent sample proportions. This value is needed to compute the Z-test statistics. The formula is as follows:

$$S_{P_1-P_2} = \sqrt{\frac{pq}{N}}$$

Where,

$p = (n_1 + n_2/N) = (5000 + 150/21{,}000) = 245.$

$q = 1 - p = 1 - 245 = 755.$

n_1 = Number in the first sample = 5,000.

n_2 = Number in the first sample = 150.

N = Total sample size taken = (20,000 + 1000) = 21,000.

$$S_{P_1-P_2} = \sqrt{\frac{0.245(0.755)}{21000}} = 0.003$$

The Z-test can now be computed as follows:

$$Z = \frac{P_1 - P_2}{S_{P_1-P_2}} = \frac{0.25 - 0.15}{0.003} = \frac{0.10}{0.003} = 33.33$$

Step 4: The confidence interval around the Z-test statistic should be computed.

A confidence interval is computed by using the per cent difference between the two independent groups ($P_1 - P_2$=0.10), the Z value corresponding to a given alpha α level for a two-tailed region of region (Z = 2.58) and then the standard deviation of the sampling distribution or standard error of the test statistics ($S_{P_1\text{-}P_2}$=0.003)

CI_{99} = 0.10 +/ – (2.58) (0.003)

CI_{99} = 0.10 +/ – (0.008)

CI_{99} = (0.092, 0.108)

Step 5: The Z-test static results should be interpreted

Our interception is based upon a test of the null hypothesis and a 99 per cent confidence interval around the computed Z-test statistic. Since the computed Z = 33.33 is greater than the Z value, Z = 2.33 at the 0.1 level of significance, we reject the null statistical hypothesis in favor of alternative statistical hypothesis. The probability that the observed difference in the sample proportions of 10 per cent would have occurred by chance is less than 0.01. We can therefore conclude that the urban city had a greater percentage of high school students smoking cigarettes than the rural city. The Type I error was set at 0.01, so that we can be fairly confident in our interception.

The confidence interval was calculated as 0.92 to 0.108, indicating that we can be 99 per cent confident that this interval contains the difference between the population proportions from which the samples were taken. Moreover, the narrowness of the confidence interval gives the idea that how much the difference in the independent sample proportions might vary from random sample to random sample. Consequently, we can feel fairly confident that a 9 per cent (0.092) to 11 per cent (0.108) difference would exist between urban and rural city high school students smoking cigarettes upon repeated sampling of the population.

Z-Test for Dependent Group

The null hypothesis that there is no difference between two population proportions can also be tested for dependent samples using the Z-test statistics. The research design would involve obtaining percentage from the same sample or group twice. The research design would therefore have paired observations. Some examples of when this occurs would be:

1. Test differences in proportions of agreement in a group before and after a discussion of the death penalty.
2. Test differences in per cent passing for students who take two similar tests.
3. Test differences in proportion of the employee who support a retirement plan and the proportion that support a company daycare.

The research design involves studying the impact of diversity training on the proportion of the company employees who would favor hiring foreign workers. Before and after diversity training, employees were asked whether or not they were in favor of company hiring foreign workers. A step-by-step approach to hypothesis testing will be used as follows:

Step 1: The non-directional research question should be stated in the statistical hypothesis format.

H_0: $P_1 = P_2$ (or $P_1 - P_2 = 0$)

H_A: $P_1 \neq P_2$ (or $P_1 - P_2 \neq 0$)

Step 2: The criteria for rejecting the null hypothesis and accepting the alternative hypothesis should be determined.

Given $\alpha = 0.05$, we select the corresponding Z value which is closest to 5 per cent of the area under the normal curve (2.5 per cent in each tail of the normal curve). If our computed Z-test statistics is greater that this Z value, we would reject the null hypothesis and accept the alternative hypothesis. This establishes our region of rejection *R* or the probability areas under the normal curve where differences in sample proportions are unlikely to occur by random chance.

R: $Z \pm 1.96$

Step 3: The sample data should be collected and the Z-test statistics should be computed.

A random sample of 100 employees from a high-tech company were interviewed before and after a diversity training session and asked whether or not they favored the company hiring foreign workers. Their samples responses were as follows:

		No	Yes	
After Diversity Training	Yes	10(0.10)	20(0.20)	30(0.30)
Before Diversity Training	No	50(0.50)	20(0.20)	70(0.70)
	Total	60(0.60)	40(0.40)	100

The standard deviation of the sampling distribution of the differences in dependent sample proportion is called standard error of the difference between dependent sample proportions. This value is needed to compute the Z-test statistics. The formula is:

$$S_{\mathrm{P}_1-\mathrm{P}_2} = \sqrt{\frac{\mathrm{P}_{11}+\mathrm{P}_{22}}{\mathrm{N}}}$$

Where,

P_{11} = Percent change from before to after training (Yes → No) = 0.10

P_{12} = Percent change from before to after training (No → Yes) = 0.20

N = Total sample size =100

$$Z = \frac{P_1 - P_2}{S_{P_1-P_2}} = \frac{0.30-0.40}{0.055} = \frac{-0.10}{0.055} = -1.82$$

Step 4: The confidence interval around the Z-test statistics should be computed.

Step 5: The Z-test statistics result should be interpreted.

4.3.3 *t*-Test

Sir William S. Gosset (pen name Student) developed a significance test and through it made a significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.

Like normal distribution, *t* distribution is also symmetrical but happens to be flatter than normal distribution. Moreover, there is a different *t* distribution for every possible sample size. As the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we will use the normal to approximate the *t* distribution. Thus, when *n* is small, the *t* distribution is far from normal, but when *n* is infinite, it is identical to normal distribution.

For applying *t*-test in context of small samples, the *t* value is calculated first of all and then the calculated value is compared with the table value of *t* at certain level of significance for given degrees of freedom. If the calculated value of t exceeds the table value (say $t_{0.05}$), we infer that the difference is significant at 5 per cent level but if calculated value is t_0 is less than its concerning table value, the difference is not treated as significant.

The t-test is used when two conditions are fullfiled,

(a) The sample size is less than 30, i.e., when $n \leq 30$.

(b) The population standard deviation (σ_p) must be unknown.

In using the t-test, we assume the following:

(a) That the population is normal or approximately normal.

(b) That the observations are independent and the samples are randomly drawn samples.

(c) That there is no measurement error.

(d) That in the case of two samples, population variances are regarded as equal if equality of the two population means is to be tested.

The following formulae are commonly used to calculate the t value:

(a) **To Test the Significance of the Mean of a Random Sample**

$$t = \frac{|\,\bar{X} - \mu\,|}{S\,|\,SE_{\bar{x}}\bar{X}}$$

Where, $\bar{X}$	=	Mean of the sample
μ	=	Mean of the universe
$SE_{\bar{X}}$	=	S.E. of mean in case of small sample and is worked out as follows:

$$SE_{\bar{X}} = \frac{\sigma_s}{\sqrt{n}} = \frac{\sqrt{\dfrac{\Sigma(X_i - \bar{X})^2}{\sqrt{n}}}}{\sqrt{n}}$$

and the degrees of freedom = $(n - 1)$.

The above stated formula for t can as well be stated as under:

$$t = \frac{|\,\bar{X} - \mu\,|}{SE_{\bar{X}}} = \frac{|\,\bar{X} - \mu\,|}{\dfrac{\dfrac{\sqrt{\Sigma(X - \bar{X})^2}}{n-1}}{\sqrt{n}}} = \frac{|\,\bar{X} - \mu\,|}{\sqrt{\dfrac{\Sigma(X - \bar{X})^2}{n-1}}} \times \sqrt{n}$$

If we want to work out the probable or fiducial limits of population mean (μ) in case of small samples, we can use either of the following:

(a) Probable limits with 95 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{X}}\ (t_{0.05})$$

(b) Probable limits with 99 per cent confidence level:

$$\mu = \bar{X} \pm SE_{\bar{X}}\ (t_{0.01})$$

At other confidence levels, the limits can be worked out in a similar manner, taking the concerning table value of t just as we have taken $t_{0.05}$ in (a) and $t_{0.01}$ in (b) above.

(b) To Test the Difference between the Means of the Two Samples

$$t = \frac{|\bar{X}_1 - \bar{X}_2|}{\text{SE}_{\bar{X}_1 - \bar{X}_2}}$$

Where, $\bar{X}_1$ = Mean of the Sample 1.

$\bar{X}_2$ = Mean of the Sample 2.

$SE_{\bar{X}_1 - \bar{X}_2}$ = Standard Error of difference between two sample means and is worked out as follows:

$$SE_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sum(X_{1i} - \bar{X}_1)^2 + \sum(X_{2\,i} - \bar{X}_2)^2}{n_1 + n_2 - 2}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

and the degrees of freedom = $(n_1 + n_2 - 2)$.

When the actual means are in fraction, then use of assumed means is convenient. In such a case, the standard deviation of difference, i.e.,

$$\sqrt{\frac{\Sigma(X_{1i} + X_1)^2 + \Sigma(X_{2i} - \bar{X}_2)^2}{n_1 + n_2 - 2}}$$

can be worked out by the following short-cut formula:

$$= \frac{\sqrt{\Sigma(X_{1i} - A_1)^2} + \Sigma(X_{2i} - A_1)^2 - n_1(X_{1i} - A_2)^2 - n_2(X_{2i} - A_2)^2}{n_1 + n_2 - 2}$$

Where, A_1 = Assumed mean of Sample 1.

A_2 = Assumed mean of Sample 2.

X_1 = True mean of Sample 1.

X_2 = True mean of Sample 2.

(c) To Test the Significance of an Observed Correlation Coefficient

$$t = \frac{r}{\sqrt{1 - r^2}} \times \sqrt{n - 2}$$

Here, t is based on $(n - 2)$ degrees of freedom.

(d) To Test and Compare Sample Mean

The t-test can be used to compare a sample mean to an accepted value (of a population mean) or it can be used to compare the means of two sample sets.

t-Test to Compare One Sample Mean to an Accepted Value

The formula for the calculation of the *t*-statistic for one sample mean is as follows:

$$t = \frac{\overline{x} - \mu_0}{s / \sqrt{n}}$$

Here, *s* is the standard deviation of the sample and not the population standard deviation.

t-Test to Compare Two Sample Means

The method for comparing two sample means is very similar. The only two differences are the equation used to compute the *t*-statistic and the degrees of freedom for choosing the tabulate *t*-value. The formula is given as follows:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

In this case, we require two separate sample means, standard deviations and sample size. The formula for degrees of freedom (*df*) depends on the condition.

One-sample *t*-test: $df = n - 1$

Two-sample *t*-test: $df = \dfrac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\dfrac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\dfrac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$

Example 4.16 A one sample *t*-test is conducted on H_0: $\mu = 81.6$. The sample has $\overline{x} = 84.1$, $s = 3.1$ and $n = 25$. Find the *t*-test statistics.

Solution: The *t*-test statistics is obtained as follows:

$$\begin{aligned}
\overline{x} &= 84.1 \\
\mu &= 81.6 \\
s &= 3.1 \\
n &= 25 \\
t &= \frac{\overline{x} - \mu_0}{s / \sqrt{n}} \\
&= 84.1 - 81.6 / (3.1 / \sqrt{}25) \\
&= 2.5 / (3.1 / 5) \\
&= 2.5 / 0.62 \\
&= 4.032
\end{aligned}$$

The value of *t* as per *t*-test statistics is 4.032.

Example 4.17: Two random samples have been selected and a two sample *t*-test for the difference in population means is conducted with $H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 > \mu_2$. The results are $s_1 = 2.5$, $\overline{x}_1 = 9.7$, $n_1 = 30$, $s_2 = 2.9$, $\overline{x}_2 = 9.1$ and $n_2 = 35$. What is the value of *t*-test?

Solution: The value of *t*-test is obtained as follows:

Here,

$$\overline{x}_1 = 9.7$$

$$\overline{x}_2 = 9.1$$

$$n_1 = 30$$

$$n_2 = 35$$

$$s_1 = 2.5$$

$$s_2 = 2.9$$

We know:

$$\mathrm{t} = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n2}}}$$

$$= 9.7 - 9.1 /\ \surd\ (2.5)^2/30 + (2.9)^2 / 35$$

$$= 0.6 / \surd 6.25/ 30 + 8.41/35$$

$$= 0.6 / \surd\ 0.208 + 0.240$$

$$= 0.6 / 0.448$$

$$= 0.6 / 0.66$$

$$= 0.90$$

The value of *t* as per as per *t*-test statistics is 0.90.

Example 4.18: Comparison of the birth weights of new born babies of women who participated in an intervention with the birth weights of a group that did not participate are given below.

	Participated	**Not Participated**
Average Weight	3100 gm	2750 gm
SD	420	425
n	75	75

Calculate the value of *t* using *t*-test.

Solution: The value of *t*-test is calculated as follows:

Here,

$$\overline{x}_1 = 3100$$

$$\overline{x}_2 = 2750$$

$\text{SD}_1 = 420$

$\text{SD}_2 = 425$

$n_1 = 75$

$n_2 = 75$

Now the value of *t* can be calculated as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}\right)}}$$

$$t = \frac{3100 - 2750}{\sqrt{\left(\frac{420^2}{75} + \frac{425^2}{75}\right)}}$$

$$t = \frac{350}{\sqrt{2352 + 2408.3}}$$

$$t = 5.07$$

The value of *t* as per as per *t*-test statistics is 5.07.

Example 4.19: A sample of 10 measurements of the diameter of a sphere, gave a mean $X = 4.38$ inches and a standard deviation, $\sigma = 0.06$ inches. Find (*a*) 95 per cent and (*b*) 99 per cent confidence limits for the actual diameter.

Solution: On the basis of the given data the standard error of mean:

$$= \frac{\sigma_s}{\sqrt{n-1}} = \frac{0.06}{\sqrt{10-1}} = \frac{0.06}{3} = 0.02$$

Assuming the sample mean 4.38 inches to be the population mean, the required limits are as follows:

(a) 95 per cent confidence limits $= \bar{X} \pm SE_{\bar{X}}(t_{0.05})$ with degrees of freedom

$= 4.38 \pm .02(2.262)$

$= 4.38 \pm .04524$

i.e., 4.335 to 4.425

(b) 99 per cent confidence limits $= \bar{X} \pm SE_{\bar{X}}(t_{0.01})$ with 9 degrees of freedom

$= 4.38 \pm .02(3.25) = 4.38 \pm .0650$

i.e., 4.3150 to 4.4450.

Example 4.20: It was found that the coefficient of correlation between two variables calculated from a sample of 25 items was 0.37. Test the significance of *r* at 5 per cent level with the help of *t*-test.

Solution: To test the significance of *r* through *t*-test, we use the following formula for calculating *t* value:

$$t = \frac{r}{\sqrt{1-r^2}} \times \sqrt{n-2}$$

$$= \frac{0.37}{1-(0.37)^2} \times \sqrt{25-2}$$

$$=1.903$$

Degrees of freedom = (n–2) = (25–2) = 23

The table value of α at 5 per cent level of significance for 23 degrees of freedom is 2.069 for a two-tailed test.

The calculated value of *t* is less than its table value, hence *r is* insignificant.

t-Test for Independent and Dependent Group

Under this heading, you will learn about *t*-Test for independent and dependent group.

Independent *t*-Test

The sampling distribution of the difference between the means of two independent samples provides the basics for the testing of a mean difference hypothesis between two groups.

A typical research in which one would use the independent *t* -test might involve one group of employees receiving sales training and the second group of employees not receiving any sales training. The number of sales for each group is recorded and averaged. The null hypothesis would be stated when the average sales for the two groups are equal. The alternative hypothesis would be stated when the group receiving the sales training will on average have higher sales than the group that did not receive any sales training. If the sample data for the two groups were recorded for Sales Training where the mean = 40, standard deviation= 10, *n* = 100, then the independent *t*-test can be computed as follows:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}}$$

Thus, the independent *t*-test can be computed as follows:

$$t = \frac{50-40}{1.41} = 7.09$$

Dependent *t*-Test

The dependent *t*-test is sometimes referred to as the paired *t*-test because it uses two sets of scores on the same individuals.

A typical research situation that uses the dependent *t*-test involves a reputed measure design with one group. For example, a psychologist is studying the effect of certain motion picture upon the attribute of violence. The psychologist hypotheses that viewing the motion will cause the students attribute to be more violent. A random sample of 10 students is given an attribute towards violence inventory before viewing the motion picture. Next 10 students view the motion picture which contains graphic violence portrayed as acceptable behavior. The ten students are the given the attribute towards violence inventory after viewing the motion picture. The average attribute toward violence score for students before viewing the motion picture was 6.75, but after viewing the motion picture it is 73. There are ten pairs of scores so ten score differences are squared and summed to calculate the sum of square difference. The standard error of the dependent *t*-test is the square root of the sum of square differences divide by $N(N-1)$.

NOTES

The dependent *t*-test to investigate whether the student's attribute towards the violence changed after viewing the motion picture would be calculated as follows:

$$t = \frac{\bar{D}}{S_D}$$

The numerator in the above formula is the average difference between the post and pre means score on the attribute towards violence inventory which is 73 – 67.5 = 5.5.

The denominator is calculated as follows:

Student	Pre	Post	D	D^2
1	70	75	+ 5	25
2	60	70	+10	100
3	85	80	− 5	25
4	50	65	+15	225
5	65	75	+10	100
6	80	70	−10	100
7	90	95	+ 5	25
8	70	80	+10	100
9	40	55	+15	225
10	65	65	0	0
	$\bar{D}_1 = 67.5$	$\bar{D}_2 = 73.0$	**ΣD = 55**	**ΣD^2 = 725**

Thus the dependent *t*-test is calculated as $t = 5.5/ 2.17=2.53$

4.3.4 Sampling Distribution of Means

The sampling distribution of a statistic is the distribution of that specific statistic which is considered as a random variable when derived from a random sample of size *n*. It may be considered as the distribution of the statistic for *all possible samples from the same population* of a given size. Thus the sampling distribution depends on the underlying distribution of the population, the statistic being considered, the sampling procedure employed and the sample size used.

The mean of the sampling distribution is the mean of the population from which the scores were sampled. Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ. We will discuss here the large samples and the small samples.

Large Samples

An important principle, known as the 'Central Limit Theorem (CLT)', describes the characteristics of sample means if a large number of equal-sized samples (greater than 30) are selected at random from an infinite population. The central limit theorem states that, 'Given a population with a finite mean μ and a finite non-zero variance σ^2, the sampling distribution of the mean approaches a normal distribution with a mean of μ and a variance of σ^2/N as N, the sample size, increases. The following are some characteristic features of sample mean:

- The distribution of 'sample means' is normal and it possesses all the characteristics of a normal distribution.
- The average value of 'sample means' will be the same as the mean of the population.
- The distribution of the 'sample means' around the population mean will have its own standard deviation, known as 'standard error of mean', which is denoted as SE_M or σ_M. It is computed by the following formula:

$$SE_M = \sigma_M = \frac{\sigma}{\sqrt{N}} \qquad (1)$$

Where, σ = Standard deviation of the population.

N = The number of cases in the sample.

Since the value of σ (i.e., standard deviation of population) is usually not known, we make an estimate of this standard error of mean by using the following formula:

$$\sigma_M = \frac{\sigma}{\sqrt{N}} \qquad (2)$$

Where, σ = Standard deviation of the sample.

N = The number of cases in the sample.

To illustrate the use of Formula (2), let us assume that the mean of the attitude scores of a sample of 100 distance learners enrolled with Gauhati University towards student support services is 25 and the standard deviation is 5. The standard error of mean can be calculated accordingly:

$$SE_M = \sigma_M = \frac{5.0}{\sqrt{100}} = 0.50$$

This 'standard error of mean' may be assumed as the standard deviation of a distribution of sample means, around the fixed population mean of all distance learners. In the case of large randomly selected samples, the sampling distribution of sample means is assumed to be normal.

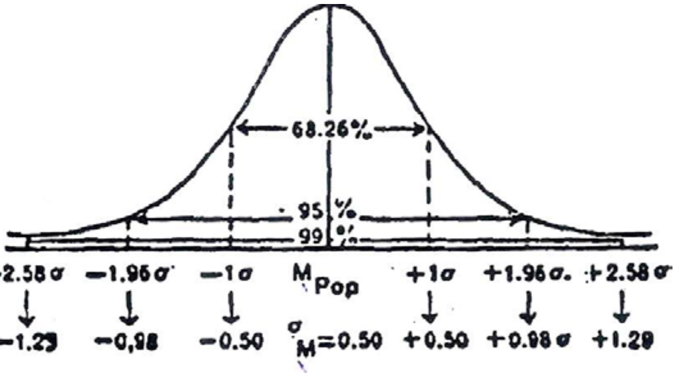


Fig. 4.5 Sampling Distribution of Means showing Variability of Obtained Means around the Population Mean in Terms of σ_M

The normal curve in Figure 4.5 shows that this sampling distribution is centred on the unknown population mean with standard deviation 0.50. The sample means often fall between the positive and the negative side of the population mean. About 2/3 of our sample means (exactly 68.26 per cent) will lie within ± 1.00 σ_M of the population mean, i.e., within a range of ±1 × 0.50 = ±0.50. Furthermore, 95 of our 100 sample means will lie within ±2.00σ_M (more exactly ±1.96 σ_M) of the population mean, i.e., 95 of 100 sample means will lie within ±1.96 × 0.50 or ±0.98 of the population mean. In other words, the probability that our sample mean of 25 does not miss the population mean (M_{Pop}.) by more than ±0.98 is 0.95. Also, 99 of our sample means will be within ±3.00σ_M (more exactly ± 2.58 σ_M) of the population mean. This indicates that 99 out of 100 sample means will fall within ± 2.58 × 0.50 or ± 1.29 of the population mean. The probability (P) that our sample mean of 25 does not miss the M_{Pop}. by more than ± 1.29 is 0.99.

Thus, the value of a population mean can be inferred from a randomly selected sample mean and can be estimated on a probability basis.

4.3.5 Confidence Intervals and Levels of Significance

When we draw a large random sample from the population to obtain measures of a variable and compute the mean for the sample, we can use the 'central limit theorem' and 'normal probability curve' to have an estimate of the population mean. We can say that M has a 95 per cent chance of being within 1.96 standard error units of M_{Pop}. In other words, a mean for a random sample has a chance of 95 per cent of being within 1.96 σ_M units from M_{Pop}. It may also be said that there is a 99 per cent chance that the sample mean lies within 2.58 σ_M units of M_{Pop}. To be more specific, it may be stated that there is a 95 per cent probability that the limits M ± 1.96 σ_M enclose the population mean and the limits M ± 2.58 σ_M enclose the population mean with 99 per cent probability. Such limits enclosing the population mean are known as the 'confidence intervals'.

These limits help us to adopt particularly two levels of confidence. One is known as 5 per cent level or 0.05 levels and the other is known as 1 per cent level or 0.01 level. The 0.05 level of confidence indicates that the probability M_{Pop}. that lies within the interval M±1.96 σ_M is 0.95 and that it falls outside of these limits is 0.05.

By saying that probability is 0.99, it is meant that M_{Pop}. lies within the interval M± 2.58 σ_M and that the probability of its falling outside of these limits is 0.01.

To illustrate, let us apply the concept to the previous problem. Taking as our limits M± 1.96 σ_M**,** we have 25 ±1.960 × 0.5 0 or a confidence interval marked off by the limits 24.02 and 25.98. Our confidence that this interval contains M_{Pop}. is expressed by a probability of 0.95. If we want a higher degree of confidence, we can take the 0.99 level of confidence for which the limits are M±2.58 σ_M or a confidence interval given by the limits 23.71 and 26.29. We may be quite confident that M_{Pop}, is not lower than 23.71 nor higher than 26.29, i.e., the chances are 99 in 100 that the M_{Pop}, lies between 23.71 and 26.19.

Small Samples

When the number of cases in the sample is less than 30, we may estimate the value of σ_M by the following formula:

$$\text{SE}_{\text{M}} = \frac{S}{\sqrt{N}} \tag{3}$$

Where,

S = Standard deviation of the small sample.

N = The number of cases in the sample.

The formula for computing S is as follows:

$$S = \frac{\sqrt{\sum x^2}}{N-1} \tag{4}$$

Where,

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

N = The number of cases in the sample.

The concept of small size was developed by William Sealy Gosset, a consulting statistician for Guinness Breweries of Dublin (Ireland) 1908. The principle is that we should not assume that the sampling distribution of means of small samples is normally distributed. He found that the distribution curves of small sample means were somewhat different from the normal curve. When the size of the sample is small then the *t*-distribution lies under the normal curve, but the tails or ends of the curve are higher than the corresponding parts of the normal curve.

Degrees of Freedom

While finding the standard deviation of small samples we use *N*–1 in the denominator instead of *N* in the basic formula for standard deviation. The difference in the two formulae may seem very little, if *N* is sufficiently large. But there is a very important difference in the 'meaning' in the case of small samples. *N*–1 is known as the 'number of degrees of freedom', denoted by '*df*'. The 'number of **degrees of freedom'** in a distribution is the number of observations or values that are independent of each other and cannot be deduced from each other. In other words, we may say that the 'degrees of freedom' connote freedom to vary.

To illustrate as to why the '*df*' used here is *N*–1, we take 5 scores, i.e., 5,6,7,8 and 9, the mean of which is 7. This mean score is to be used as an estimate of the population mean. The deviations of the scores from the mean 7 are – 2, – 1, 0, +1 and +2. A mathematical requirement of the mean is that the sum of these deviations should be zero. Of the five deviations, only 4, i.e., *N*–1 can be chosen freely (independently) as the condition that the sum is equal to zero restricts the value of the 5th deviate. With this condition, we can arbitrarily change any four of the five deviates and thereby fix the fifth. We could take the first four deviates as –2, –1, 0 and +1, which would mean that for the sum of deviates to be zero, the fifth deviate has to be +2. Similarly, we can try other changes and if the sum is to remain zero, one of the five deviates is automatically determined. Hence, only 4, i.e., (5 – 1)'s are free to vary within the restrictions imposed.

When a statistic is to be used to estimate a parameter, the number of degrees of freedom depends upon the restrictions imposed. One '*df*' is lost for each of the restrictions imposed. Therefore, the number of '*df*' varies from one statistics to another. For example, in estimating and computing the population mean (M_{Pop}) from the sample Mean (M), we lose 1 '*df*'. So, the number of degrees of freedom is (N – 1).

Let us determine the 0.95 and 0.99 confidence intervals for the population mean (M_{Pop}) of the scores 10, 15, 10, 25, 30, 20, 25, 30, 20 and 15, obtained by 10 distance learners on an attitude scale. The mean of the scores is as follows:

$$= \frac{10+15+10+25+30+20+25+30+20+15}{10}$$

$$= \frac{200}{10} = 20.00$$

Using Formula (4) we compute the standard deviation as follows:

Table 4.1 *Standard Deviation*

X	***X* = *X* – M**	***X*²**
10	–10	100
15	–5	25
10	–10	100
25	5	25
30	10	100
20	0	0
25	5	25
30	10	100
20	0	0
15	–5	25
	$\Sigma x = 0$	

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

$$= \sqrt{\frac{500}{10-1}}$$

$$= 7.45$$

From Formula (3) we compute:

$$\text{SE}_\text{M} = \frac{7.45}{\sqrt{10}}$$

$$= 2.36$$

For estimating the M_{Pop} from the sample mean of 20.00, we determine the value of '*t*' at the selected points using appropriate number of degrees of freedom. The available '*df*' for determining t is $N-1$ or 9. With 9 '*df*', we read that '*t*' = 2.26 at 0.05 level and 3.25 at 0.0l level. From the first t-value we know that 95 of our 100 sample means will lie within ± 2.26 SE_M or ± 2.26 × 2.36 of the population mean and 5 out of 100 falls outside of these limits. The probability (P) that our sample mean 20.00 does not miss the M_{Pop}. pop by more than ± 2.26 × 2.36 or ± 5.33 is 0.95. From the second t-value, we know that 99 per cent of our sample mean will lie between M_{Pop} and 3.25 SE_M or ± 3.25 × 2.36, and that 1 per cent fall will beyond these limits. So, the probability (P) that our sample mean of 20.00 does not miss the M_{Pop} by more than ±3.25 × 2.36 or ± 7.67 is 0.99.

Taking our limits as M ±2.26 SE_M, we have 20.00 ± 2.26 × 2.36 or 14.67 and 25.33 as the limits of the 0.95 confidence interval. The probability (P) that M_{Pop} is not less than 14.67 or greater than 25.33 is 0.95. Taking the limits M ± 3.25 SE_M, we have 20.00 ± 3.25 × 2.36, or 12.33 and 27.67 as the limits of the 0.99 confidence interval and the probability (P) so that M_{Pop} is not less than 12.33 and not greater than 27.67 is 0.99.

The use of small samples to build generalizations in educational research should be made cautiously as it is difficult to ensure that a small sample adequately represents the population from which the sample is drawn. Furthermore, conclusions drawn from small samples are usually unsatisfactory because of the great variability from sample to sample. In other words, large samples drawn randomly from the population will provide a more accurate basis than will small samples for inferring population parameters.

4.3.6 Two-Tailed and One-Tailed Tests of Significance

Suppose a null hypothesis were set up that there was no difference other than a sampling error difference between the mean height of two groups, A and B. We would be concerned only with the difference and not with the superiority or inferiority in height of either group. To test this hypothesis, we apply two-tailed test as the difference between the obtained means of height of two groups may be as often in one direction (plus) as in the other (minus) from the true difference of zero. Moreover, for determining probability, we take both tails of sampling distribution.

For a large sample two-tailed test, we make use of a normal distribution curve. The 5 per cent area of rejection is divided equally between the upper and the lower tails of this curve and we have to go out to ±1.96 on the base line of the curve to reach the area of rejection as shown in the Figure 4.6.

NOTES

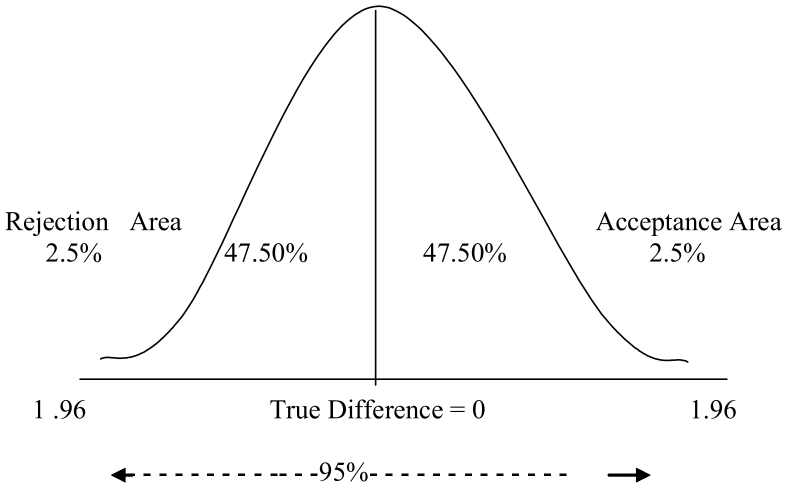


Fig. 4.6 *A Two-Tailed Test at 0.05 Levels (2.5 Per Cent at Each Level)*

Similarly, if we have 0.5 per cent area at each end of the normal curve where 1 per cent area of rejection is to be divided equally between its upper and lower tails, it is necessary to go out to ± 2.58 on the base line to reach the area of rejection as shown in the Figure 4.7.

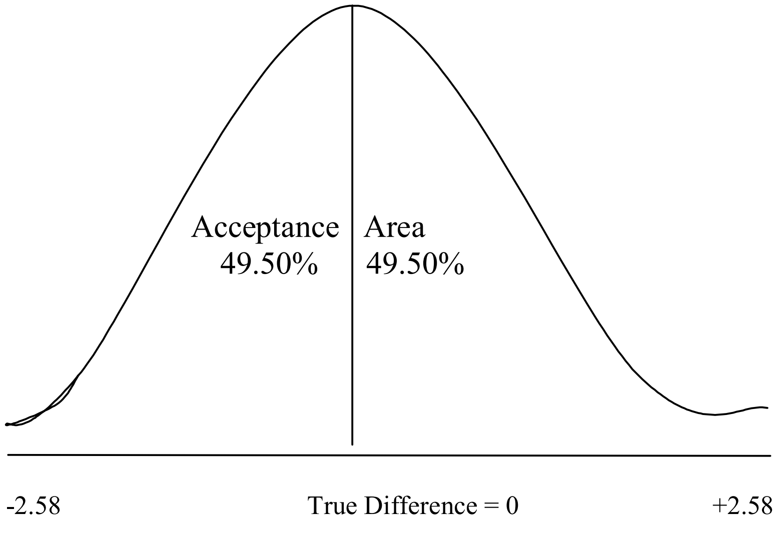


Fig. 4.7 *A Two-Tailed Test at 0.01 level (0.5 Per Cent at Each Level)*

In the case of the above example, a null hypothesis was set up where there was no difference other than a sampling error difference between the mean creative thinking M.Ed. score of males and females. Thus, we were concerned, with a difference and not in superiority or inferiority of either group in the creative thinking ability. To test this hypothesis, we applied 'two-tailed test' as the difference between the two means might have been in one direction (plus) or in the other (minus) from the true difference of zero and we took both tails of sampling distribution in determining probabilities.

NOTES

As is evident from the above example, we make use of a normal distribution curve in the case of a large sample 'two-tailed test'. The 5 per cent area of rejection is equally divided between the upper and lower tails of the curve and we have to go out to ±1.96 on the base line of the curve to reach the area of rejection.

Similarly, we have 0.5 per cent area at each end of the normal curve when 1 per cent of rejection is to be divided equally between its upper and lower tails and it is necessary to go out to ± 2.58 on the base line to reach the area of rejection.

In the above problem, if we change the null hypothesis as: male group of M.Ed. have significantly higher creative thinking than that of the female group; or male group have significantly lower creative thinking than the female group of M.Ed. course, then each of these hypotheses indicates a direction of difference. In such situations, the use of 'one-tailed test' is made. For such a test, the 5 per cent area or 1 per cent area of rejection is either at the upper tail or at the lower tail of the curve, to be read from 0.10 column (instead of 0.05) and 0.02 column (instead of 0.0l).

Application of *t*-Test for Testing the Significance of Difference Between Two Independent Small Samples

The frequency distribution of small sample means drawn from the same population forms a *t*-distribution, and it is reasonable to expect that the sampling distribution of the difference between the means computed from two different populations will also fall under the category of *t*-distribution. Fisher provided the formula for testing the difference between the means computed from independent small samples as follows.

$$t = \frac{[M_1 - M_2]}{\frac{\sqrt{\left(\Sigma x_1^2 + \Sigma x_2^2\right)}}{N_1 + N_2 - 2}\left(\frac{N_1 + N_2}{N_1 \times N_2}\right)} \qquad (5)$$

Where,

M_1 and M_2	=	Means of two samples.
$\Sigma x_1{}^2$ and $\Sigma x_2{}^2$	=	Sums of squares of the deviations from the means in the two samples.
N_1 and N_2	=	Number of cases in the two samples.
df	=	Degrees of freedom = $N_1 + N_2 - 2$.

To illustrate the use of the Formula (5), let us test the significance of the difference between mean scores of 7 boys and 10 girls in an intelligence test as illustrated in Table 4.2.

Table 4.2 Scores of 7 Boys and 10 Girls in an Intelligence Test

	$(N_1 = 7)$			$(N_2 = 10)$	
Boys X_1	x_1	x_1^2	**Girls** X_2	x_2	x_2^2
13	0	0	10	–4	16
14	1	1	16	2	4
11	–2	4	12	–2	4
12	–1	1	13	–1	1
15	2	4	18	4	16
13	0	0	13	–1	1
13	0	0	19	5	25
			14	0	0
			13	–1	1
			12	–2	4
$\Sigma X_1 = 91$		$\Sigma X_1^2 = 10$	$\Sigma X_2 = 140$		$\Sigma X_2^2 = 72$

$$M_1 = \frac{91}{7} = 13$$

$$M_2 = \frac{140}{10} = 14$$

$df = N_1 + N_2 - 2 = 7 + 10 - 2 = 15$

Using Formula (5) we compute t as follows:

$$t = \frac{[14-13]}{\sqrt{\left(\frac{10+72}{7+10-2}\right)\left(\frac{7+10}{7\times 10}\right)}}$$

$$= \frac{1}{\sqrt{\left(\frac{82}{15}\right)\left(\frac{17}{10}\right)}}$$

$$= \frac{1}{\sqrt{9.29}}$$

$$= 0.33$$

To test the significance of difference between the two means by making use of the 'two-tailed test' (null hypothesis, i.e., no differences between the two groups), we look for the t-critical values for rejection of null hypothesis for (7 + 10 – 2) or 15 df. These t-values are 2.13 at 0.05 and 2.95 at 0.01 levels of significance. Since the obtained t-value 0.33 is less than the table value necessary for the rejection of the null hypothesis at 0.05 levels for '*df*' 15, the null hypothesis is accepted and it may be concluded that there is no significant difference in the mean intelligence scores of males and females. If we change the null hypothesis as: boys will have higher

NOTES

intelligence scores than girls or males will have lower intelligence scores than females, then each of these hypotheses indicates a direction of difference rather than simply the existence of the difference. So, we make use of 'one-tailed test'. For given degrees of freedom, i.e., 15, the 0.05 level is read from the 0.10 column ($^p/_2 = 0.05$) and the 0.01 level from 0.02 column($^p/_2 = 0.01$) of the '*t*' table. In the one-tailed test, for 15 '*df*' *t*-critical values at 0.05 and 0.01 levels, as read from the 0.10 and the 0.02 columns are 1.75 and 2.60, respectively. Since the computed *t*-value of 0.33 does not reach the table value at 0.05 levels (i.e., 1.75 for 0.l0), we may conclude that the difference in two groups is present merely because of chance factors.

CHECK YOUR PROGRESS

3. What is Z-test?
4. What is a dependent *t*-test?
5. What is 'the number of degrees of freedom'?

4.4 SUMMARY

- Among all the probability distributions, the normal probability distribution is by far the most important and frequently used continuous probability distribution. This is so because this distribution fits well in many types of problems. This distribution is of special significance in inferential statistics since it describes probabilistically the link between a statistic and a parameter (i.e., between the sample results and the population from which the sample is drawn).
- Some of the characteristics of the normal distribution or that of a normal curve include: (a) It is a symmetric distribution. (b) The curve is asymptotic to the base line which means that it continues to approach but never touches the horizontal axis. (c) The variance (σ^2) defines the spread of the curve.
- We can have several normal probability distributions but each particular normal distribution is being defined by its two parameters viz., the mean (μ) and the standard deviation (σ). There is, thus, not a single normal curve but rather a family of normal curves.
- The area under the normal curve (often termed as the standard normal probability distribution table) is organized in terms of standard variate (or Z) values. It gives the values for only half the area under the normal curve, beginning with Z = 0 at the mean. Since the normal distribution is perfectly symmetrical the values true for one half of the curve are also true for the other half.
- Some of the areas of application of normal distribution include: (a) Random processes such as body temperature of a healthy adult. (b) Approximation of binomial distribution. (c) Standardization (d) Composite scores.

- The Normal Probability Curve (NPC), simply known as normal curve, is a symmetrical bell-shaped curve. This curve is based upon the law of probability and discovered by French mathematician Abraham Demoivre (1667–1754) in the 18th century. In this curve, the mean, median and mode lie at the middle point of the distribution. The total area of the curve represents the total number of cases and the middle point represents the mean, median and mode.
- The NPC or Normal Probability Curve has several features which are essential to understand for its use. The major characteristics are limited. Some of them include: (a) It is a bell shaped curve. (b) The measures of central tendency are equal, i.e., mean, mode and median concentrate on one point. (c) The height of the curve is 0.3989.
- The Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by means of a normal distribution. Typically, it is a statistical test used to determine whether two population means are different when the variances are known and the sample
- The Z-score is the number of standard deviations which are away from the mean and can be used to compare different scores when the mean and standard deviation of the population are known. These comparisons are based on the assumptions that the distribution of the population is normal and that the two scores are drawn from tests that measure the same construct(s).size is large.
- Sir William S. Gosset (pen name Student) developed a significance test and through it made a significant contribution to the theory of sampling applicable in case of small samples. When population variance is not known, the test is commonly known as Student's *t*-test and is based on the *t* distribution.
- Like normal distribution, *t* distribution is also symmetrical but happens to be flatter than normal distribution. Moreover, there is a different t distribution for every possible sample size. As the sample size gets larger, the shape of the *t* distribution loses its flatness and becomes approximately equal to the normal distribution. In fact, for sample sizes of more than 30, the *t* distribution is so close to the normal distribution that we will use the normal to approximate the *t* distribution. Thus, when n is small, the *t* distribution is far from normal, but when n is infinite, it is identical to normal distribution.
- The sampling distribution of a statistic is the distribution of that specific statistic which is considered as a random variable when derived from a random sample of size n. It may be considered as the distribution of the statistic for all possible samples from the same population of a given size. Thus the sampling distribution depends on the underlying distribution of the population, the statistic being considered, the sampling procedure employed and the sample size used.
- The 'number of degrees of freedom' in a distribution is the number of observations or values that are independent of each other and cannot be deduced from each other. In other words, we may say that the 'degrees of freedom' connote freedom to vary.

4.5 KEY TERMS

- **Normal Probability Curve (NPC):** Simply known as normal curve, it is a symmetrical bell shaped curve based upon the law of probability discovered by French mathematician Abraham Demoivre (1667–1754) in the 18th century.
- **Z-test:** Any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution.
- **Z-score:** It is the number of standard deviations which are away from the mean and can be used to compare different scores when the mean and standard deviation of the population are known.
- ***t*-test:** Any statistical hypothesis test in which the test statistic follows a Student's *t* distribution, if the null hypothesis is supported.

4.6 ANSWERS TO 'CHECK YOUR PROGRESS'

1. Three characteristics of normal distribution are:
   (a) It is a symmetric distribution.
   (b) The curve is asymptotic to the base line which means that it continues to approach but never touches the horizontal axis.
   (c) The variance (σ^2) defines the spread of the curve.
2. A Normal Probability Curve (NPC), simply known as normal curve, is a symmetrical bell-shaped curve. This curve is based upon the law of probability and discovered by French mathematician Abraham Demoivre (1667–1754) in the 18th century. In this curve, the mean, median and mode lie at the middle point of the distribution. The total area of the curve represents the total number of cases and the middle point represents the mean, median and mode.
3. The Z-test is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by means of a normal distribution. Typically, it is a statistical test used to determine whether two population means are different when the variances are known and the sample size is large.
4. The dependent *t*-test is sometimes referred to as the paired *t*-test because it uses two sets of scores on the same individuals. A typical research situation that uses the dependent *t*-test involves a reputed measure design with one group.
5. The 'number of degrees of freedom' in a distribution is the number of observations or values that are independent of each other and cannot be deduced from each other. In other words, we may say that the 'degrees of freedom' connote freedom to vary.

4.7 QUESTIONS AND EXERCISES

Short-Answer Questions

1. Why is the normal distribution of special significance in inferential statistics?
2. Which two parameters define the normal distribution?
3. How is the area under the normal curve measured?
4. What is normal probability curve?
5. What is the significance of Z-test?
6. What is Z-score?
7. How is one-sample Z-test calculated?
8. What do you mean by Z-test for independent and dependent groups?
9. Who developed the *t*-test? When is it used?
10. How will you use *t*-test for independent and dependent groups?
11. What is the significance of sampling distribution of means?

Long-Answer Questions

1. State the distinctive features of the normal probability distributions.
2. Explain the circumstances when normal probability distribution can be used.
3. Discuss the various applications of normal distribution.
4. In a distribution exactly normal, 7 per cent of the items are under 35 and 89 per cent are under 63. What are the mean and standard deviation of the distribution?
5. Fit a normal distribution to the following data:

Height in Inches	Frequency
60–62	5
63–65	18
66–68	42
69–71	27
72–74	8

6. Discuss the significance of Z-test with the help of examples.
7. Explain how *t* value is calculated in *t*-test.
8. Discuss the steps that should be followed for Z-test of an independent group.

4.8 FURTHER READING

NOTES

Mood, Alexander M., Franklin A. Graybill and Duane C. Boes. 1974. *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Gupta, S.C. 2005. *Fundamentals of Statistics*, 17th edition. Mumbai: Himalaya Publishing House.

Walker, H. M. and J. Lev. 1953. *Statistical Inference*. New York: Henry Holt.

Health, R. W. and N. M. Downie. 1970. *Basic Statistical Methods, 3rd Edition*. New York: Harper International.

UNIT 5 ANALYSIS OF VARIANCE

NOTES

Structure

5.0 INTRODUCTION

In this unit, you will learn about analysis of variance, its types, basic principles and assumptions. You will also learn about non-parametric tests such as Chi-square (χ^2) Test, Mann-Whitney *U* test, Rank-difference methods (both ρ and T), Coefficient of concordance (*W*), Median Test, Kruskal-Wallis *H* Test and Friedman test.

In business decisions or in education, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the differences among various population means. The methodology used for such types of determinations is known as ANalysis Of VAriance or ANOVA. This technique is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the *F*-Test. The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. It helps to know whether any of the differences between the means of the given sample are significant.

Non-parametric tests are called distribution-free tests as they do not require any assumption regarding the shape of the population distribution from where the sample is drawn. However, some non-parametric tests do depend on a parameter such as median but they do not require a particular distribution for their application. These tests could also be used for the small sample sizes where the normality assumption does not hold true.

5.1 UNIT OBJECTIVES

NOTES

After going through this unit, you will be able to:

- Discuss Analysis Of Variance (ANOVA), its basic principle and its assumptions
- Explain the various important non-parametric tests
- Describe the uses of parametric and non-parametric tests

5.2 ANALYSIS OF VARIANCE: PARAMETRIC TESTS AND WHEN TO USE THEM

In business decisions, we are often involved in determining if there are significant differences among various sample means, from which conclusions can be drawn about the differences among various population means. For example, we may be interested to find out if there are any significant differences in the average sales figures of four different salesman employed by the same company, or we may be interested to find out if the average monthly expenditures of a family of four in five different localities are similar or not, or the telephone company may be interested in checking, whether there are any significant differences in the average number of requests for information received in a given day among the five areas of City (Under Study), and so on. The methodology used for such types of determinations is known as ANalysis Of VAriance or ANOVA. This technique is one of the most powerful techniques in statistical analysis and was developed by R. A. Fisher. It is also called the *F*-Test.

There are two types of classifications involved in the analysis of variance. The one-way analysis of variance refers to the situations when only one fact or variable is considered. For example, in testing for differences in sales for three salesman, we are considering only one factor, which is the salesman's selling ability. In the second type of classification, the response variable of interest may be affected by more than one factor. For example, the sales may be affected not only by the salesman's selling ability, but also by the price charged or the extent of advertising in a given area.

5.2.1 Basic Principle of ANOVA

The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples. In terms of variation within the given population it is assumed that the values of (X_{ij}) differ from the mean of this population only because of random effects i.e., there are influences on (X_{ij}) which are unexplainable, whereas in examining differences between populations we assume that the difference between the mean of the *j*th population and the grand mean is attributable to what is called a 'specific factor' or what is technically described as treatment effect. Thus, while using ANOVA, we assume that each of the samples is drawn from a normal population and that each of these populations has the same variance. We also assume that all factors other than the one or more being tested are effectively controlled. This, in other words, means that we assume the absence of many factors that might affect our conclusions concerning the factor(s) to be studied.

Thus, a composite procedure for testing simultaneously the difference between several sample means is known as the ANOVA. It helps us to know whether any of the differences between the means of the given sample are significant. If the answer is yes, we examine pairs (with the help of the *t*-test) to see just where the significant difference lie. If the answer is no, we do not proceed further.

In such a test, as the name implies, we usually deal with the analysis of the variances. Variances are simply the arithmetic average of the squared deviation from their means. In other words, it is the square of standard deviation (Variance = σ^2). Variance has a quality which makes it especially useful. It has an additive property, which the standard deviation with its square root does not possess. Variance on this account can be added up and broken down into components. Hence, the term 'analysis of variance' deals with the task of analyzing of breaking up the total variance of a large sample or a population consisting of a number of equal groups or sub-samples into two components (two kinds of variances), given as follows:

- **'Within Groups' Variance:** This is the average variance of the members of each group around their respective group means, i.e., the mean value of the scores in a sample (as members of each group may vary among themselves).
- **'Between Groups' Variance:** This represents the variance of group means around the total or grand mean of all groups, i.e., the best estimate of the population mean (as the group means may vary considerably from each other).

The technique of analysis of variance is applied to determine if any two of the seven means differ significantly from each other by a single test, known as *F*-test, rather than 21 *t*-tests. The *F*-test makes it possible to determine whether the sample means differ from one another (between group variance) to a greater extent than the test scores differ from their own sample means (within group variance) using the ratio given below:

$$F = \frac{\text{Variance between the groups}}{\text{Variance within groups}}$$

Assumptions for the Analysis of Variance (*F*-Test)

Certain basic assumptions underlying the technique of analysing variance are as follows:

- The population distribution should be normal. This assumption is, however, not so important. The study of Norton (Guilford, 1965) also points out that '*F*' is rather insensitive to variations in the shape of population distribution.
- All groups with a certain criterion or of the combination of more than one criterion should be randomly chosen from the sub-population having the same criterion or having the same combination of more than one criterion. For

example, if we wish to select two groups from two schools, one belonging to a rural area school and the other to an urban area school, we must, choose the groups randomly from the respective schools.

- The sub-groups under investigation must have the same variability. In other words, there should be homogeneity of variance.

5.2.2 One-Way ANOVA

One-way analysis of variance, also abbreviated as one-way ANOVA, is a technique used to compare means of two or more samples using the F distribution. This technique can be used only for numerical data.

Example 5.1: To illustrate the use of F-test, let us consider an example of 20 students who have been randomly assigned to four groups of five each, to be taught by different methods, i.e., A, B, C and D. Their performance scores on an achievement test, administered after the completion of experiment are given in Table 5.1.

Table 5.1 *Achievement Test Scores of the Four Groups Taught through Four Different Methods*

	Methods or Groups				
	A	**B**	**C**	**D**	
	(X_1)	**(X_2)**	**(X_3)**	**(X_4)**	
	14	19	12	17	
	15	20	16	17	
	11	19	16	14	
	10	16	15	12	
	12	16	12	17	
ΣX	62	90	71	77	300
ΣX^2	786	1634	1025	1207	4652

We may compute the analysis of variance using the following steps:

1. Correction $= \dfrac{(\Sigma X)^2}{N} = \dfrac{(300)^2}{200} = 4500$

2. Total sum of squares (Total SS)

$$= \Sigma X^2 - \text{Correction}$$
$$= (786 + 1634 + 1025 + 1207) - 4500$$
$$= 4652 - 4500$$
$$= 152$$

3. Sum of squares between means of treatments (Methods) A, B, C and D (between means):

$$= \frac{(\Sigma X_1)^2}{N_1} + \frac{(\Sigma X_2)^2}{N^2} + \frac{(\Sigma X_3)^2}{N^3} + \frac{(\Sigma X_4)^2}{N^4} - \text{Correction}$$

$$= \frac{(62)^2}{5} + \frac{(90)^2}{5} + \frac{(71)^2}{5} + \frac{(77)^2}{5} - 4500$$

$$= 4582.8 - 4500$$

$$= 82.8$$

NOTES

4. Sum of squares within treatments (Methods) A, B, C and D (SS within means):

   = Total SS – SS between means

   = 152 – 82.8

   = 69.2

5. Calculation of variances from each SS and analysis of the total variance into its components.

Each SS becomes a variance when divided by the degrees of freedom (df) allotted to it. There are 20 scores in all, in Table 5.1, and hence there are (N – 1) or (20 – 1) = 19 'df' in all. These 19 'df' are allocated in the following ways:

If N = Number of scores in all and K = number of treatments or groups, we have 'df' for total SS = N–1 = 20 – 1 = 19, '*df*' for within treatments = $N - K$ = 20 – 4 = 16; and 'df' for between the means of treatments = K – 1 = 4 – 1 = 3.

The variance among means of treatments is 82.8/3 or 27.60; and the variance within means is 69.2/16 or 4.33.

The summary of the analysis of variance may be presented in tabular form as shown in Table 5.2.

Table 5.2 *Summary of Analysis of Variance*

Source of Variance	***df***	**Sum of Squares (SS)**	**Mean Square (Variance)**
Between the means of treatment	3	82.8	27.60
Within treatment	16	69.2	4.33
Total	19	152.0	

Using formula,

$$F = \frac{27.60}{4.33} = 6.374$$

In the present problem, the null hypothesis asserts that four sets of scores are in reality the scores of four random samples drawn from the same normally distributed schools, and that the means of the four groups A, B, C, and D will differ only through fluctuations of sampling. For testing this hypothesis, we divided the 'between means' variance by the 'within treatments' variance and compared the resulting variance ratio, called F, with the F-values. The F value of 6.374 in the present case is to be checked for table value for 'df' 3 and 16 (the degrees of freedom for numerator and denominator). The table values for 0.05 and 0.01 levels of significance are 3.24 and 5.29. Since the computed F-value of 6.374 is greater than the table values, we reject the null hypothesis and conclude that the means of the four groups differ significantly.

5.2.3 Two-Way ANOVA

NOTES

We have studied one way ANOVA involving four different methods of teaching. In two-way analysis of variance classification, an estimate of population variance, i.e., total variance is supposed to be broken up into (i) Variance due to adjustment, (ii) Variance due to anxiety alone, and (iii) The residual variance called interaction variance (Adj × Anx), where A = Adjustment and Anx = Anxiety.

Example 5.2: A study has been conducted on anxiety and adjustment with the help of 2 × 2 factorial designs. It has four conditions and the score is given below in the Table.

Score

Adjustment			
		High	Low
Anxiety	High	A	B
	Low	C	D

ABCD are four experimental conditions. Calculate:

(i) Is there any significant difference between the means of the two conditions of adjustment?

(ii) Is there any significant difference between the means of the two conditions of anxiety?

(iii) Is there any interaction between two independent variables (adjustment and anxiety)?

Solution: We first construct the table as follows:

A Condition X_1	B Condition X_2	C Condition X_3	D Condition X_4	X_1^2	X_2^2	X_3^2	X_4^2
6	4	6	6	136	16	36	36
6	3	8	6	36	09	64	36
6	6	5	6	36	36	25	36
6	6	8	6	36	36	64	36
5	4	6	4	25	16	36	16
6	3	6	5	36	09	36	25
6	4	6	6	36	16	36	36
5	3	6	6	25	09	36	36
6	5	6	6	36	25	36	36
6	4	6	5	36	16	36	25
$\Sigma X_1 = 58$	$\Sigma X_2 = 42$	$\Sigma X_3 = 63$	$\Sigma X_4 = 56$				
$X_1^2 = 338$	$X_2^2=188$	$X_3^2=405$	$X_4^2 = 318$				
$N_1 = 10$	$N_2 = 10$	$N_3 = 10$	$N_4 = 10$				

$$C = \frac{\left(\Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \Sigma X_4\right)^2}{N}$$

$$C = \frac{\left(58 + 42 + 63 + 56\right)^2}{40} = \frac{47961}{40}$$

$$= 1199.02$$

Total SS $= \Sigma X_1{}^2 + \Sigma X_2{}^2 + \Sigma X_3{}^2 + \Sigma X_4{}^2 + \ldots. -$ Correction

$= 338 + 188 + 405 + 318 - 1199.02$

$= 1249 - 1199.02 = 49.98$

Among SS $= \dfrac{\left(\Sigma X_1\right)^2}{N_1} + \dfrac{\left(\Sigma X_2\right)^2}{N_2} + \dfrac{\left(\Sigma X_3\right)^2}{N_3} + \dfrac{\left(\Sigma X_4\right)^2}{N_4} -$ Correction

$= = \dfrac{(58)^2}{10} + \dfrac{(42)^2}{10} + \dfrac{(63)^2}{10} + \dfrac{(56)^2}{10} -$ Correction

$= 3364 + 1764 + 3969 + 3136/40 -$ Correction

$= 336.4 + 176.4 + 396.9 + 313.6 -$ Correction

$= 1223.3 - 1199.02 = 24.28$

Within SS $=$ Total SS – Among SS

$= 49.98 - 24.28 = 25.70$

SS between amount of first IV (Adjustment)

$= \dfrac{\left(\Sigma X_1 + \Sigma X_3\right)^2}{N_1 + N_3} + \dfrac{\left(\Sigma X_2 + \Sigma X_4\right)^2}{N_2 + N_4} -$ Correction

$= \dfrac{\left(58 + 63\right)^2}{10 + 10} + \dfrac{\left(42 + 56\right)^2}{10 + 10} -$ Correction

$= 1212.25 - 1199.02$

$= 13.23$

SS between amount of second IV (Anxiety)

$= \dfrac{\left(\Sigma X_1 + \Sigma X_2\right)^2}{N_1 + N_2} + \dfrac{\left(\Sigma X_3 + \Sigma X_4\right)^2}{N_3 + N_4} -$ Correction

$= \dfrac{\left(58 + 42\right)^2}{10 + 10} + \dfrac{\left(63 + 56\right)^2}{10 + 10} - 1199.0211$

$= \dfrac{10000}{20} + \dfrac{14161}{20} - 1199.0211$

$= 500 + 708.05 - 1192.02$

$= 1208.05 - 1192.02$

$= 9.03$

Interaction SS = Among SS – Between SS for first IV – Between second IV = 24.28 – 13.23 – 9.03 = 2.02.

Summary: Analysis of Variance

Source of Variation	Sum of Square	df	Mean Square	F	Result
Between Adj	13.23	1	13.23	18.63	Significant at 0.01 level
Between Anx	9.03	1	9.03	12.71	Significant at 0.01 level
Intraction: Adj × Anx	2.02	1	2.02	2.84	NS
Within group error	25.70	36	00.71	–	–
Total	49.98	39	–	–	–

$$F \text{ Ratio} = \frac{\text{Mean Square between group}}{\text{Mean Square with in group}} = \frac{2.02}{0.71}$$

$$= 2.84$$

In this way we have calculated the rest *F* Ratio also.

Result: After seeing the above table we conclude that our first two hypotheses have been accepted and that there is a significant difference whereas our third hypothesis that there is interaction between the two independent variables (Adjustment and Anxiety) is not true.

ANOVA Technique in Context of Two-Way Design when Repeated Values are There

In case of a two-way design with repeated measurements for all the categories, we can obtain a separate independent measure of inherent or smallest variations. For this measure, we can calculate the sum of squares and degrees of freedom in the same way as we have worked out the sum of squares for variance within samples in the case of one-way ANOVA, *SS* total, *SS* between columns and *SS* between rows can also be worked out as stated above. We then find left-over sums of squares and left-over degrees of freedom which are used for what is known as **interaction variation**. Interaction is the measure of inter-relationship among the two different classifications. After making all these computations, ANOVA table can be set up for drawing inferences.

Check Your Progress

1. What is the basic principle of ANOVA?
2. List any two assumptions for the Analysis of Variance (*F*-Test).

5.3 NON-PARAMETRIC TESTS: WHEN TO USE NON-PARAMETRIC TESTS IN EDUCATION

The important non-parametric statistics are as follows:

- Chi-square (χ^2) Test
- Mann-Whitney *U* test
- Rank-difference methods (both ρ and T)
- Coefficient of concordance (*W*)
- Median Test
- Kruskal-Wallis *H* Test
- Friedman test

5.3.1 Chi-square (χ^2) Test

The chi-square is one of the most important non-parametric statistics, which is used for several purposes. For this reason, Guilford (1956) has called it the general-purpose statistic. It is a non-parametric statistic because it involves no assumption regarding the normalcy of distribution or homogeneity of the variances. The chi-square test is used when the data is expressed in terms of frequencies of proportions or percentages. The chi-square applies only to discrete data. However, any continuous data can be reduced to the categories in such a way that they can be treated as discrete data and then, the application of chi-square is possible. The formula for calculating χ^2 is given as follows:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \quad ...(5.1)$$

where χ^2 = chi-square; f_o = obtained or observed frequency; and f_e = expected frequency or theoretical frequency. There are several uses of the chi-square test.

First, chi-square may be used as a test of equal probability hypothesis. By equal probability hypothesis one means the probability of having the frequencies in all the given categories as equal. Suppose, for example, 100 students answer an item on an attitude scale. The item has five categories of response options—strongly agree, agree, neutral, disagree and strongly disagree. According to the equal probability hypothesis, the expected frequency of responses given by 100 students would be 20 in each. The chi-square test would test whether or not the equal probability assumption is tenable. If the value of the chi-square test is significant, the equal probability hypothesis becomes untenable and if the value of the chi-square is not significant, the equal probability hypothesis becomes tenable.

The second use of the chi-square test is in testing the significance of the **independence hypothesis**. By independence hypothesis is meant that one variable is not affected by, or related to, another variable and hence, these two variables are independent. The chi-square is not a measure of the degree of relationship in these conditions. It merely provides an estimate of some factors other than chance (or sampling error), which account for the possible relationship. Generally, in dealing with data related to independent hypothesis, they are first arranged in a contingency table. When observations on two variables are classified in a two-way table, data are called the contingency data and the table is known as the contingency table. Independence in a contingency table exists only when each tally exhibits a different event or individual.

The third important use of chi-square is in testing a hypothesis regarding the normal shape of a frequency distribution. When chi-square is used in this connection, it is commonly referred to as a test of goodness-of-fit.

The fourth use of chi-square is in testing the significance of several statistics. For example, for testing the significance of the phi-coefficient, coefficient of concordance, and coefficient of contingency, one converts the respective values into chi-square values.

If the chi-square value appears to be a significant one, one should also take their original values as significant. To illustrate this, a 3 × 3 contingency table (table 5.3), which shows data of 200 students who were classified into three classes on the basis of their educational qualification is used. The students' educational attainments are measured in the course of their study by classifying them as superior, average or inferior.

***Table 5.3** The Use of Chi-square in a 3×3 Contingency Table*

	Superior	Average	Inferior	
Master	30 (25)	15 (15)	5 (10)	50
Bachelor	25 (25)	10 (15)	15 (10)	50
Intermediate	45 (50)	35 (30)	20 (20)	100
	100	60	40	200

(The figures in brackets show expected frequency)

The question posed is: Is educational achievement related to educational qualification? The obtained data have been shown above. The first step in calculating χ^2 as a test of significance of independence or the relationship between educational achievement and educational qualification is to compute the expected frequency. The null hypothesis is that these two variables are not related or are independent, and if this hypothesis is true, the expected frequencies should be as follows:

Cells of Table	Expected Frequency
Upper left	$(100\times50)/200=25$
Upper middle	$(60\times50)/200=15$
Upper right	$(40\times50)/200=10$
Middle left	$(100\times50)/200=25$
Middle middle	$(60\times50)/200=15$
Middle right	$(40\times50)/200=10$
Lower left	$(100\times100)/200=50$
Lower middle	$(60\times100)/200=30$
Lower right	$(40\times100)/200=20$
	$\Sigma = 200$

After calculating expected frequency for each cell, the chi-square may be calculated in the following manner:

f_o	f_e	$f_o - f_e$	$(f_o - f_e)^2$	$\frac{(f_o - f_e)^2}{f_e}$
30	25	+5	25	1
15	15	0	0	0
5	10	–5	25	2.5
25	25	0	0	0
10	15	–5	25	1.67
15	10	+5	25	2.5
45	50	–5	25	0.5
35	30	+5	25	0.83
20	20	0	0	0
$\Sigma f_o = 200$	$\Sigma f_e = 200$	$\Sigma(f_o - f_e) = 0$		$\Sigma = 9.00$
$df = (r - 1)(K = 1) = (3 - 1)(3 - 1) = 2 \times 2 = 4$				

Entering the probability table of chi-square, one finds that the value of chi-square for d.f. = 4 at the 0.05 level should be 9.488. As the obtained chi-square is below it ($p >> 0.05$), one concludes that the null hypothesis is retained. Hence, the two variables, namely, educational qualification and educational attainment in the present study are found to be independent. For calculating d.f. in a chi-square test, the formula as noted above is $(r - 1)(k - 1)$ where r = the number of rows and k is the number of columns.

Chi-square

When the data has been arranged in a 2×2 contingency table (where d.f. = 1), we need not calculate the expected frequency in the manner described above. In such a situation, the chi-square can be directly calculated with the help of the following equation:

$$\chi^2 = \frac{N[|AD - BC|]^2}{(A+B)(C+D)(A+C)(B+D)} \qquad ...(5.2)$$

where A, B, C and D are symbols for frequency of four cells in a 2×2 table; N = total number of frequencies; bars (II) indicate that in subtracting BC from AD the sign is ignored.

Suppose the researcher wants to know whether or not the two given items in the test are independent. Both items have been answered in 'Yes' or 'No' form. The test was administered to a sample of 400 students and the obtained data were as follows:

Chi-square in a 2×2 Table

		Item No. 6		
		Yes	No	
Item No. 10	No	180 A	120 B	300
	Yes	90 C	10 D	100
		270	130	400

According to the formula:

$$\text{d.f.} = (r-1)(k-1) = (2-1)(2-1) = 1$$

Entering the probability table of chi-square, we find that for d.f., the value of chi-square at the 0.001 level should be 10.827. As the obtained value of the chi-square is much above it, we conclude that item nos. 6 and 10 are not independent, that is, they are related.

Sometimes it happens that with 1 d.f., any one of the expected cell frequencies becomes less than 5. In such a situation a correction called **Yates' correction for continuity** is applied. Some writers have suggested that Yates' correction for continuity should be applied when any of the expected frequencies goes below 10. Where frequencies are large, this correction makes no difference but where frequencies are small, Yates' correction is significant. Yates' correction consists in reducing the absolute value of difference between f_o and f_e by 0.5, that is, each f_o which is larger than f_e is decreased by 0.5 and each f_o which is smaller than f_e is increased by 0.5. The formula for chi-square in such a situation is given as follows:

$$\chi^2 = \frac{N\left[|AD - BC| - \frac{N}{2}\right]^2}{(A+B)(C+D)(A+C)(B+D)} \qquad \text{...(5.3)}$$

where subscripts are defined as usual. Suppose, 60 students (50 boys and 10 girls) were administered an attitude scale. The items were to be answered in 'Yes' and 'No' form.

Their frequencies towards 10 items are presented in the table below. The question is: Do the opinions of boys and girls differ significantly?

Chi-square with Yates' Correction in a 2×2 Table

	Yes	No	
Boys	20 A	30 B	50
Girls	3 C	7 D	10
	23	37	60

According to Equation (5.3):

$$\chi^2 = 60\frac{\left[|(20)(7)-(30)(3)|-\frac{60}{2}\right]^2}{(50)(10)(23)(37)}$$

$$= \frac{60|140-90|-30]^2}{425500} = \frac{60\times 400}{425500} = \frac{24000}{425500} = 0.056$$

In the above example, the expected frequency (23×10/60) is less than 5. Hence, chi-square has been calculated by Equation (5.3). Entering the table for chi-square, we find that for d.f. = 1, the value of chi-square at the 0.05 level should be 3.841. Since the obtained value is less than it ($P >> 0.05$), one concludes that the opinions of boys and girls do not differ significantly.

5.3.2 Mann-Whitney *U* Test

The Mann-Whitney *U* test is a non-parametric substitute for the parametric *t* test. This test was independently proposed by Mann and Whitney. The Mann-Whitney *U* test is used when the researcher is interested in testing the significance of difference between two independently drawn samples or groups. For applications of the *U* test it is essential that the data be obtained on ordinal measurement, that is, they must have been obtained in terms of rank. Where the data have been obtained in terms of scores, for application of the Mann-Whitney *U* test, it is essential that those scores be converted into rank without much loss of information. It is not necessary for the application of the Mann-Whitney *U* test that both groups must have unequal size. However, this test can also be applied to groups having equal size.

Here the calculation of the Mann-Whitney *U* test, which are concerned with larger sample size of more than 20 cases is given:

$$U = N_1N_2 + \frac{N_1(N_1+1)}{2} - \Sigma R_1 \qquad ...(5.4)$$

$$U = N_1N_2 + \frac{N_2(N_2+1)}{2} - \Sigma R_2 \qquad ...(5.5)$$

Table 5.4 presents the scores of two groups on the Lie scale. Group I has 10 subjects and Group II has 21 subjects. The first step is to rank all the scores in one combined distribution in an increasing order of size. The lowest score (taking both sets of scores together) is 7 (second column) and hence, it is given a rank of 1. The nest score is 8, which is again in the second column and it has been given a rank of 2. The third score from below is 10 (in the first column), which has been given a rank of 3. In this way, ranking is continued until all scores receive ranks. Subsequently, the two columns of ranks are summed. At this point, a check on arithmetical calculation is imposed. The check is that the sums of these two columns must be equal to $N(N+1)/2$.

Check: $\Sigma R_1 + \Sigma R_2 = \dfrac{(N)(N+1)}{2} = 88.5 + 407.5 = 496;\ \dfrac{(31)(32)}{2} = 496$

NOTES

Hence, we can proceed:

(by Equation 5.4)

$$U = N_1 N_2 + \frac{N_1(N_1+1)}{2} - \Sigma R_1 = (10)(21) + \frac{(10)(10+1)}{2} - 88.5 = 176.5$$

(by Equation 5.5)

$$U = N_1 N_2 + \frac{N_2(N_2+1)}{2} - \Sigma R_2 = (10)(21) + \frac{(21)(21+1)}{2} - 407.5 = 33.5$$

Table 5.4 *Calculation of the Mann-Whitney U Test from Larger Sample Sizes*

Gr. I **($N_1 = 10$)**	**Gr. II** **($N_2 = 21$)**	R_1	R_2
18	32	7	13
14	40	15	18
30	31	11	12
10	39	3	16.5
39	15	16.5	6
26	8	9	2
27	47	10	19
19	33	8	14
35	52	15	22
11	48	4	20
	7		1
	50		21
	61		27
	58		24
	53		23
	59		25
	60		26
	65		30
	63		28
	67		31
	64		29
		$\Sigma R_1 = 88.5$	$\Sigma R_2 = 407.5$

It is the lower value of U test that one wants. For testing the significance of the obtained U, its values are converted into Z score as shown:

$$Z = \frac{U - \dfrac{N_1 N_2}{2}}{\sqrt{\dfrac{(N_1)(N_2)(N_1 + N_2 + 1)}{12}}} \qquad ...(5.6)$$

$$= \frac{176.5 - \frac{(10)(21)}{2}}{\sqrt{\frac{(10)(21)(10+21+1)}{12}}} = \frac{71.5}{23.664} = 3.02$$

NOTES

A *Z* score from +1.96 to +2.58 is taken to be significant at the 0.05 level of significance and if the *Z* score is greater than even +2.58, one takes it to be significant at the 0.01 level. Since the obtained *Z* is 3.02, on can take the value of the Mann-Whitney *U* to be a significant one. Rejecting the null hypothesis, one can conclude that the two groups differ significantly on the measures of the Lie scale.

5.3.3 Rank-difference Methods

The methods of correlation based upon rank differences are very common among behavioural scientists. These are two most common methods which are based upon the differences in ranks assigned on the *X* and *Y* variables. One is the Spearman rank-difference method and the other is the Kendall rank-difference method. The Spearman rank-difference method symbolized by ρ (read as rho) is a very popular method of computing the correlation coefficient between two sets of ranks or between two sets of scores converted into ranks. The method has been named after Spearman who discovered it. This method is applicable when the number of pairs of scores or ranks is preferably small, that is, 30 or below.

The equation is:

$$\rho = 1 - \frac{6\sum D^2}{N(N^2-1)} \qquad ...(5.7)$$

where ρ = Spearman's rank-difference correlation coefficient; *D* = difference between rank$_1$ and rank$_2$; and *N* = Number of pairs of ranks or scores.

To illustrate the calculation of ρ consider the data given in Table 5.5 that follows, which show the scores of 12 students on the intelligence test (*X*) as well as on the educational test (*Y*). The first step is to rank both sets of scores separately giving the highest score a rank of 1, the next highest score a rank of 2, and so on. Then, keeping the algebraic signs in view, the difference between two sets of ranks is computed. This is noted under column *D*. Subsequently, each difference is squared and noted under column D_2. Substituting the values in the equation, we get a ρ of –0.185. Following the table of Guilford (1956: Table L, p. 549), we can test the significance of the obtained ρ. Since the obtained value of ρ is less than the value given at the 0.05 level (ρ > 0.05) for *N* = 12, one can accept the null hypothesis and can conclude that *X* and *Y* are independent and whatever correlation has been found is due to the chance factor.

NOTES

***Table 5.5** Illustration of the Spearman Rank-Difference Correlation*

X	Y	$Rank_1$	$Rank_2$	$D\ (R_1 - R_2)$	D^2
47	68	8.5	1	+7.5	56.25
50	60	5.5	2.5	+3	9.00
70	54	2	7	–5	25.00
72	53	1	8	–7	49.00
46	60	10	2.5	–7.5	56.25
50	55	5.5	6	–0.5	0.25
42	48	11	9	+2	4.00
58	30	3	12	–9	81.00
55	45	4	10	–6	36.00
36	43	12	11	+1	1.00
49	59	7	4	+3	9.00
47	56	8.5	5	+3.5	12.25
				$\Sigma D = 0.0$	$\Sigma D^2 = 339.00$

$$\rho = 1 - \frac{6(339)}{12(122-1)} = 1 - \frac{2034}{1716} = 1 - 1.185 = -0.185$$

Another method of computing the rank-difference correlation has been developed by Kendall. The method is known as Kendall's τ for which the formula is as follows.

$$\tau = \frac{S}{(1/2)N(N-1)} \qquad \text{...(5.8)}$$

where τ = Kendall's τ, S = actual total; and N = number of objects or scores which have been ranked.

Suppose 12 students have been administered two tests and their scores are presented in the following table.

Scores of 12 students on X and Y Test

	A	B	C	D	E	F	G	H	I	J	K	L
X	20	26	17	16	15	23	22	24	19	28	30	10
Y	70	80	40	45	38	49	77	76	72	47	36	35

The first step is to rank both sets of scores giving the highest score a rank of 1, the next higher a rank of 2, and so on. The following table presents the ranks based upon two sets of the scores given in that table above. Subsequently, the ranks of the X test are rearranged in a way that they appear in a natural order like 1, 2, 3.

Ranks based upon two sets of scores given in the above table

	A	B	C	D	E	F	G	H	I	J	K	L
X	7	3	9	10	11	5	6	4	8	2	1	12
Y	5	1	9	8	10	6	2	3	4	7	11	12

NOTES

Accordingly, ranks on the *Y* test are adjusted. The following table presents the ranks in a rearranged order. Subsequently, the value of *S* is computed. For this, we start with the rank on the *Y* test from the left side.

Rearranged order of ranks

	A	B	C	D	E	F	G	H	I	J	K	L
X	1	2	3	4	5	6	7	8	9	10	11	12
Y	11	7	1	3	6	2	5	4	9	8	10	12

The first rank on the left side is 11. Count the number of ranks which are above 11 and the number of ranks which are below 11, separately. Only one rank (that is, 12) falling at the right of the first rank on the *Y* test is above 11 and the remaining 10 ranks fall below it. Hence, its contribution to *S* would be equal to 1 – 10. Likewise, the second rank on the *Y* test is 7. The four ranks falling right of 7, are above 7 and 6 ranks are below it. Hence, its contribution to *S* would be 4 – 6.

Identical procedures are repeated for other ranks on the *Y* test. Thus:

$S = (1 - 10) + (4 - 6\ ) + (7 - 1) + (4 - 3\ ) + (6 - 0) + (4 - 1) + (4 - 0) + (2 - 1) + (2 - 0) + (1 - 0) = (-9) + (-2) + (9) + (6) + (1) + (6) + (3) + (4) + (1) + (2) + (1) = 33 - 11 = 22.$

Substituting this in the formula given in Equation (5.8):

$$\tau = \frac{S}{\frac{1}{2}N(N-1)} = \frac{22}{\frac{1}{2}12(12-1)} = \frac{22}{66} = 0.333$$

The significance of τ is tested by converting it into a *Z* score, the formula for which is as follows:

$$Z = \frac{\tau}{\sqrt{\frac{2(2N+5)}{9N(N-1)}}} \qquad ...(5.9)$$

Hence

$$Z = \frac{0.33}{\sqrt{\frac{2[(2)(12)+5)]}{9(12)(12-1)}}} = \frac{0.33}{\sqrt{0.0488}} = \frac{0.3}{0.2209} = 1.4938$$

Since the obtained *Z* score is less than 1.96, one can say that this is not significant even at the 0.05 level. Accepting the null hypothesis, one can say that the given set of scores is not correlated. According to Siegel (1956), τ has one advantage over ρ, and that is that the former can be generalized to partial correlation. If both τ and ρ are computed from the same data, the answer will not be the same and hence, numerically, they are not equal.

NOTES

5.3.4 Coefficient Concordance, *W*

The coefficient of concordance symbolized by the letter *W* has been developed by Kendall and is a measure of correlation between more than two sets of ranks. Thus, *W* is a measure of correlation between more than two sets of rankings of events, objects and individuals. When the investigator is interested in knowing the inter-test reliability, *W* is chosen as the most appropriate statistic. One characteristic of *W* which distinguishes it from other methods of correlation is that it is either zero or positive. It cannot be negative. *W* can be computed with the help of the formula given below:

$$W = \frac{S}{\frac{1}{12}K^2(N^3 - N)} \qquad ...(5.10)$$

where W = coefficient of concordance; S = sum of squares of deviations from the mean of R_j; K = number of judges or sets of rankings; and N = number of objects or individuals which have been ranked. Suppose four teachers (A, B, C and D) ranked 8 students on the basis of performance shown in the classroom. The ranks given by the four teachers are presented in the following table. The details of the calculations have also been shown.

Ranks given by four teachers to eight students on the basis of classroom performance

Teachers	Students							
	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)
A	3	4	7	5	8	6	2	1
B	2	3	6	4	8	7	1	5
C	1	3	5	6	8	7	2	4
D	3	4	2	5	7	6	1	8
R_j	9	14	20	20	31	26	6	18

$$\text{Mean of } R_j = \frac{9+14+20+20+31+26+6+18}{8} = \frac{144}{8} = 18$$

$$S = (9-18)^2 + (14-18)^2 + (120-18)^2 + (20-18)^2 + (31-18)^2$$
$$+ (26-18)^2 + (6-18)^2 + (18-18)^2$$
$$= (-9)^2 + (-4)^2 + (2)^2 + (2)^2 + (13)^2 + (8)^2 + (-12)^2 + (0)^2 = 482$$

Now substituting in Equation (8.21):

$$W = \frac{482}{\frac{1}{12}(4)^2(8^3-8)} = \frac{482}{672} = 0.717 = 0.72$$

when $N > 7$, the significance of *W* is tested by converting its value into χ^2 with the help of the following equation: $\chi^2 = K(N-1)W$. Thus, $X^2 = 4(8-1) \times 0.72 = 20.16$ and d.f. in this situation is always equal to $N - 1$. Hence d.f. $= 8 - 1 = 7$.

From the probability table for chi-square we find that the value of chi-square for d.f. = 7 at 0.05 level of significance should be 18.475. Since the obtained value of the chi-square exceeds this required value, one can take this value of *W* as a significant one. Thus rejecting the null hypothesis, one can say that there is an overall significant relationship in ranking done by the four teachers.

5.3.5 Median Test

The median test is used to see if two groups (not necessarily of same size) come from the same population or from populations having the same median. In the median test, the null hypothesis is that there is no difference between the two sets of scores because they have been taken from the same population. If the null hypothesis is true, half of the scores in both the groups should lie above the median and the remaining half of the scores should lie below the median. The following table presents the scores of two groups of students in an arithmetic test. The first step in the computation of a median test is to compute a common median for both distributions taken together.

Scores of 30 Students on an Arithmetic Test

Gr. B (*N* = 16)	16, 17, 8, 12, 14, 9, 7, 5, 20, 22, 4, 26, 27, 5, 10, 19
Gr. B (*N* = 14)	28, 30, 33, 40, 45, 47, 40, 38, 42, 50, 20, 18, 18, 19

For computing the common median, both the distributions are pulled together as shown in the following table:

Scores	*f*
49–53	1
44–48	2
39–43	3
34–38	1
29–23	2
24–28	3
19–23	5
14–18	5
9–13	3
4–8	5
	N = 30

$$\text{Median} = 1 + \frac{(N/2 - F)i}{f_m}$$

$$= 18.5 + \frac{(30/2 - 13)5}{5}$$

$$= 20.5$$

Subsequently, a 2×2 contingency table is built as follows:

Now, the chi-square test can be applied. For computing chi-square from a 2×2 table, we may follow the equation for Chi-square (χ^2). Yates' correction is not needed here because none of the cells contain an expected frequency less than 5.

	Above Mdn	Not above Mdn	
Gr. A	A 3	B 13	16
	C 10	D 4	14
	13	17	30

Now substituting the values in Equation (5.2) we get:

$$\chi^2 = \frac{30[|3(4)-(13)(10)|]^2}{(16)(14)(13)(17)}$$

$$= \frac{30[|12-130|]^2}{49504} = \frac{417720}{49504} = 8.438 = 8.44$$

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

From the probability table for chi-square, we find that for d.f. = 1 the chi-square value at the 0.01 level should be 6.635. Since the obtained value of the chi-square exceeds this value ($p < 0.01$), we can reject the null hypothesis and conclude that the two samples have not been drawn from the same population or from populations having equal medians.

5.3.6 Kruskal-Wallis *H* Test

The primary difference between the *F* test and the Kruskal-Wallis *H* test on the one hand and the Friedman test on the other hand is that the *F* test is a parametric analysis of variance, whereas the *H* test and Friedman test are the non-parametric analysis. The *H* test is a one-way non-parametric analysis of variance and the Friedman test is a two-way non-parametric analysis of variance.

Table 5.6 *H Test from Scores Obtained by Three Groups on Lie Scale*

Gr.A	**(N = 6)**	**Gr. B**	**(N = 8)**	**Gr. C**	**(N = 10)**
15	(14)	17	(15)	20	(17)
10	(9)	9	(8)	25	(19)
8	(3)	8	(6)	13	(12)
5	(4)	14	(13)	11	(10)
6	2)	2	(1)	26	(20)
4		8	(6)	24	(18)
		12	(11)	36	(24)
		18	(16)	30	(23)
				29	(22)
				27	(21)
R_j = 38		R_j = 76			R_j = 186

The *H* test is used when the investigator is interested in knowing whether or not groups of the independent samples have been drawn from the same population.

If the obtained data does not fulfill the two basic parametric assumptions, namely, the assumptions of normality and the assumption of homogeneity of variances, the *H* test is the most appropriate statistic. The equation for the *H* test is as given as follows:

$$H = \frac{12}{N(N+1)}\left[\sum \frac{R_j^2}{N_j}\right] - 3(N+1) \qquad \text{...(5.11)}$$

Where N = number in all samples combined; R_j = sum of ranks in j sample; and N_j = number in j sample.

Data to illustrate the calculation of the *H* test have been given in Table 8.6. Three groups of students were administered a Lie Scale and their scores are presented in Table 5.6. The first step is to combine all the scores from all of the groups and rank them with the lowest score receiving a rank of 1 and the largest score by rank *N*. Ties are treated in usual fashion in ranking subsequently, the sum of ranks in each group or column is found and the *H* test determines where these sums of ranks are so disparate that the three groups cannot be regarded as being drawn from the same population. The ranks assigned to each score earned by the member of the group are given in brackets. Now, substituting the value in the equation given in 5.11. we get:

$$H = \frac{12}{(24)(25)}\left[\sum \frac{(38)^2}{6} + \frac{(76)^2}{8} + \frac{(186)^2}{10}\right] - 3(24+1)$$

$$= \frac{53067.192}{600} - 75 = 88.445 - 75 = 13.445 = 13.44$$

When each sample has six or more than six cases, the *H* test is interpreted as chi-square. In such a situation, d.f. = number of groups or samples minus one. So, here d.f. = 2. Entering the probability table for chi-square, we find that for d.f. = 2, the value of chi-square at the level of significance should be 9.210. Since the obtained value of *H* test exceeds this required value, it can be said that the *H* value is significant. Rejecting the null hypothesis, one concludes that the samples are independent and that they have not been drawn from the same population.

5.3.7 Friedman Test

The Friedman test is a two-way non-parametric analysis of variance. When the groups are matched, doubts exist about the two basic parametric assumptions, namely, the assumption of normality and the assumption of homogeneity of variances, the investigator resorts to the Friedman test for testing whether or not the samples have been drawn from the same population. To illustrate this, the calculations of the Friedman test have been presented in Table 5.7. The following equation calculates the Friedman test is as follows:

$$X_r^2 = \frac{12}{NK(K+1)}(R_j)^2 - 3N(K+1) \qquad \text{...(5.12)}$$

where X_r^2 = Friedman test; N = number of rows; K = number of columns; and R_j = separate sums of ranks of each column.

Table 5.7 *Friedman Test of Two-way ANOVA from Scores of Three Groups on Recall Test*

	I	II	III	IV	V
Gr. A	12(4)	8(2)	4(1)	10(3)	16(5)
Gr. B	10(3)	7(2)	6(1)	11(4)	17(5)
Gr. C	7(2)	83)	4(1)	12(5)	10(4)
R_j	9	7	3	12	14

The first step in calculation of the Friedman test is to rank each score in each row separately giving the lowest score in each row a rank of 1 and the next lowest score in each row a rank of 2, and so on. The ranking can also be done in reversed order, that is, giving the highest score in each row a rank of 1, the next highest score in each row a rank of 2, and so on. Rank assigned to each score in each row is given in parentheses. The Friedman test is applied to determine whether or not the rank totals symbolized by R_j differs significantly. Now, substituting the values in the equation of chi-square, we get

$$X_r^2 = \frac{12}{(3)(5)(5+1)}\left[(9)^2+(7)^2+(3)^2+(14)^2\right]-3(3)(5+1)$$

$$= 63.866 - 54 = 9.87$$

When the number or rows (N) and the number of columns (K) are too small, the significance of the Friedman test can be ascertained with the help of special tables (Siegel, 1956). For example, when $K = 4$, $N = 2$ to 4 or when $K = 3$, $N = 2$ to 9, the significance of the Friedman test can be done through these special tables. But when the number of rows and the number of columns are greater than those said above, the Friedman test is interpreted as the chi-square test. In the present example, the significance of the Friedman test would be interpreted in terms of chi-square. The d.f. is always equal to $K - 1$ for chi-square applied as a test of significance of the Friedman test. Hence d.f. in the present example would be $K - 1 = 5 - 1 = 4$. Entering the table for d.f. 4, we find that the chi-square should be 9.488 at the 0.05 level of significance. Since the obtained value of the Friedman test exceeds this value ($p < 0.05$), one rejects the null hypothesis and conclude that three matched groups differ significantly.

CHECK YOUR PROGRESS

3. Name any three important non-parametric tests.
4. When is the Mann-Whitney U test used?
5. What is the Friedman Test?

5.4 SUMMARY

- ANalysis Of VAriance or ANOVA is one of the most powerful techniques in statistical analysis and was developed by R.A. Fisher. It is also called the F-Test. There are two types of classifications involved in the analysis of variance. The one-way analysis of variance refers to the situations when only one fact or variable is considered. In the second type of classification, the response variable of interest may be affected by more than one factor.
- The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.
- In case of a two-way design with repeated measurements for all the categories, we can obtain a separate independent measure of inherent or smallest variations. For this measure, we can calculate the sum of squares and degrees of freedom in the same way as we have worked out the sum of squares for variance within samples.
- In the case of one-way ANOVA, SS total, SS between columns and SS between rows can also be worked out as stated above. We then find left-over sums of squares and left-over degrees of freedom which are used for what is known as 'interaction variation'. Interaction is the measure of inter-relationship among the two different classifications. After making all these computations, ANOVA table can be set up for drawing inferences.
- The important non-parametric statistics are as follows: (a) Chi-square (χ^2) Test (b) Mann-Whitney *U* test (c) Rank-difference methods (both ρ and T) (d) Coefficient of concordance (*W*) (e) Median Test (f) Kruskal-Wallis *H* Test (g) Friedman test.
- The chi-square is one of the most important non-parametric statistics, which is used for several purposes. For this reason, Guilford (1956) has called it the general-purpose statistic. It is a non-parametric statistic because it involves no assumption regarding the normalcy of distribution or homogeneity of the variances. The chi-square test is used when the data are expressed in terms of frequencies of proportions or percentages.
- The Mann-Whitney *U* test is a non-parametric substitute for the parametric *t* test. This test was independently proposed by Mann and Whitney. The Mann-Whitney *U* test is used when the researcher is interested in testing the significance of difference between two independently drawn samples or groups.
- The methods of correlation based upon rank differences are very common among behavioural scientists. These are two most common methods which are based upon the differences in ranks assigned on the *X* and *Y* variables. One is the Spearman rank-difference method and the other is the Kendall rank-difference method.

- The coefficient of concordance symbolized by the letter *W* has been developed by Kendall and is a measure of correlation between more than two sets of ranks. Thus, *W* is a measure of correlation between more than two sets of rankings of events, objects and individuals. When the investigator is interested in knowing the inter-test reliability, *W* is chosen as the most appropriate statistic.
- The median test is used to see if two groups (not necessarily of same size) come from the same population or from populations having the same median. In the median test, the null hypothesis is that there is no difference between the two sets of scores because they have been taken from the same population.
- The primary difference between the *F* test and the Kruskal-Wallis *H* test on the one hand and the Friedman test on the other hand is that the *F* test is a parametric analysis of variance, whereas the *H* test and Friedman test are the non-parametric analysis. The *H* test is a one-way non-parametric analysis of variance and the Friedman test is a two-way non-parametric analysis of variance.
- The Friedman test is a two-way non-parametric analysis of variance. When the groups are matched, doubts exist about the two basic parametric assumptions, namely, the assumption of normality and the assumption of homogeneity of variances, the investigator resorts to the Friedman test for testing whether or not the samples have been drawn from the same population.

5.5 KEY TERMS

- **ANOVA:** Its full form is analysis of variance and it tests for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.
- **One-way ANOVA:** It is a technique used to compare means of two or more samples using the *F* distribution and can be used only for numerical data.
- **Chi-square (χ^2) Test:** It is a non-parametric statistic because it involves no assumption regarding the normalcy of distribution or homogeneity of the variances.
- **Mann-Whitney *U* Test:** It is used when the researcher is interested in testing the significance of difference between two independently drawn samples or groups.

5.6 ANSWERS TO 'CHECK YOUR PROGRESS'

1. The basic principle of ANOVA is to test for differences among the means of the populations by examining the amount of variation within each of these samples, relative to the amount of variation between the samples.
2. Two basic assumptions underlying the technique of analysing variance are as follows:
   - The population distribution should be normal. This assumption is, however, not so important. The study of Norton (Guilford, 1965) also points out that 'F' is rather insensitive to variations in the shape of population distribution.
   - The sub-groups under investigation must have the same variability. In other words, there should be homogeneity of variance.
3. The three important non-parametric tests are as follows:
   - Chi-square (χ^2) Test
   - Mann-Whitney *U* test
   - Rank-difference methods (both ρ and T)
4. The Mann-Whitney U test is used when the researcher is interested in testing the significance of the difference between two independently drawn samples or groups.
5. The Friedman test is a two-way non-parametric analysis of variance. When the groups are matched, doubts exist about the two basic parametric assumptions, namely, the assumption of normality and the assumption of homogeneity of variances, the investigator resorts to the Friedman test for testing whether or not the samples have been drawn from the same population.

5.7 QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the two types of classification involved in the analysis of variance?
2. Why is the two-way ANOVA technique used?
3. Discuss one-way ANOVA in detail with examples.
4. State the uses of Chi-square (χ^2) Test.
5. List the importance of rank-difference methods.

Long-Answer Questions

1. Explain the test to determine the differences within the factor under the one-way ANOVA.

NOTES

2. Explain the two-way ANOVA technique in the context of repeated and non-repeated value designs.
3. A manufacturing company has purchased three new machines of different makes and wishes to determine whether one of them is faster than the others in producing a certain output. Five hourly production figures are observed at random from each machine and the results are given below:

Observations	A_1	A_2	A_3
1	25	31	24
2	30	39	30
3	36	38	28
4	38	42	25
5	31	35	28

Use ANOVA and determine whether the machines are significantly different in their mean speed. (Given at 5% level, $F_{2, 12} = 3.89$)

4. What is Coefficient Concordance, *W*? How is it different from other tests?
5. Describe the Friedman Test with suitable examples.

5.8 FURTHER READING

Mood, Alexander M., Franklin A. Graybill and Duane C. Boes. 1974. *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Trivedi, K.S. 1994. *Probability and Statistics with Reliability, Queuing and Computer Science Applications*. New Delhi: Prentice-Hall of India.

Mendenhall, William, Robert J. Beaver and Barbara M. Beaver. 2005. *Introduction to Probability and Statistics*, 12th edition. California: Duxbury Press.

Gupta, S.C. 2005. *Fundamentals of Statistics*, 17th edition. Mumbai: Himalaya Publishing House.