Unit I: Introduction

Basic Concept of Statistics:

The word "Statistics" has been derived from the Latin word "**Status**", French word **"Statistique"**, Italian word **"Statista"** and German word **"Statistik"** each of which means 'political state. States used Statistics only to keep information for government purposes regarding the population, 'poverty or wealth' of the country, the number of polices, soldiers, fiscal policies etc. Statistics was regarded as the science of statecraft. So it was used in collection of the information.

The word "Staistics" is used in singular as well as plural sense.

(i) Quantitative information of facts or simply data.

(ii) Statistical methods for handling numerical data.

The first is used in the plural sense. Statistics in the plural sense, it means the quantitative information or the numerical set of data or numerical facts collected systematically. The second is used in the singular sense. In the singular sense, it means the statistical methods and techniques adopted for collection, presentation, analysis and the interpretation of the numerical data.

Definition of Statistics

There is not a single definition of Statistics since the field of applicability has been widening day by day. As the field of applicability is extended, the definition of Statistics needs to be modified. Hence, it has been defined in a wider domain. Some comprehensive definitions are as follows.

Definition of Statistics in Singular sense

In the singular sense, it is also known as 'statistical method' for the collection, presentation, analysis and interpretation of numerical data.

The most comprehensive definition given by Croxton and Cowden is:

"Statistics may be defined as the science of the collection, presentation, analysis and interpretation of numerical data."

Definition of Statistics in Plural sense

In the plural sense, of all definitions Prof. Horace Secrist's definition is the most comprehensive. According to him

"Statistics are aggregates of facts affected to a marked extent by multiplicity of causes, numerically expressed, enumerated or estimated according to reasonable standard of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

Statistics may also be classified into two parts, which are as follows.

Theoretical Statistics or Mathematical Statistics
Applied Statistics

Theoretical Statistics or mathematical statistics

Theoretical Statistics can further be subdivided into two parts.

- Descriptive Statistics
- Inferential Statistics

Descriptive Statistics

Descriptive Statistics merely describes the data and consists of methods and techniques used in collection, organization, presentation of data using table and chats, diagram, graph etc., summarizing data using measure of central tendency, dispersion, skewness, kurtosis etc. and analysis of data in order to describe

various features and characteristics of such data is called descriptive statistics. Summarized results obtained from descriptive statistics which describe the data but cannot be used to generalize. Average, rates, ratios, percentages etc. are the basic summary statistics or descriptive statistics for summarizing the data depending on their nature. Hence, some most frequently used descriptive statistical measures are measures of central tendency, measures of dispersion, measures of skewness, measures of kurtosis etc.

Inferential Statistics

Inferential Statistics deals with the methods of drawing (or inferring) conclusions about the characteristics of the population based upon the results of the sample taken from the population.

In other words, Statistics which deals with sample selection from population and statistical technique used to draw conclusion about population on the basis of statistical measures obtained from sample is called inferential statistics.

It is used in estimation of parameters and statistical testing of hypothesis.

Applied Statistics

This consists of massive application of theoretical or mathematical Statistics in the different areas such as biology, physics, astronomy, metrology, chemistry, medical science, sociology, psychology, business, economics , agriculture, Econometrics, bio-statistics or biometry, etc. The statistical tools and methods are used in order to solve many more practical problems in diversified area. Besides that applied statistics has been used in decision making problems.

Functions of Statistics

Statistics has been playing a vital role almost every area in the universe. Some of important functions are listed below.

- **To help classification of data:** When data are collected, they are not comprehensive initially. To make these data comprehensive, it is useful to classify the data utilizing statistical principles. Hence, statistics helps in classifying data according to the nature of data.
- **Statistics simplifies complexity: S**tatistics consists of aggregate of numerical facts. Huge facts and figures are difficult to remember. The complex mass of figures can be made simple and understandable with the help of various statistical techniques. Statistical methods such averages, dispersion, graphs, diagram etc. make huge mass of figures easily understandable. So, the function of statistics reduces the complexity of the huge mass of figures to a simpler form.
- **Statistics facilities comparison:** The science of statistics does not mean only counting but also comparison. Statistics contributes different statistical procedure and methods such as average, ratios, percentages, rates, coefficients etc. for comparing the relevant values with each other. Just for example, height (in cm.) and weight (in lbs.) of children can be compared by converting them into relative statistical measures.
- **Statistics present facts in a definite form:** One of the important functions of statistics is to present the general statements in a precise and definite form. The conclusion stated numerically is definite and hence more convincing than the conclusions stated qualitatively.
- **To determine relationship between different phenomena:** Statistical techniques such as correlation and regression etc. are useful to measure the relationship between the variables. One can easily observe the correlation between income and expenditure of a family, ages and playing habits of students etc.
- **To help in formulation of policies:** Statistics helps in formulating the policies in different fields such as economics, business wage and income, tax and export, development, and expenditure etc. are based on the statistical inference drawn from these areas. The government policies are also framed on the basis of statistics. In fact, without statistics, suitable policies can not be framed.

- **Statistics helps in predicting or forecasting future trends:** As statistical methods are highly useful tools in analyzing the past data and predicting some future trends. So, while preparing suitable policies and plans, it is necessary to have knowledge of use of statistical techniques like time series analysis, regression analysis etc. On the basis of past or present values of variables, other values of variables for future can be predicted with the help of statistical technique like time series analysis, regression analysis. Hence, Statistics helps for forecasting future trends.
- **Statistics helps in formulating and testing hypothesis:** Statistical methods are helpful not only in estimating the present and forecasting future but also helpful in formulating and testing the hypothesis for the development of new theories. As Statistics is the science which is most useful and applicable to formulate and testing of hypothesis for the new theories and also for verification of existing one by calculating statistical measures from the sample data.

- **To draw valid inferences or conclusions:** One of the indispensable sciences for valid conclusions about any phenomena based on data is the statistics. Generally, it is very hard to enumerate each and every unit of universe because of different reasons. Therefore, certain units are selected from the universe and they are studied. Applying different statistical tools on these sample data, statisticians can draw inferences about the whole universe. Thus statistics helps to draw valid conclusions.

Importance and Scope of Statistics

The sufficient quantitative information regarding the various business activities mentioned above can be collected with the help of statistical methods. This information can be of great help in formulating suitable polices that manager can extract suitable information from the collected data and used in making decisions.

In recent years, the development in statistical studies has considerably increased its scope and importance. So, there is hardly any walk of life, which has not been affected by statistics. It has become one of the indispensable parts in almost all areas. It is used in the field of physical sciences, Biological sciences, Medical Sciences, Industry, Economics sciences, Social Sciences, Management Sciences, Information Technology, Agriculture, Insurance, Business, management, accounting, finance, marketing, production, computer, information technology, social sciences, Engineering and many other areas. Attempts have been made to discuss some of the uses of Statistics in some areas which are discussed below.

Statistics in Physical Sciences

The immense use of statistics has been found in physical sciences. In physical sciences, a large number of instruments are used for measurements purpose on the same item and scientists are encountered with the variations in these measurements. So, in order to have an idea about the degree of accuracy, the statistical techniques such as estimation theory i.e. point estimation, interval estimation-confidence intervals and confidence limits are used to assign certain limits within which true value of the phenomenon may be expected to lie. The desire for precision was first felt in physical sciences and this led the science to express the facts under study in quantitative form. The statistical theory with the powerful tools of sampling, estimation, testing of hypothesis, design of experiments etc. are most effective for the analysis of quantitative expression of all fields of study.

There is an increasing use of statistics in most of the physical sciences such as astronomy, geology, engineering, physics, metrology etc. For example, In astronomy, the theory of Gaussian 'Normal Law of Errors' for the study of the movement of stars and planets is developed by using the principle of least squares. Gauss also used the normal curve to describe the theory of accidental errors of measurement for calculating the orbits of heavenly bodies.

Statistics in Biological Sciences

Statistics has been using importantly in biological sciences for analyzing the biological data in different prospective of the subject matter of interest. Statistical methods used in exploration of biological phenomenon are called Bio–statistics. Biostatistics is the application of statistics to problems in the biological sciences, health, and medicine. Sir Francis Galton (1822-1911), a British Biometrician developed the association between statistical methods and biological theories in his work on 'Regression' in the connection with the inheritance of stature. According to Prof. Karl Pearson' (1857-1936), who pioneered the study of 'correlation Analysis' the whole theory of heredity rests on statistical basis.

Medical Sciences

In medical sciences also, there is huge application of statistics. In medical sciences, the statistical tools for collection, presentation and analysis of observed factual data relating to the causes and incidence of diseases are of paramount importance. The most important application of statistics in medical sciences lies in test of significance (more precisely t-test) for testing the efficacy of a manufacturing drug, injection or medicine for controlling or curing specific ailments. Comparative studies for the effectiveness of different medicines by different concerns also be made with statistical techniques (t & F tests of significance. Bio statistics (sometimes known as biometrics), most generally, is the application of statistics to biology and most commonly to medicine. Hence, in medical science, Statistics is used

- To compare the efficacy of a particular drug to the patients
- To compare the action of two different drugs.
- To compare variability between patients, association between variables such as age and sex, height and weight, age and blood pressure, and to find out the association between two attributes such as cancer and smoking
- To identify signs and symptoms of a disease
- To find and establish the normal range and averages of variables such as blood sugar, pulse rate, BP, cholesterol level etc.
- To help in providing methods for the definition of normal or healthy and indicating at what point the measure some bodily characteristics becomes abnormal.

Statistics in Industry

Statistics plays a crucial role in industry. In industry, Statistics is extensively used in 'Quality control'. In production engineering to find out whether the product is confirming to the specifications or not. Statistical tools, such as inspection plan, control chart etc are used. The main objective in any production process is to control the quality of the product so that it conforms to specifications. This is called process control and is obtained through powerful technique of control charts and inspection plans which is based on setting of 3σ control limits which has its basis on the theory of probability and normal distribution and inspection plans are based on the special kind of sampling theory which are very important aspect of statistical theory.

Statistics in Economics Sciences

Statistics has widely used in economics. Statistical data and advanced techniques of statistical analysis immensely useful in the solution of a variety of economic problems such as production, consumption, distribution of income and wealth, wages, prices profits, savings, investment, unemployment, poverty etc

Time series analysis, index numbers, forecasting techniques, demand analysis are some of the very powerful statistical tools which are used immensely in the analysis of economic data and also for economic planning. For example, time series analysis is extremely used in economic statistics for the study of the series relating to prices, production, and consumption of commodities, money in circulation, bank deposits and bank clearing, sales in department store etc.

So, statistics has widely used in

i. the proper understanding of the economic problems
ii. Formulation of economic policies
iii. the evaluation of the effectiveness of economic policies.

iv. the development of the economic theory.

The increasing interaction of mathematics and statistics with economics led to the development of new discipline called Econometrics.

Statistics in Social Sciences

According to Bowley, "Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations." Statistics has employed in political science, sociology, anthropology, population studies etc. Every social phenomenon is affected to a marked extent by a multiplicity of factors which bring out the variation in observations from to time, place to place and objective to objective. Statistical tools such as regression and correlation analysis are most frequently used in the study of social science research. Sampling techniques and estimation theory are very powerful and indispensible tools for conducting any social surveys. Statistics has been used mostly in the field of Demography which is one of the important fields of social science for studying fertility, morality, population growth, marriages, migration etc.

Statistics in Business and Management Sciences

Statistics influence the operations of business and management in many dimensions. Statistical applications include the area of production, marketing, promotion of product, sales, purchase, financing, banking, distribution, accounting, marketing research, manpower planning, forecasting, research and development etc. 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 judgment. In such situations, Statistics are used to analyze this vast data base for extracting relevant information.

Statistical knowledge is very helpful to the businessman. Statistical techniques are used to formulate different plans and policies, to estimate the market fluctuations, changes in the demand conditions etc. It is also helpful for businessman in forecasting the future trends and tendencies. Hence, for becoming a successful businessman, ideas in statistics are essential.

Some of typical areas of business operations where Statistics have extensively and effectively used are as follows:

Marketing: Statistical analysis are frequently used in providing information for making decision in the field of marketing it is necessary first to find out what can be sold and the to evolve suitable strategy, so that the goods which to the ultimate consumer. A skill full analysis of data on production purchasing power, man power, habits of consumer, transportation cost should be consider to take any attempt to establish a new market.

Production: In the field of production statistical data and methods play a very important role. The production of any item depends upon the demand of that item and the demand must be forecasted using statistical techniques. Similarly, the decisions about what to produce? How to produce? When to produce? How much to produce? are based largely upon the feedback surveys that are analyzed statistically.

Finance: The financial organization discharging their finance function effectively depend very heavily on statistical analysis.

Banking: Banking institutes have found if increasingly to establish sesearch department within their organization for the purpose of gathering and analysis information, not only regarding their own business but also regarding general economic situation and every segment of business in which they may have interest.

Investment: Statistics have been almost indispensible in making a sound investment .so, Statistics greatly assists investors in making clear and valued judgment in his investment decision in selecting securities which are safe and have the best prospects of yielding a good income.

Purchase: The purchasing department of an organization makes decisions regarding purchase of raw materials and other supplies from different venders. Therefore, to frame suitable purchase policies such as what to buy? At what price to buy? What quantity to buy? When to buy? Where to buy? Whom to buy? Statistical data are used.

Control: the management control process combines statistical and accounting methods in making the

overall budget for the coming year including sales, materials, labor and other costs and net profits and capital requirement.

Accounting: A very important application of statistics in accountancy is the ' Method of Inflation Accounting' which consists in evaluating the accounting records based on the historical costs of assets after adjusting for the change in the purchasing power of money and is obtained through the powerful statistical tools of price index number and price deflators. The regression analysis theory is of immense help in' Cost Accounting' in forecasting cost or price for any given value of the dependent variable. With the statistical tools of regression the effect of changes in future prices on the cost of production can be predicted.

Statistics in Information Technology

The availability of computers has resolved the problem of data storage as well as the problem of data processing (or process of converting data into meaningful information). The availability of internet facilities or other electronic mailing devices has eased the communication of data or information. But computer alone is not sufficient to resolve all the problems related to data such as the problems that arise during the time of data collection, data analysis and interpretation of data or information. Thus, there is inter-linkage between Statistics, computer and information technology.

Statistics in Agriculture:

The term 'Analysis of variance' was introduced by Prof. R.A. Fisher in 1920 to deal with the problem in the analysis of agronomical data, which is a powerful statistical tool for tests of significance. Statistical techniques are widely used for executing different agricultural surveys, estimating the production of various livestock products like milk, egg, wool and yield estimation of crops etc. The measurement of variations in the entire agricultural surveys carried out at different time points, analysis of the effect of manure, irrigation etc. in the production and the productivity of crops, different statistical techniques such as minimum variance unbiased estimates techniques, regression analysis , design of experiment techniques etc. are applied.

Statistics in Insurance

The main basis of modern theory of statistics is the probability, and it is the backbone of Insurance. The idea of life insurance developed by during the end of the seventeenth century after the preparation of the life tables by Edmund Hally in 1961. Life tables are indispensable for the solution of all problems concerning the duration of human life. Life tables, which are based on the scientific use of statistical methods i.e. probability and mathematical expectation are the key –stone or the pivot on which the whole science of insurance hinges. Life tables form the basis for determining the rates premiums and annuity necessary to various amounts of life insurance. Insurance can be looked upon as one of the pioneer branches of commerce and Business which has used statistics for very long time Thus, the success of an insurance company largely depends on the accuracy of the statistical data which is used for the construction of life tables.

Statistics in Education and Psychology

Statistics has been used very widely in education and psychology too. Statistics has been most frequently used approaches in education and psychology regarding to in the scaling of mental tests and other psychological tests, comparison of intelligent Quotient (I.Q.) test scores, for measuring the reliability and validity of test scores, in item Analysis and factor analysis. For research in this area, statistical techniques such as Cronbach's Alpha, Mc-Nemer's test etc. have been used. Moreover, advanced statistical analysis such as Factor Analysis and Principal Component Analysis are also used in data analysis for rigorous research of this area. The vast applications of statistical data and statistical theories have given rise to a new discipline called 'Psychometry'.

Statistics in Planning

The modern age as termed as the 'Planning'. Planning is necessary for efficient workmanship and in formulating future policies. Statistics provides the valued interpretation of facts and figures relevant to

planning. Planning depends on the forecasting and Statistics provides the necessary tools of estimation and forecasting .So, Statistics is indispensible in planning.

Statistics in Research

One cannot think of understanding any research activities without using Statistics. Primarily, statistical techniques are used for collecting information in any research. Besides, statistical methods are used for analysis and interpretation of research findings. Thus, there is hardly any advanced research studies where Statistics is not being used. It is used in all spheres of human activities.

Limitations of Statistics

The field of Statistics, though widely used in all areas of human knowledge and widely applied in variety of disciplines such as business, economics, research, physical sciences, biological sciences, medical sciences , social sciences, computer sciences etc. has its own some limitations. Some of these limitations are as follows:

Statistics does not deal with individuals

Statistics deals with an aggregate of facts and does not give any specific recognition to the individual items of a series. So, it deals only with the mass of phenomena or group of individuals and indicates the characteristics of whole group. A single item or the isolated figure can not be regarded as statistics. For example , the marks obtained by a student in Statistics is 70 does not constitute Statistics but the average marks of a group of students in Statistics is 70 forms statistics.

Statistics does not deal qualitative characteristics directly:

Statistical methods can only be applied to numerically expressed data. Statistics is not applicable to qualitative characteristics directly such as honesty, intelligent, integrity, poverty, knowledge, goodness, beauty etc. 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 and then statistical techniques are used. For example intelligence may be compared to some degree by comparing IQs or some other scores in certain intelligence tests.

Statistical laws are not exact

In statistical work 100% accuracy is rare because statistical laws are true only on the average. Generally statistical laws are probabilistic in nature .They are not exact as the laws of Physics, Mathematics, Chemistry. It is also said that Statistics itself is not an absolute measure. It provides precise result minimizing error as much as possible.

Statistics can be misused

Only the experts or statistician can handle statistical data properly. It is likely to be misused the Statistics by non-statistical persons in handling data and interpreting the result. Even expert person or statistician may come to fallacious conclusion if data are biased or inadequate. Hence, the famous statement that "figures don't lie but the liars can figure" is a testimony to the misuse of statistics.

Variable

In Statistics, a variable or characteristic may be defined as an attribute that describe a person, place, thing or idea under study. Therefore, a characteristic or measurement that may vary from one biological entity to another or place to place or time to time is called variable. In other words, a variable is a characteristic that varies from one person or thing to another .Therefore, a quantitative or qualitative characteristic that varies from observation to observation in the same group is called a variable. For example, the variable may be height, weight, age, blood pressure, pulse rate, blood sugar, temperature, gender, income, expenditure , sales, profit, operating cost, ethnicity, religion, occupation, hair colour, eye colour, knowledge, nationality, religion, pain etc. Generally, variables are denoted by X or Y or Z.

Types of variable

The statistical data can be divided into two broad categories. So, there are two types of variables

1. Categorical variable or Qualitative variable
2. Numerical variable or Quantitative variable

Qualitative variable

Qualitative variable is a variable or characteristic which cannot be measured in quantitative form but can only be identified by name or categories. The qualitative variables are just categorized. It is also called categorical variable. Categorical variables take a value that is one of several possible categories. As naturally measured, categorical variables have no numerical meaning .The data are classified by counting the individuals having the same characteristics or attributes and but not by measurement. For Examples: Hair color (black, grey ,white) , gender: male/female, religion , nationality , marital status, disease: present/absent, eye color (black, brown ,blue), vaccine: vaccinated/ not vaccinated , smoke: smoking/not smoking, stages of cancer (I, II, III, or IV), blood group (A, B, AB, O) , degree of pain (minimal, moderate, severe or unbearable) etc.

The qualitative variable if it has only two categories is binary or dichotomous variable such as male/female, married/single, live/death etc. Data of this type are called nominal data and ordered categorical data such as pain: mild/moderate/severe and cigarette consumption: non-smoker/light smoker/ heavy smoker, socioeconomic status: lower, middle, upper are also called ordinal data.

Qualitative variable may be further classified into two categories

(i) Nominal variable
(ii) Ordinal variable

Nominal variable

Qualitative variables which can be categorized into various categories such that the numbers or symbols 0, 1, 2 etc. assigned have no numerical meaning are called nominal variables. It describes the differences but not the differences between the numbers. It does not satisfy ordinary arithmetic properties such as addition, subtraction, multiplication, division etc.

For examples, gender (male/female), marital status (married/single), profession (Teacher, doctor, engineer, officer etc.), religion (Hindu, Muslim. Buddhist etc.), hair color (black, grey, white) etc.

Ordinal variable: Qualitative variables which can be ordered and ranked categorical data are called ordinal variables. For examples, pain: mild/moderate/severe; cigarette consumption: non-smoker/light smoker/ heavy smoker; socioeconomic status: lower, middle, upper; attitude towards certain fact: positive, negative, bad etc.

Quantitative variable: A quantitative variable is one that can be measured and expressed numerically. Therefore, quantitative variable describes the characteristics in term if a numerical value, which are expressed in units of measurements. For example, height, weight, age. Income, pulse rate, blood pressure, level of hemoglobin in blood, number of children in a family, temperature records etc. These data may be represented by interval and ratio scale. Quantitative Variables can be divided into two types based on the nature of the characteristics. They are:

1. Discrete variable
2. Continuous variable

Discrete variable: A variable is said to be discrete if it takes only countably many values (i.e. whole numbers). It is a variable whose values are obtained by counting.

For example,
No. of goals scored in a football match.
No. of children in a family : 1, 2, 3, 4, 6
WBC count: 2ooo, 3010, 4060, 5050
No. of Bed in hospital wards : 50, 75, 100, 500, 1000.

Continuous variable: A variable is said to be continuous if it takes all possible real values (i.e. whole number as well as fractional values) within certain range. It is a variable whose values are obtained by measuring. For examples: height, weight, age, income, pulse rate, blood pressure, level of hemoglobin in blood, cholesterol etc.
i.e.
Weight of children (in kg): 8, 9. 9.5, 9.8, 10, 10.3
Height of patients (in feet): 4.6, 4.9, 5, 5.4, 5.6, 5.8, 6

Types of Data and Data collection

By definition of Statistics, data is an aggregate of facts which can numerically be expressed. So, Data is one of the main sources of information. The first step of statistical approach (or study) is collection of data. Collection of data means the methods that are to be used for getting necessary information from units under investigation. The process of getting necessary information from the units under investigation is called collection of data. The method of data collection depends upon the nature, objective, and the scope of inquiry. The quality of information depends upon the source of data, the methods of collection of data and recording system.

There are mainly two types of data on the basis of collection procedures. They are as follows

1. Primary data
2. Secondary data

Primary data

Primary data are those fresh and original data, which are collected and recorded by the Investigator or researcher. Therefore, the data which are originally collected by investigator or researcher or an agent for the first time for the purpose of statistical enquiry (or investigation) are known as primary data. They are the first hand, unique, original, reliable and accurate in character. Primary data are collected for specific purpose of study of the investigator or researcher. For example, if a public health student wants to study the fertility behavior of women in Jumla, the data collected for this purpose by him/her are primary data.

Methods of collecting primary data

There are various methods of collecting primary data.

1. Interview method
   (a) Direct personal interview method
   (b) Indirect personal interview method
   (c) Telephone interview method
2. Information through correspondence
3. Mailed questionnaire method

4. Schedule sent through enumerators
5. Internet
6. Observation method.

Interview method

The process of obtaining the information through interview with respondents by the enumerator is called interview method of obtaining data.

Direct personal interview method

In this method, the investigators or researchers meet the respondents (or informants) personally and asked the necessary questions and extract the required data from them. The persons from whom informations are collected are known as informants. Since the researcher collects the information, the information is accurate, fresh and reliable. This method is effective for small number of respondents. It suits best for intensive study of the limited field.
This method is suitable in the following situations

- When the area of inquiry is limited.
- When the data is to be kept secret
- When pure and original data is needed.

Merits

- Information obtained by this method is generally accurate.
- Investigator or researcher can collect other supplementary information related with the study.
- The accuracy of data can be verified in the spot as the investigator is being involved in collecting data
- People willingly supply information because they are approached personally.
- Sensitive questions can be asked because informants feel at home with the interviewer.

Demerits

- It is costly and time consuming
- It is not useful for wide area of inquiry
- There is chance of getting biased information

Indirect personal interview method

This method is applied when the informants hesitate to provide information directly. The information regarding property, personal habitats such as smoking habitat, drug addicts, using family planning measures etc. In this method, information can generally be collected with the help of third person who is well known and familiar with the respondents. The third person here is said to be witness.

It is suitable

- When the direct sources of information are not available
- When the informants are reluctant (i.e. to hesitate) to give information.
- When the field of investigation is very wide.

Merits

- This method is comparatively economic.
- It is convenient way of gathering sensitive information with the help of witness.
- It is applicable for wider area of inquiry.

Demerits

- The witness may be biased to provide accurate information.
- There will be lack of interest to provide information about respondents.

Information through correspondence

This is a correspondence method. In this method, the investigator appoints the correspondents in different places to collect information. The correspondents collect information and send to central office or to the investigator and finally central office or investigator utilize the data. This method is more suitable in the field of news media.

Merits

- This method is economic (cheap) and appropriate for extensive investigations.
- The information from wide area can be easily obtained.

Demerits

- Information through this method may not be uniform.
- It is time consuming method if there is no facility of communication such as telephone, internet etc.

Mailed questionnaire method

Under this method, a list of questions is prepared and is sent to all the informants (or respondents) by post. The list of questions is technically called questionnaire .The respondents (or informants) are requested to answer these questions and return to the investigator within a specified time. A covering letter accompanying the questionnaire explains the purpose of the investigation and the importance of correct information. This method is appropriate in those cases where the informants are literates and are spread over a wide area.

Merits:

1. It is relatively cheap.
2. It is preferable when the informants are spread over the wide area.
3. The information collected through this method is accurate since the educated respondents provide information.

Demerits:

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

2 It has high chance of non-response.

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

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

Schedule sent through enumerators

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

Merits

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

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

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

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

5. It is most popular methods.

Demerits:

- This is time consuming and expensive method

Telephone interview method

If enumerator or researcher collects the information through telephone is called telephone interview method. It is becoming more popular these days. If the respondents have telephone access, this method is may be very effective. It is so rapid and cheap method of data collection.

Merits

- Information obtained by this method is generally accurate.
- It is rapid and cheap.
- Response rate is high.

Demerits

- People without access of telephone will be ignored even if they are very important.
- It is not useful for wide area of inquiry
- There is chance of getting biased information.

Internet

In this age of information technology, internet is also gaining popularity as a method of data collection. This method enables the investigator to get an access to the data directly or by inquiry with the concerned individual or agency to get required data.

Merits:

- This method covers the whole world geographically.
- The data collected can be the most recent and up to date.

Demerits

- Investigator who does not have an access to the internet will be unable to use it.
- This method might be costly.

Observation method

The method by which researcher or enumerator collects the information observing the information, observing phenomenon about issue or individual or commodity culture or experimental units etc.

Merits

- Non-response rate is absent.
- Confusion can be reduced at spot.
- It consists high reliability.

Demerits

- It is expensive.
- It covers less area.

Problems involved in collecting primary data

There are various problems that have to be faced while collecting primary data. The nature of problems depends upon the different situations. Some of the problems are as follows.

- Generally there is lack of time and money while collecting primary data.
- Transportation facility is essential for the enumerators or investigators to and from the research field. This is one of the major problems for under developed countries like Nepal.
- There is high degree of non-response error from illiterate respondents.

Secondary Data:
Secondary data are those data which have been already collected and analyzed by some earlier agency for its own use; and later the same data are used by a different agency. The data is a primary for those persons or institutions that collect them but the same data become secondary for another. Actually secondary data are the data, which are borrowed from others who have collected them for some other purpose. This type of data is not original in nature. The degree of accuracy of this type of data is comparatively less than that of the primary data.

Sources of Secondary data:
In most of the studies the investigator finds it impracticable to collect first-hand information on all related issues and as such he makes use of the data collected by others. In order to collect secondary data, the following sources may be used.

1 Published sources, and
2 Unpublished sources

Published Sources:
The various sources of published data are:

- Reports and publications of ministries, departments of the government.
- Reports and publications of reputed INGO'S such as UNDP, ADB, UNESCO, WHO, World Banks etc.
- Reports and publications of reliable NGO'S, NDHS (Nepal Demographic Health Survey), Journals, periodicals etc.
- Publications brought out by research agencies, research scholars, etc.

In recent years, the internet has become one of the important sources of data. Almost all reputed companies, business firms, or institutions have internet web sites and provides public access.

Unpublished Sources

All statistical material is not always published. There are various sources of unpublished data such as:

- Records maintained by various Government and private offices.
- Records maintained by research institutions, research scholars, etc.
- Records updated by the department's institutions for their internal purpose.

Merits of secondary data

- It saves time and cost.
- The scope of inquiry can be increased in terms of area and time period to be covered.
- Much of the secondary data available has been collected for many years and therefore it can be used to plot trends.
- It helps the government in making decisions and planning future policy

Demerits of secondary data

- Data may not be in the exact form of the requirement of the researcher.
- Some information is often omitted or some categories are pooled.

Precaution in using secondary data
The secondary data available from any organizations or agencies can not directly be used for the statistical investigation because the data may be erroneous, not suitable and inadequate to the problem

under study. So, before using the secondary data to the problem under study, the investigator should be more careful on the following factors

- Reliability of data
- Suitability of data and
- Adequacy of data

Reliability of data.

The investigator or researcher should be very careful about the reliability of data. If data are not reliable, even it is treated and interpreted properly; it may not give good result. It means that the result may miss guide. So, the investigator or researcher must be serious about the following points.

- The reliability, integrity and experience of the organization or institutions.
- The reliability of sources of information and
- The methods applied for the collection and analysis of the data

It is essential that the collecting agency or organization must be unbiased in collecting data. Data should be collected in normal time. That is the time should be free from abnormalities, political upheavals or any abnormalities.

Suitability of data

Even if the data are reliable, it should be used after confirming its suitability. It is essential to see whether the collected data are suitable for the purpose the inquiry or not. In this case, the investigator should compare the objectives, nature and scope of the given inquiry with the original investigation. Investigator should confirm various terms and units which are clearly defined and uniform or not throughout the earlier investigation.

Adequacy of data

Even if the data are reliable and suitable, it is necessary to see whether it is adequate or not for the inquiry .Thus, in order to arrive at conclusion free from limitations and accuracies, the secondary data must be subjected through scrutiny and editing before accepted for use.

Advantages of Secondary Data

1. It saves time and cost.
2. If specially trained persons collect it, the quality of secondary data is better.
3. The scope of inquiry can be increased in terms of area and time period to be covered.

Disadvantages of Secondary data

1. Data may not be in the exact form of the requirement of the researcher.

2. Some information is often omitted or some categories are pooled.

Difference between primary and secondary data

The difference between primary and secondary data is basically depends on the mode of collection of data. The data which are primary for one agency is treated as secondary for other and vice versa. However, the main differences between primary and secondary data are as follows:

Primary data	Secondary data
• Primary data are original in the sense that they are personally collected by the investigator or researcher involving himself / herself.	• Secondary data are not original in the sense that they are collected by someone other than the investigator or researcher.
• Primary data collection is more expensive and exhaustive.	• Secondary data are readily available at less expense.
• Primary data are collected as per requirement of the investigator.	• Secondary data might have been collected with different objectives.
• Primary data may be influenced by personal prejudice of the investigator.	• Secondary data may not be influenced by the personal prejudice of the investigator.

MEASURES OF DISPERSION

Introduction

The measure of central tendency like mean, median and mode give only the idea about the concentration of the items around the central values but are unable to give the clear information about the variability or the scatterness of the items from the central values is called dispersion and its measure is the measure of dispersion or the measure of variation.

Central tendency just measures the central value of the given items which is supposed to represent mass of the data. However, it may not be true if the values of the items highly scattered or spread. Thus, to ensure the reliability of the computed central value, measure of dispersion is used. Thus, measures of dispersion are statistical tools i.e. descriptive statistical measures which are used to measure the variation or spread or scatterness or deviation of data from the central value. So, it gives an idea of homogeneity or heterogeneity of the distribution. For example, let us consider the following three series A, B and C with the same number of observations i.e.n= 5

	Observations	Total	Mean ($\overline{X}$)	Median (Md)	Mode (Mo)
A	20 20 20 20 20	100	20	20	20
B	18 19 20 20 23	100	20	20	20
C	12 16 20 20 32	100	20	20	20

If we carefully study the above table, we will notice that all the three series A, B and C have same number of observations i.e. n = 5 and have the same mean, median and mode but they differ completely (or entirely) in formation or structure. Series A is stationary or constant and shows no dispersion or variability.Series B is slightly dispersed and series C is relatively more dispersed.In other words, series B is more homogeneous than series C and alternatively series C is more heterogeneous than series B.Though three series A, B and C have same averages, they can not be similar because they are differently constituated.

Thus, only the measures of central tendency are inadequate to describe (or characterize) the distribution completely. Therefore, the measures of central tendency must be supported and supplemented by some other measures.One of such measure is dispersion or variability.

Definition

The definitions of the prominent statisticians about the dispersion are given below:

"Dispersion or spread is the degree of the scatter or variation of the variables about a central value." **B.C. Brooks & W.F.L. Dick**

Objectives of the Measures of Dispersion

The main objectives of measuring variability or Dispersion are as follows:

(i) To determine the reliability of an average

(ii) To control the variation of the data from the central value.

(i) To compare two or more series or sets of data with regard to their variability.

(ii) To facilitate the use of other statistical measures for further analysis of data.

(iii) To help in devicing a system of quality control.

Characteristics for an ideal (or good) measure of Dispersion (i.e. Requisites of an ideal or good measure of dispersion)

The following are the requisites to be satisfied by different measures of dispersion to become an ideal measure.

(i) It should be rigidly defined.

(ii) It should be simple to understand and easy to calculate.

(iii) It should be based on all the observations

(iv) It should be suitable for further mathematical treatment.

v) It should be least affected by fluctuations of sampling.

(vi) It should not be much affected by extreme observations.

Absolute and Relative Measures of Dispersion

There are two types of dispersion: (1) absolute and (2) relative.

Absolute measures of Dispersion: The measures of dispersion which are dependent to original units of measurement of data are said to be absolute measures of dispersion. The absolute measures of dispersion can be used only for comparing the variability or dispersion of two or more distributions having in the same unit. If the distributions are given in different units, then for comparison, the absolute measures of dispersion cannot be used.

E.g. Income x: Rs. 800

Income y: Rs. 300

$\therefore$ Absolute measure of dispersion = Rs. 800 – Rs.300 = Rs. 500

Relative measures of Dispersion: The measures of dispersion which are pure numbers, independent of units of measurement of the data are said to be relative measures of dispersion. Thus, the relative measures of dispersion will not have any unit of measurement and are obtained by taking a ratio or percentage of an absolute measure of dispersion to a suitable average. i.e. the ratios of two absolute values.

$\therefore$ Relative value = $\frac{Rs.800}{Rs.300} = \frac{8}{3}$

Therefore, for comparing the variability of two or more than two distributions even if they are measured in the same units, for convenience results, the relative measures of dispersion instead of the absolute measures of dispersion are computed.

Therefore, the various measures of dispersion are as follows.

1. Range
2. Quartile deviation or Semi-interquartile range
3. Mean deviation or Aerage deviation.
4. Standard deviation
5. Lorenz curve

Note: But Lorenz curve is beyond of our syllabus.

Thus, the abosolute and relative measures of dispersion are given in the following table:

Absolute measure	Relative measure
1. Range 2. Quatile deviation or Seim-interquartile range 3. Mean deviation 4. Standard deviation	1. Coefficient of range 2. Coefficient of quartile deviation. 3. Coefficient of mean devaiation. 4. Coefficient of variation
5. Graphical method: Lorenz curve.	

Range

Range is the simplest of all the measures of dispersion. It is defined as the difference between largest (maximum) value and smallest (minimum) value for the given observations of the distribution. If the largest and the smallest observation of the distribution are L and S respectively. Then, the range is denoted by R.

Range (R) = Largest item – Smallest item = L – S.

Where, X_{max} = L = Larest item or observation

X_{min} = S = Smallest item or observation

In case of grouped frequency distribution i.e. continuous frequency distribution, range is defined as the upper limit of the highest class and the lower limit of the smallest class.

Range is an absolute measure of dispersion and depends upon the units of measurement. if we want ot compare the variability of two or more distributions with same units of measurement, we may use but for different units, we may not use.Its relative measure of dispersion is known as coefficient of range and the coefficient of range is given by

$$\textbf{Coefficient of range} = \frac{\mathbf{L - S}}{\mathbf{L + S}}$$

It is a relative measure of dispersion.

For individual Series

[redacted]

(a) Obtain the range and coefficient of range for the following data:

Height (in cm): 155, 160, 158, 156, 162, 165

Solution: We have,

Largest value (L) = 165 cm

Smallest value (S) = 155 cm

Range (R) = 165 – 155 cm = 10 cm

$$\text{Coefficient of range} = \frac{L - S}{L + S} = \frac{165 - 155}{165 + 155} = \frac{10 \text{ cm}}{320 \text{ cm}} = 0.0312$$

(b) Calculate the range and coefficient of range from the following data:

2, 3,5,7,12,15,8,20

Solution:

Largest value (L)= 20

Smallest value (S) = 2

Range (R) = 20 –2 = 18

Coefficient of range = $\frac{L - S}{L + S} = \frac{20-2}{20+2} = 0.818$

For Discrete Series

[redacted]

Find range and coefficient of range from the following wage distribution:

Daily wages (in Rs.)	50	110	80	120	130
No. of works	20	15	30	15	8

Solution: We have given,

Largest item (L) =130

Smallest item (S) = 50

Range (R) = L – S = 130 – 50 = Rs. 80

Coefficient of range = $\frac{L - S}{L + S} = \frac{130 - 50}{130 + 50} = \frac{80}{180} = 0.444$

For Continous Series

[redacted]

Find range and coefficient of range from the following grouped frequency distribution:

Marks	0-30	30-60	60-90	90-120	120-150
No. of students	10	20	18	25	5

Solution: We have given,

Largest item (L) = 150

Smallest item (S) = 0

Range (R) = L – S = 150 – 0 = 150 marks

Coefficient of range = $\frac{150-0}{150+0} = \frac{150}{150} = 1$

Merits of Range

- It is rigidly defined.
- It is simple to understand and easy to calculate.
- Its computation is very fast. Hence, if we want to know a quick rather than a very accurate picture of variability, we may compute range.

- It is mostly used in calculating the range of meteorological data such as data related to temperature, rainfall etc.
- Range is used in industry for statistical quality control of manufactured product by the construction of R-chart i.e. control chart for range.
- Range is also useful in studying the variations in the prices of stocks and shares and other commodities that are sensitive to price changes from one period to another.

Demerits of Range

- It is not based on all the observations. That is, it is only considers the largest and smallest values of the distribution.
- It can not be calculated in case of open end classes frequency distribution.
- It is much affected by fluctuation of sampling.
- It is affected by extreme observations.
- It is not suitable for further mathematical treatments.

Quartile Deviation or Semi-interquartile Range

Quartile deviation is a measure of dispersion based on the upper quartile Q_3 and lower quartile Q_1. The difference between the upper quartile Q_3 and lower quartile Q_1 is known as inter-quartile range.

Inter-quartile range = $Q_3 - Q_1$

The half of the inter-quartile range is called semi-interquartile range, which is also known as quartile deviation. Generally, it is denoted by quartile deviation (Q.D.) and is given by

Quartile deviation (Q.D.) = $\dfrac{Q_3 - Q_1}{2}$

Where,

Q_3 = Third quartile or upper quartile

Q_1 = First quartile or lower quartile

Quartile deviation is an absolute measure of dispersion. The relative measure based on the lower and upper quartiles known as coefficient of quartile deviation. For comparative studies of variability of two or more than two distributions, we need relative measure of dispersion which is known as coefficient of quartile deviation and is given by

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

Note:

- Quartile deviation (Q.D.) is the most suitable or appropriate measure of dispersion for open end classes.
- Less the coefficient of Q.D. implies more will be the uniformity or less will be the variability.
- Greater the coefficient of Q.D. implies less will be the uniformity or greater will be the variability.

For Individual Series

[redacted]

From the following data, obtain interquartile range, Q.D. & coefficient of Q.D.

Daily production: 25, 20, 23, 18, 22, 17, 26

Solution:

Arranging the given data in ascending order, we get,

17, 18, 20, 22, 23, 25, 26

Now,

First quartile (Q_1) = Value of $\left(\frac{n+1}{4}\right)^{th}$ item

= Value of $\left(\frac{7+1}{4}\right)^{th}$ item

= Value of 2^{nd} item

$Q_1 = 18$

Again,

Third quartile (Q_3) = Value of $\left(\frac{3(n+1)}{4}\right)^{th}$ item

= Value of $\left(\frac{3(7+1)}{4}\right)^{th}$ item

= Value of 6^{th} item

$Q_3 = 25$

Interquartile range = $Q_3 - Q_1$

= 25 – 18

= 7

Quartile deviation or Semi-interquartile range (Q.D.) = $\frac{Q_3 - Q_1}{2} = \frac{25-18}{2} = \frac{7}{2} = 3.5$

$\therefore$ Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{25-18}{25+18} = \frac{7}{43} = 0.163$

[redacted]

Find the interquartile range, Q.D. and Coefficient of Q.D. from the following series:

X: 9 10 5 6 7 2 8 4

Solution

Arranging the given data in ascending order, we get,

X: 2 4 5 6 7 8 9 10

Now,

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

Again,

Q_1 = Value of $\frac{(n+1)^{th}}{4}$ item

= Value of $\left(\frac{9}{4}\right)^{th}$ item = Value of 2.25^{th} item

= Value of 2^{nd} item + 0.25(3^{th} item - 2^{nd} item)

So, Q_1 = 4 + 0.25 (5 – 4) = 4.25

Q_3 = Value of $\frac{3(n+1)^{th}}{4}$ item

= Value of $\left(\frac{3 \times 9}{4}\right)^{th}$ item

= Value of 6.75^{th} item

= Value of 6^{th} item + 0.75(7^{th} item - 6^{th} item)

Q_3 = 8 + 0.75 (9 – 8) = 8.75

Interquartile range = $Q_3 - Q_1$

= 8.75 – 4.25

= 4.5

Quartile deviation (or Semi-interquartile range) Q.D. = $\frac{Q_3 - Q_1}{2} = \frac{4.5}{2} = 2.25$

Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{4.5}{13} = 0.346$

For Discrete Series

[redacted]

Findout the Q.D. and coefficient of Q.D. from the following distribution:

Size	10	12	15	18	20	22
No. of parts	6	14	20	16	8	6

Solution:

Computation of less than cumulative frequency

Size	No. of parts (f)	Less than c.f.
10	6	6
12	14	20
15	20	40
18	16	56
20	8	64
22	6	70
	N = $\sum f$ = 70	

Now,

First quartile (Q_1) = Value of $\left(\frac{N + 1}{4}\right)^{th}$ item

= Value of $\left(\frac{70 + 1}{4}\right)^{th}$ item

= Value of $(17.75)^{th}$ item

= 12

Third quartile (Q_3) = Value of $3\left(\frac{N+1}{4}\right)^{th}$ item

= Value of $\left(3 \times \frac{71}{4}\right)^{th}$ item

= Value of $(53.25)^{th}$ item

$Q_3 = 18$

Interquartile range = $Q_3 - Q_1$

= 18 – 12

= 6

Quartile deviation (or Semi-interquartile range) Q.D. = $\frac{Q_3 - Q_1}{2} = \frac{18-12}{2} = 3$

Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{18-12}{18+12} = \frac{6}{30} = \frac{1}{5} = 0.2$

For Continuous Series

[redacted]

Calculate the interquartile range, Q.D. and Coefficient of Q.D. from the following groups:

Daily earning (Rs.)	50-100	100-200	200-250	250-300	300-400
No. of Persons	5	7	8	6	2

Solution:

X	f	c.f.
50-100	5	5
100-200	7	12
200-250	8	20
250-300	6	26
300-400	2	28

Here,

For $Q_1 = \left(\frac{N}{4}\right)^{th}$ item $= \left(\frac{28}{4}\right)^{th}$ item $= 7^{th}$ item

Q_1 lies in the class 100-200.

L = 100, f = 7, c.f. = 5, h = 100

Now,

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

$$= 100 + \frac{7-5}{7} \times 100 = 100 + \frac{200}{7} = 100 + 28.57 = 128.57$$

For $Q_3 = \frac{3 \times 28}{4} = 21$

Q_3 lies in the class 250-300.

L = 250, f = 6, c.f. = 20, h = 50

Now, $Q_3 = L + \frac{\frac{N}{4} - c.f.}{f} \times h$

$$= 250 + \frac{21-20}{6} \times 50 = 100 + \frac{50}{6} = 100 + 8.33 = 258.33.$$

Interquartile range = $Q_3 - Q_1$
= 258.33 – 128.57
= 129.76

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{258.33 - 128.57}{2} = \frac{129.76}{2} = Rs.64.88$$

$$\text{Coefficient of Q.D.} = \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{129.76}{386.9} = 0.33$$

Calculate suitable or appropriate measure of dispersion from the following data. Support the choice for the measure.

Height	below 2	2-5	5-10	10-15	15 & above
Frequency	5	7	8	6	2

Solution: Since, the given frequency distribution has open end classes, therefore quartile deviation (Q.D.) is the appropriate measure of dispersion.

Size	Frequency	Less than c.f.
below 2	3	3
2-5	8	11
5-10	11	22

10-15	12	34
15 & above	6	40
Total	N = $\sum f$ = 40	

For first quartile (Q_1):

$$\frac{N}{4} = \frac{40}{4} = 10$$

Q_1 lies in class is 2-5

Now,

L = 2, c.f. = 3, f = 8, h = 3

$$Q_1 = L + \frac{\frac{N}{4} - \text{c.f.}}{f} \times h = 2 + \frac{10 - 3}{8} \times 3 = 2 + 2.625 = 4.625$$

For third quartile (Q_3):

$$\frac{3N}{4} = \frac{3\times40}{4} = 30$$

Q_3 lies in class is 10-15

L = 10, c.f. = 22, f = 12, h = 5

$$Q_3 = \frac{\frac{3N}{4} - \text{c.f.}}{f} \times h = 10 + \frac{30 - 22}{12} \times 5 = 10 + 3.33 = 13.33$$

Quartile deviation (Q.D.) = $\frac{Q_3 - Q_1}{2} = \frac{13.33-4.625}{2} = 4.3525$

[redacted]

The distribution of fortnightly wages of 280 employees of an undertaking is as follows

Fortnightly wages (in Rs.)	Less than 200	200-399	400-599	600-799	800-999	1000-1199	1200-1399	1400 & above
No.of employees	12	16	38	78	80	35	14	7

Calculate the appropriate measures of dispersion and its coefficient from the following distribution and support for your choice of measure.

Solution: Since, the given frequency distribution has open end classes; therefore quartile deviation (Q.D.) is the appropriate measure of dispersion.

Fortnightly wages (in Rs.)	No. of employees (f)	Less than c.f.
Less than 200	12	12
200-399	16	28
400-599	38	66
600-799	78	144
800 -999	80	224
1000-1199	35	259
1200 -1399	14	273
1400 & above	7	280
Total	N = $\sum f$ = 280	

Correction factor = $\frac{400-399}{2} = \frac{1}{2} = 0.5$

For first quartile (Q_1):

$\frac{N}{4} = \frac{280}{4} = 70$

Q_1 lies in class is 600-799 but its exclusive class is 599.5-799.5

Now,

L = 599.5, c.f. = 66, f = 78, h = 200

$$Q_1 = L + \frac{\frac{N}{4} - c.f.}{f} \times h = 599.5 + \frac{70-66}{78} \times 200 = 609.7564$$

For third quartile (Q_3):

$\frac{3N}{4} = \frac{3\times280}{4} = 210$

Q_3 lies in class is 800-999 but its exclusive class is 799.5-999.5

L = 799.5, c.f. = 144, f = 80, h = 200

$$Q_3 = \frac{\frac{3N}{4} - \text{c.f.}}{f} \times h = 799.5 + \frac{210-144}{80} \times 200 = 964.5$$

Quartile deviation (Q.D.) = $\frac{Q_3 - Q_1}{2} = \frac{964.5 - 609.7564}{2}$ = Rs. 177.3718

Coefficient of Q.D. = $\frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{964.5 - 609.7564}{964.5 + 609.7564} = 0.22534$

[redacted]

Calculate the quartile deviation (or semi-interquartile range) and coefficient of quartile deviation from the following data.

Marks	Below 20	Below 40	Below 60	Below 80	Below 100
No. of students	8	20	50	70	80

OR

Marks (Below)	20	40	60	80	100
No. of students	8	20	50	70	80

Solution:

Marks	f	Less than c.f.
below 20	8	8
20-40	12	20
50-60	30	50
60-80	20	70
80 -100	10	80
Total	N = $\sum f$ = 80	

For first quartile (Q_1):

$\frac{N}{4} = \frac{80}{4} = 20$

Q_1 lies in class is 20 - 40

Now,

L = 50, c.f. = 8, f = 12, h = 20

$$Q_1 = L + \frac{\frac{N}{4} - c.f.}{f} \times h = 20 + \frac{20-8}{12} \times 20 = 40$$

For third quartile (Q_3):

$$\frac{3N}{4} = \frac{3\times80}{4} = 60$$

Q_3 lies in class is 60-80

L = 60, c.f. = 50, f = 20, h = 20

$$Q_3 = \frac{\frac{3N}{4} - c.f.}{f} \times h = 60 + \frac{60-50}{20} \times 20 = 70$$

Quartile deviation (Q.D.) = $\frac{Q_3-Q_1}{2} = \frac{70-40}{2}$ = 15 marks

Coefficient of Q.D. = $\frac{Q_3-Q_1}{Q_3+Q_1} = \frac{70-40}{70+40}$ = 0.272727

[redacted]

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

Marks in Statistics	Above 0	Above 10	Above 20	Above 30	Above 40	Above 50	Above 60	Above 70
No. of Stufents	150	140	100	80	74	70	30	14

OR

Marks in Statistics (More than)	0	10	20	30	40	50	60	70
No. of Stufents	150	140	100	80	74	70	30	14

Solution:

Marks in Statistics	f	Less than c.f.
0-10	10	10
10-20	40	50
20-30	20	70
30-40	6	76
40 -50	4	80

50-60	40	120
60-70	16	136
70 & more	14	150
Total	N = $\sum f$ = 150	

For first quartile (Q_1):

$\frac{N}{4} = \frac{150}{4} = 37.5$

Q_1 lies in class is 10 - 20

Now,

L = 10, c.f. = 10, f = 40, h = 10

$$Q_1 = L + \frac{\frac{N}{4} - \text{c.f.}}{f} \times h = 10 + \frac{37.5-10}{40} \times 10 = 16.875$$

For third quartile (Q_3):

$\frac{3N}{4} = \frac{3\times150}{4} = 112.5$

Q_3 lies in class is 50-60

L = 50, c.f. = 80, f = 40, h = 10

$$Q_3 = \frac{\frac{3N}{4} - \text{c.f.}}{f} \times h = 50 + \frac{112.5-80}{40} \times 10 = 58.125$$

Quartile deviation (Q.D.) = $\frac{Q_3-Q_1}{2} = \frac{58.125-16.875}{2}$ = 20.625 marks

Coefficient of Q.D. = $\frac{Q_3-Q_1}{Q_3+Q_1} = \frac{58.125-16.875}{58.125+16.875} = 0.55$

[redacted]

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

Central Value	6	9	12	15	18
Frequency	14	24	38	20	4

Solution:

Class size (h) = Difference between two successive mid. values

= 9 – 6

h = 3

$\frac{h}{2}$ = 1.5

Subtract the value of $\frac{h}{2}$ from each mid. value to get lower limit and add $\frac{h}{2}$ to the same mid. value to get upper limit.

C.I.	f	Less than c.f.
4.5 - 7.5	14	14
7.5-10.5	24	38
10.5-13.5	38	76
13.5-16.5	20	96
16.5 -19.5	4	100
Total	N = $\sum f$ = 100	

For first quartile (Q_1):

$\frac{N}{4} = \frac{100}{4} = 25$

Q_1 lies in class is 7.5 – 10.5

Now,

L = 7.5, c.f. = 14, f = 24, h = 3

$$Q_1 = L + \frac{\frac{N}{4} - \text{c.f.}}{f} \times h = 7.5 + \frac{25-14}{24} \times 3 = 8.875$$

For third quartile (Q_3):

$\frac{3N}{4} = \frac{3\times100}{4} = 75$

Q_3 lies in class is 10.5-13.5

L = 10.5, c.f. = 38, f = 38, h = 3

$$Q_3 = \frac{\frac{3N}{4} - \text{c.f.}}{f} \times h = 10.5 + \frac{75-38}{38} \times 3 = 13.42105$$

Quartile deviation (Q.D.) = $\frac{Q_3-Q_1}{2} = \frac{13.42105-8.875}{2}$ = 2.273025

Relative measure of quartile deviation is

Coefficient of Q.D. = $\frac{Q_3-Q_1}{Q_3+Q_1} = \frac{13.42105-8.875}{13.42105+8.875}$ = 0.203895

Merits of Quartile Deviation

- It is only of the dispersion, which can be calculated for the frequency distribution having open-end classes.
- It is rigidly defined.
- It is simple to understand and easy to calculate.
- It is not affected by extreme observations.
- It is very useful when it as desired to know variability in the central half part of the data.

Demerits of Quartile Deviation

- Q.D. is not based on all the observations. Since it ignores 25% of the data at the lower end and 25% of the data at the upper end of the distribution.
- It is much affected by fluctuations of sampling.
- It is not suitable for further mathematical treatment.

Mean Deviation or Average deviation

Range and Quartile deviation are not the better measure of dispersion as all the items are not included in both cases and they do not show the variations of the items or observations from an average. Thus, they completely ignore the composition of the distribution. So, to overcome both these drawbacks, the other measure of dispersion is developed, which is known as mean deviation or average deviation or mean absolute deviation. Generally, it is denoted by M.D.

Mean deviation is defined as the arithmetic mean of the absolute (modulas) deviations of the items or observations taken from their central values. The central values generally are taken as mean, median or mode. If the average deviation is taken from mean, median, or mode, then they are said to be mean deviation from mean, mean deviation from median or mean deviation from mode respectively.

In this case of

1Individual Series: Let $x_1, x_2, \ldots\ldots x_n$ be the n variate values of a variable X. Then, the mean deviation is defined by the following formula:

We have,

M.D. from Mean = $\frac{\sum|X-\overline{X}|}{n}$

M.D. from Median = $\frac{\sum|X-Md|}{n}$

2. Discrete Series: Let $x_1, x_2, x_3, \ldots\ldots, x_n$ be the n variate values of a variable X and $f_1, f_2, f_3, \ldots\ldots f_n$ be their corresponding frequencies. Then, M.D. is defined by the following formula:

M.D. from Mean = $\frac{\sum f|X - \overline{X}|}{N}$

M.D. from Median = $\frac{\sum f|X - Md|}{N}$

Where, X = value of variable.

3. Continuous series (grouped frequency distribution)

Formula as same in discrete series but X is middle value of the class.

M.D. from Mean = $\frac{\sum f|X - \overline{X}|}{N}$

M.D. from Median = $\frac{\sum f|X - Md|}{N}$

Coefficient of Mean deviation from mean and median

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

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

[redacted]

Find mean deviation from mean for the following data:

Profit (in '000' Rs.): 20, 23, 24, 27, 31

Solution:

For mean deviation from mean:

Computation of Sum of Values

Profit (in '000' Rs.) (X)	$\|X-\overline{X}\|$
20	5
23	2
24	1
27	2
31	6
$\sum X$ = 125	$\sum\|X - \overline{X}\|$ = 16

Now,

$$\text{Mean } (\bar{X}) = \frac{\sum X}{n} = \frac{125}{5} = 25$$

$$\text{M.D. from Mean} = \frac{\sum |X - \bar{X}|}{n} = \frac{16}{5} = \text{Rs.3.2 (in 000)}$$

$$\text{Coefficient of M.D. from Mean} = \frac{\text{M.D. from Mean}}{\text{Mean}} = \frac{3.2}{25} = 0.1258$$

From the following weight distribution, find mean deviation from mean, median and mode, and their corresponding coefficient of M.D.:

Variable	10	12	15	16	20
No. of parts	6	14	20	13	7

Solution:

Computation of Sum of Values for M.D.

Weight (X)	No. of parts(f)	d = X-15	fd	$\lvert X-\bar{X}\rvert$	$f\lvert X-\bar{X}\rvert$	c.f.	$\lvert X-Md\rvert$	$f\lvert X-Md\rvert$
10	6	–5	–30	–4.6	27.6	6	5	30
12	14	–5	–42	–2.6	36.4	20	3	42
15	20	0	0	0.4	8	40	0	0
16	13	1	13	1.4	18.2	53	1	13
20	7	5	35	5.4	37.8	60	5	35
	N= $\sum f$ =60		$\sum fd$ = –24		$\sum f\lvert X-\bar{X}\rvert$ =128			$\sum f\lvert X-Md\rvert$ =120

Now,

$$\text{Mean } (\bar{X}) = A + \frac{\sum fd}{N} = 15 - \frac{24}{60} = 15 - 0.4 = 14.6$$

$$\text{M.D. from Mean} = \frac{\sum f|X - \bar{X}|}{N} = \frac{128}{60} = 2.13$$

Coefficient of M.D. from Mean = $\frac{\text{M.D. from Mean}}{\text{Mean}} = \frac{2.13}{14.6} = 1.2$

Again,

Median (M_d) = value of $\left(\frac{N+1}{2}\right)^{th}$ item = value of $\left(\frac{60+1}{2}\right)^{th}$ item

= value of $(30.5)^{th}$ item

$\therefore$ $M_d = 15$

M.D. from Median = $\frac{\sum f|X - Md|}{N} = \frac{120}{60} = 2$

Coefficient of M.D. from Median = $\frac{\text{M.D. from Median}}{\text{Median}} = \frac{2}{15} = 0.133$

Calculate M.D. and coefficient of M.D. from Mean from the following distribution**:**

Class interval	0-20	20-40	40-50	50-60	60-80
Frequency	4	6	10	8	2

Solution:

Computation of Sum of Values

Class interval	Frequency (f)	Mid-value (X)	$d' = \frac{X - 45}{5}$	fd'	$\|X - \bar{X}\|$	$f\|X - \bar{X}\|$
0-20	4	10	–7	–28	31.67	126.68
20-40	6	30	–3	–18	11.67	70.02
40-50	10	45	0	0	3.33	33.3
50-60	8	55	2	16	13.33	106.64
60-80	2	70	5	10	28.33	56.66
Total	N = 30			$\sum fd' = -20$		$\sum f\|X - \bar{X}\| = 393$

Now,

Mean ($\bar{X}$) = A + $\frac{\sum fd'}{N} \times h = 45 + \frac{(-20)}{30} \times 5 = 45 - \frac{10}{3} = 41.67$

$$\therefore \text{M.D. from Mean} = \frac{\sum f\left|X-\overline{X}\right|}{N} = \frac{393.3}{30} = 13.11$$

$$\text{Coefficient of M.D. from Mean} = \frac{\text{M.D. from Mean}}{\text{Mean}} = \frac{13.11}{41.67} = 0.3146$$

Merits of Mean Deviation

- It is rigidly defined.
- It is simple to understand and easy to calculate.
- Mean deviation is taken form median will be minimum. It means, it is more reliable than range and quartile deviation.
- It is less affected by extreme observations as compared with standard deviation.

Demerits of Mean Deviation

- It ignores the negative sign.
- It can not be calculated exactly for frequency distribution having open end classes.
- It does not give satisfactory result when mean deviation taken from mode , when mode is ill-defined.
- It is rarely used in sociological studies.

Standard Deviation

Standard deviation is said to be the best measure of dispersion (or ideal measure of dispersion) as it satisfies almost all the requisites (or characteristics) of an ideal or a good measure of dispersion. It was first suggested by Karl Pearson in 1893. It is usually denoted by Greek alphabet σ (sigma) and defined as "the positive square root of the arithmetic mean of the square of the deviations of the given set of observations from their arithmetic mean."

Merits of Standard deviation

- It is rigidly defined and based on all the observations.
- It is least affected by fluctuations of sampling as compared to other measures of dispersion.
- It is suitable for further mathematical treatment.
- It is mostly useful in sampling distributions, testing of hypothesis, skewness and in correlation.
- The standard deviation is most useful in judging the representative of mean i.e. the distribution with the smallest standard deviation is said to have the most representative mean.
- It is the most useful statistical tool for further statistical analysis.

Demerits of Standard Deviation

- It is not easy to calculate as compared to other measures of dispersion.
- It can not be calculated exactly for the frequency distributions with open-end classes.
- It gives more weightage to extreme values and less to those which are nearer to mean.

Variance

The square of the standard deviation is known as variance. It is denoted by σ^2 and given by

$\therefore \sigma^2 = V(X)$ Where V(X) = variance of variable x

$\Rightarrow \sigma = \sqrt{V(X)}$

It is clear that variance is always positive and measured in terms of square units of given data. The concept of variance is very much useful in an advanced statistical work and has very important applications in inferential statistics.

Coefficient of Standard Deviation

The standard deviation divided by their arithmetic mean is known as coefficient of standard deviation

Coefficient of S.D. = $\frac{\sigma}{\bar{X}}$

Coefficient of Variation (C.V.)

100 times the coefficient of standard deviation is called coefficient of variation. In other words, the coefficient of standard deviation expressed in percentage is known as coefficient of variation. Symbolically,

C.V. = $\frac{\sigma}{\bar{X}}$ × 100%

According to Karl Pearson, coefficient of variation is the "percentage variation in the mean". It is a relative measure of dispersion, so it is independent of units of measurement. It is always expressed in percentage. Therefore, C.V. can betterly be used to compare two or more than two distributions with regard to their variability, consistency, uniformity, homogeneity, equitability, stability etc.

Coefficient of variation (C.V.) is applicable for the comparison of variability of two or more than two distributions (series) as follows

Less C.V. is considered as	**More C.V. is considered as**
More consistent	Less consistent
More homogeneous	Less homogeneous
More uniform	Less uniform
More stable	Less stable
More representative to mean	Less representative to mean
More equitable	Less equitable
Less variable	More variable
Less disparity	More disparity

Note: As interpretation of C.V. is based on its magnitude only, therefore, if C.V. comes out to be negative, the negative value of C.V. will be interpreted taking its absolute value. i.e. $|C.V.|$

For Individual Series

(i) S.D. (σ) = $\sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$ **- direct method**

(ii) S.D. (σ) = $\sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$ **- Short cut method**

Where, d = X – a a = Assumed mean

Standard deviation (S.d.) = σ

n = number of observations

Calculate the standard deviation and variance from the following frequency distribution.

2 4 6 8 10 12

Solution:

X	X^2
2	4
4	16
6	36
8	64
10	100
12	144
$\sum X = 42$	$\sum X^2 = 364$

Standard deviation (σ) = $\sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$

= $\sqrt{\frac{364}{6} - \left(\frac{42}{6}\right)^2}$

$= \sqrt{60.666 - 49}$

$= 3.41565$

Variance $(\sigma^2) = (3.41565)^2 = 11.6666$

Find S.D of the following series by using

(a) Direct method

(b) Short cut method (or assumed mean method)

X: 4, 6, 8, 14, 18

Also, compute variance.

Solution:

(a) **Direct method**

Computation of sum of values

X	**X^2**
4	16
6	36
8	64
14	196
18	324
$\sum X = 50$	$\sum X^2 = 636$

S.D. $(\sigma) = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{636}{5} - \left(\frac{50}{2}\right)^2} = \sqrt{27.2} = 5.21$

OR

(b) Short cut method

Computation of sum of values

X	**d = X – 8**	**d^2**
4	–4	16
6	–2	4

8	0	0
14	6	36
18	10	100
	$\sum d = 10$	$\sum d^2 = 156$

S.D. $(\sigma) = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{156}{5} - \left(\frac{10}{5}\right)^2} = \sqrt{27.2} = 5.21.$

Also, variance (σ^2) = (S.D.)^2 = $(5.21)^2$ = 27.2.

[redacted]

The running capacity of two car is given below, state which is more consistent and why?

Car A	250	255	280	290	295	300
Car B	280	282	290	295	298	295

Solution:

To test the consistency, the coefficient of variation (C.V.) is compared.

Calculation of Coefficient of Variation (C.V.)

For Car A

X	d = X - 290	d^2
250	-40	1600
255	-35	1225
280	-10	100
290	0	0
295	5	25
300	10	100
	$\sum d$ = -70	$\sum d^2$ = 3050

$\bar{X}$ = a + $\frac{\sum d}{n}$ = 290 + $\frac{(-70)}{6}$ = 278.33

S.d. $(\sigma) = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{3050}{6} - \left(\frac{-70}{6}\right)^2}$

$= \sqrt{508.33 - 11.67} = \sqrt{372.22} = 19.29$

Coefficient of variation, C.V. (Car A) = $\frac{\sigma}{\bar{X}} \times 100 = \frac{19.29}{278.33} \times 100 = 6.93\%$

For Car B

Y	d = Y - 290	d^2
280	-10	100
282	-8	64
290	-0	0
295	5	25
298	8	64
295	2	4
	$\sum d = 0$	$\sum d^2 = 278$

$\bar{Y} = b + \frac{\sum d}{n} = 290 + \frac{(0)}{6} = 290$

S.d. $(\sigma) = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{278}{6} - \left(\frac{0}{6}\right)^2}$

$= \sqrt{46.33 - 0} = \sqrt{46.33} = 6.81$

Coefficient of variation, C.V. (Car B) $= \frac{\sigma}{\bar{Y}} \times 100 = \frac{6.81}{290} \times 100 = 2.35\%$

Since, C.V. (CarA) > C.V. (Car B). Hence, we conclude that car B is more consistent than car A.

[redacted]

Following are marks secured by Miss A and Miss B in 10 tests 100 marks each.

Test	1	2	3	4	5	6	7	8	9	10
Marks secured by A	48	52	54	60	62	64	68	68	72	74
Marks secured by B	44	48	55	58	63	66	67	70	73	80

If the consistency of performance is the criterion for awarding a prize. Who should be awarded by the prize?

Solution: For the consistency of performance, the coefficient of variation (C.V.) is compared.

i.e. to test the consistency, the coefficient of variation (C.V.) is compared.

Calculation of Coefficient of Variation (C.V.)

For Miss A

Marks (X)	X^2
48	2304
52	2704
54	2916
60	3600
62	3844
64	4096
68	4624

68	4624
72	5184
74	5476
$\sum X$ = 622	$\sum X^2$ = 39372

$\overline{X} = \frac{\sum X}{n} = \frac{622}{10} = 62.2$

S.d. $(\sigma) = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{39372}{10} - \left(\frac{622}{10}\right)^2}$

$= \sqrt{3937.2 - 3868.84} = \sqrt{68.36} = 8.268$

Coefficient of variation, C.V. (Miss A) $= \frac{\sigma}{\bar{X}} \times 100 = \frac{8.268}{62.2} \times 100 = 13.292\%$

For Miss B

Marks (X)	X^2
44	1936
48	2304
55	3025
58	3364
63	3969
66	4356
67	4489
70	4900
73	5329
80	6400
$\sum X$ = 624	$\sum X^2$ = 40072

$\overline{X} = \frac{\sum X}{n} = \frac{624}{10} = 62.4$

S.d. $(\sigma) = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2} = \sqrt{\frac{40072}{10} - \left(\frac{624}{10}\right)^2}$

$= \sqrt{4007.2 - 3893.76} = \sqrt{113.44} = 10.650$

Coefficient of variation, C.V. (Miss B) $= \frac{\sigma}{\bar{X}} \times 100 = \frac{10.650}{62.4} \times 100 = 17.067\%$

Since, C.V. (Miss A) < C.V.(Miss B). Hence, if the consistency of performance is the criteria for awarding a prize, Miss A should be awarded by the prize.

[redacted]

Following are the marks obtained by two students A and B in 10 test of 100 marks each.

Test 1	1	2	3	4	5	6	7	8	9	10
Marks of A	54	90	86	58	62	82	78	66	70	64
Marks of B	58	85	64	70	73	79	82	61	67	76

(a) Who is better?

(b) Who is intelligent?

(c) If the consistency of performance is the criteria for awarding a prize. Who should get the prize?

Solution:

Arrange the marks of A and B according to ascending order

For student A			For student B		
X	d = X – 70	d^2	X	d = X – 70	d^2
54	– 16	256	58	– 12	144
58	– 12	144	61	– 9	81
62	– 8	64	64	– 6	36
64	– 6	36	67	– 3	9
66	– 4	16	70	0	0
70	0	0	73	3	9
78	8	64	76	6	36
82	12	144	79	9	81
86	16	256	82	12	144
90	20	400	85	15	256
Total	10	1380		15	796

(a) $\overline{X}_A = A + \frac{\sum d}{n} = 70 + \frac{10}{10} = 71$ marks

$$\overline{X}_B = A + \frac{\sum d}{n} = 70 + \frac{15}{10} = 71.5 \text{ marks}$$

Since, $\overline{X}_A < \overline{X}_B$, therefore, student B is better than student A.

(b) For intelligen, median (Md) is calculated.

For student A

$$\text{Md (A)} = \text{value of } \left(\frac{n+1}{2}\right)^{th} item$$

$$= \text{value of } \left(\frac{10+1}{2}\right)^{th} item$$

$$= \text{value of } (5.5)^{th} \text{ item}$$

$$\therefore \quad M_d(A) = \frac{66 + 70}{2} = 68 \text{ marks}$$

Similarly,

For student B,

$$\text{Md (A)} = \text{value of } \left(\frac{n+1}{2}\right)^{th} item$$

$$= \text{value of } \left(\frac{10+1}{2}\right)^{th} item$$

$$= \text{value of } (5.5)^{th} \text{ item}$$

$$\text{Md (B)} = \frac{70 + 73}{4} = 71.5 \text{ marks}$$

Since, $M_d(A) < \text{Md (B)}$

Hence, B is more intelligent than A

(c) For the consistency of performance, the coefficient of variation (C.V.) is compared.

For student A

$$\text{S.D. } \sigma(A) = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2} = \sqrt{\frac{1380}{10} - \left(\frac{10}{10}\right)^2} = \sqrt{137} = 11.70.$$

$$\text{C.V. (A)} = \frac{\sigma(A)}{\overline{X}_A} \times 100 \ \% = \frac{11.70}{71} \times 100\% = 16.48\%$$

For student B

$$\text{S.D. } \sigma(B) = \sqrt{\frac{\sum d^2}{n} - \left(\frac{\sum d}{n}\right)^2}$$

$$= \sqrt{\frac{796}{10} - \left(\frac{10}{10}\right)^2} = \sqrt{77.35} = 8.79.$$

$$\text{C.V. (B)} = \frac{\sigma(B)}{\overline{X}_B} \times 100 \ \% = \frac{8.79}{71.5} \times 100\% = 12.29\%.$$

Conclusion: Since C.V.(student A) > CV(student B) . Therefore, if the consistency of performance is the criteria for awarding a prize, student B should get the prize.

For Discrete Series

Direct method

$$\text{S.D } (\sigma). = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$

Short cut method

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

Where, d = X – a

a = Assumed mean

N = $\sum f$ = Total frequency

[redacted]

Find standard deviation (S.D.) and variance of the following data.

Variable (X)	10	14	15	18	20
Frequency (f)	3	5	7	6	4

Solution:

Let a = 16

Value (X)	Frequency (f)	d = X-16	fd	fd^2
10	3	–6	–18	108
14	5	–2	–10	20
16	7	0	0	0
18	6	2	12	24
20	4	4	16	64
	N = $\sum f$ = 25		$\sum fd$ = 0	$\sum fd^2$ = 216

S.D. $(\sigma) = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$

$= \sqrt{\frac{216}{25} - \left(\frac{0}{25}\right)^2}$

$= \sqrt{\frac{216}{25} - 0}$

$\therefore \sigma = 2.939$

Variance (σ^2) = $(2.939)^2$ = 8.637

[redacted]

Students' age in the regular daytime BCA program and the morning program of Peoples campus are described by two samples. If the homogeneity in age of the class is a positive factor in learning make suggestion, with reason, which of the two groups will be easier to teach?

Regular BCA Program		Morning BCA Program	
Age	No. of Students	Age	No. of Students
23	9	27	10
29	2	31	8
28	5	30	5
22	10	29	4
30	1	28	6
21	4	33	5
25	11	34	5
26	6	35	11

27 24	3 9	36 32	2 4
Total	60	Total	60

Solution:

Now, we have to find the coefficient of variation (C.V.) for each program to compare the homogeneity.

For Regular BCA Program

Let a = 25

Age (X)	No. of Students (f)	d = X – 25	fd	fd^2
23	9	– 2	– 18	36
29	2	4	8	32
28	5	3	15	45
22	10	– 3	– 30	90
30	1	5	5	25
21	4	– 4	– 16	64
25	11	0	0	0
26	6	1	6	6
27	3	2	6	12
24	9	– 1	– 9	9
Total	$\sum f = N = 60$		$\sum fd = -33$	$\sum fd^2 = 319$

Now,

Mean $(\overline{X})$ = a + $\frac{\sum fd}{N}$ = 25 + $\frac{-33}{60}$ = 25 – 0.55 = 24.45.

S.D. (σ) = $\sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$ = $\sqrt{\frac{319}{60} - \left(\frac{-33}{60}\right)^2}$ = $\sqrt{5.317 - 0.303}$

= $\sqrt{5.014}$ = 2.239.

C.V. (Regular BCA program) = $\frac{\sigma}{\overline{X}}$ × 100 = $\frac{2.239}{24.45}$ × 100 = 9.16%.

For Morning BCA Program

Let a = 31

Age (X)	No. of Students (f)	d = X – 31	fd	fd^2
27	10	– 4	– 40	160
31	8	0	0	0
30	5	– 1	– 5	5
29	4	– 2	– 8	16
28	6	– 3	– 18	54
33	5	2	10	20
34	5	3	15	45
35	11	4	44	176
36	2	5	10	50
32	4	1	4	4
Total	$\sum f = N = 60$		$\sum fd = 12$	$\sum fd^2 = 530$

Now,

Mean $(\overline{X}) = a + \frac{\sum fd}{N} = 31 + \frac{12}{60} = 31 - 0.2 = 31.2.$

S.D. $(\sigma) = \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2} = \sqrt{\frac{530}{60} - \left(\frac{12}{60}\right)^2} = \sqrt{8.833 - 0.04}$

$= \sqrt{8.793} = 2.965.$

C.V. (Morning BCA program) $= \frac{\sigma}{\overline{X}} \times 100\% = \frac{2.965}{31.2} \times 100 = 9.50\%$

Since, C.V.(Regular BCA program) < C.V.(Morning BCA program). Hence, if the homogeneity in age of the class is a positive factor in learning we conclude that it is easier to teach in Regular BCA programme than morning BCA program.

For Continuous Series

Direct method

S.D (σ). $= \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$

Short cut method

S.D (σ). $= \sqrt{\frac{\sum fd^2}{N} - \left(\frac{\sum fd}{N}\right)^2}$

Where, $d = X - a$

a = Assumed mean

$N = \sum f =$ Total frequency

Step-deviation method

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

Where, $d' = \frac{X-a}{h}$

X = mid. value

a = assumed mean

h = class size (For unequal class size h is taken common factor)

Note:

(i) Formula for calculating s.d. (σ) is same in case of continuous series and discrete series but X is middle value of the class intervals in case of continuous series but X is the given value of variable for discrete series.

(ii) h is class size but for unequal class size h is taken as common factor for continuous series.

.

Calculate the standard deviation and coefficient of variation from following data

Income(Rs)	20-25	25-30	30-35	35-40	40-45	45-50
No.of persons	170	110	80	45	40	35

Solution:

Let a = 32.5

Income (in Rs.)	No. of persons (f)	Mid Value (X)	d' = $\frac{X-32.5}{5}$	fd'	fd'^{2}
20-25	170	22.5	–2	–340	680
25-30	110	27.5	–1	–110	110
30-35	80	32.5	0	0	0
35-40	45	37.5	1	45	45
40-45	40	42.5	2	80	160
45-50	35	47.5	3	105	315
	N= $\sum f$ = 480			$\sum fd'$ = –220	$\sum fd'^2$ = 1310

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

$= \sqrt{\frac{1310}{480} - \left(\frac{-220}{480}\right)^2} \times 5$

$= \sqrt{2.729 - 0.210} \times 5$

$= \sqrt{2.519} \times 5$

= Rs.7.935

For coefficient of variation (C.V.):

Mean $(\bar{X}) = a + \frac{\sum fd\prime}{N} \times h$

$= 32.5 \frac{(-220)}{480} \times 5$

$= 30.20833$

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

$= \frac{7.935}{30.208} \times 100$

$= 26.267\%$

.

Findout the SD and variance from the following distribution.

Wages (in Rs.)	0-20	20-40	40-60	60-80	80-100
No. of workers	2	3	4	3	2

Solution:

Wages (in Rs.)	No. of workers (f)	Mid Value (X)	$d' = \frac{X - 50}{20}$	fd'	fd'^2
0-20	10	10	–2	–20	40
20-40	12	30	–1	–12	12
40-60	15	50	0	0	0
60-80	8	70	1	8	8
80-100	5	90	2	10	20
	$N = \sum f = 50$			$\sum fd' = -14$	$\sum fd'^2 = 80$

S.D. $(\sigma) = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times h = \sqrt{\frac{80}{25} - \left(\frac{-14}{25}\right)^2} \times 20$

$= 20 \times \sqrt{3.2 - 3136} =$ Rs. 33.98.

Variance $(\sigma^2) = (33.98)^2 = 1154.64$

.

Calculate dispersion from the following data.

Marks (Less than)	10	20	30	40	50	60	70

Frequency	5	20	40	62	70	76	80

OR

Marks	Below 10	Below 20	Beow30	Below 40	below 50	Below 60	Below 70
Frequency	5	20	40	62	70	76	80

Solution:

Since, Standard deviation is an ideal measure of dispersion, so for dispersion, standard deviation (S.D.) is calculated.

Let a = 35

Marks below	Less than c.f.	Frequency (f)	Mid value (X)	$d' = \frac{X - 35}{10}$	fd'	fd'^2
0-10	5	5	5	–3	–15	45
10-20	20	20 – 5 = 15	15	–2	–30	60
20-30	40	40 – 20 = 20	25	–1	–20	20
30-40	62	62 – 40 = 22	35	0	0	0
40-50	70	70 – 62 = 8	45	1	8	8
50-60	76	76 – 70 = 6	55	2	12	24
60-70	80	80 – 76 = 4	65	3	12	36
		N = 80			$\sum fd' = -33$	$\sum fd'^2 = 193$

Now

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

$$= \sqrt{\frac{193}{80} - \left(\frac{-33}{80}\right)^2} \times 10$$

$$= 10 \times \sqrt{2.4125 - .1702} \times 10 = \sqrt{2.2423} \times 10$$

$$= 1.497 \times 10 = 14.97.$$

[redacted]

Calculate an ideal measure of dispersion (i.e. standard deviation) and mean of the following series:

Marks (More than)	0	10	20	30	40	50	60
Frequency	80	76	70	62	40	20	5

OR

Marks	Above 0	Above 10	Above 20	Above 30	Above 40	Above 50	Above 60
Frequency	80	76	70	62	40	20	5

Solution:

Since, almost all the requisites to be good measure of dispersion followed by the standard deviation (S.D.). So, standard deviation is called an ideal measure of dispersion

Let a = 35

Marks	More than c.f.	Frequency (f)	Mid value (X)	$d' = \frac{X - 35}{10}$	fd'	fd'^2
0-10	80	4	5	–3	–12	36
10-20	76	6	15	–2	–12	24
20-30	70	8	25	–1	–8	8
30-40	62	22	35	0	0	0
40-50	40	20	45	1	20	20
50-60	20	15	55	2	30	60
60-70	5	5	65	3	15	45
		N = 80			$\sum fd' = 33$	$\sum fd'^2 = 193$

Now,

$$\text{Mean } (\bar{X}) = A + \frac{\sum fd'}{N} \times h = 35 + \frac{(33)}{80} \times 10 = 35 + 4.125 = 39.825 \text{ marks}$$

$$\text{Standard deviation } (\sigma) = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times h = \sqrt{\frac{193}{80} - \left(\frac{33}{80}\right)^2} \times 10$$

$= \sqrt{2.4125 - .1702} \times 10 = \sqrt{2.2423} \times 10 = 1.497 \times 10 = 14.97$ marks

[redacted]

The following table shows the monthly expenditure of ward no.1 and ward no. 2 of Kathmandu Metropolitan City in certain locality.

Expenditure (in 000 Rs.)	0-5	5-10	10-15	15-20	20-25	25-30
No. of families (ward no.1)	5	12	50	20	10	3
No. of families (ward no.2)	7	15	40	18	12	8

(iii) Which ward of people has uniform expenditure?

Solution: The ward of people has more uniform expenditure whose Coefficient of variation (C.V.) is less**.** **For ward no.1**

Expenditure (in 000 Rs.)	No. of families (ward no.1) f	mid. value (x)	$d' = \frac{x-12.5}{5}$	fd'	fd'^2
0-5	5	2.5	-2	-10	20
5-10	12	7.5	-1	-12	12
10-15	50	12.5	0	0	0
15-20	20	17.5	1	20	20
20-25	10	22.5	2	20	40
25-30	3	27.5	3	9	27
	N = 100			$\sum fd' =$ 27	$\sum fd'^2 =$ 119

$\bar{X} = a + \frac{\sum fd'}{N} \times h = 12.5 + \frac{27}{100} \times 5 = 27.35$

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

$= \sqrt{\frac{119}{100} - \left(\frac{27}{100}\right)^2} \times 5 = 5.284648$

C.V. (Ward no. 1) $= \frac{\sigma}{\bar{X}} \times 100 = \frac{5.284648}{27.35} \times 100 = 19.3223\%$

For ward no.2

Expenditure (in 000 Rs.)	No. of families (ward no.1) f	mid. value (x)	$d' = \frac{x-12.5}{5}$	fd'	fd'^2
0-5	7	2.5	-2	-14	28
5-10	15	7.5	-1	-15	15
10-15	40	12.5	0	0	0
15-20	18	17.5	1	18	18
20-25	12	22.5	2	24	48
25-30	8	27.5	3	24	72
	N = 100			$\sum fd' =$ 37	$\sum fd'^2 =$ 181

$\bar{X} = a + \frac{\sum fd'}{N} \times h = 12.5 + \frac{37}{100} \times 5 = 14.35$

$\sigma = \sqrt{\frac{\sum fd'^2}{N} - \left(\frac{\sum fd'}{N}\right)^2} \times h = \sqrt{\frac{181}{100} - \left(\frac{37}{100}\right)^2} \times 5 = 6.467418$

C.V.(Ward no. 2) $= \frac{\sigma}{\bar{X}} \times 100 = \frac{6.467418}{14.35} \times 100 = 45.06911\%$

Since, C.V.(Ward no. 1) < C.V.(Ward no. 2). Therefore, people of ward 1 has more uniform expenditure than ward no. 2.

The following table gives the two bike models and their corresponding life:

Life (in years)		0-2	2-4	4-6	6-8	8-10
No. of bikes	Model T_1	1	9	12	11	8
	Model T_2	5	7	11	19	9

Which model of bike has greater uniformity?

Solution:

We have to compute coefficient of variation (C.V.) to determine the uniformity.

Computation of Sum of Values for Mean and S.D

. . **For Model T_1**

Life (in years)	No. of bikes (f)	mid. value (x)	$d' = \frac{x-5}{2}$	fd'	fd'^2
0-2	1	1	-2	-2	4
2-4	9	3	-1	-9	9
4-6	12	5	0	0	0
6-8	11	7	1	11	11
8-10	8	9	2	16	32
	N = $\sum f$ = 41			$\sum fd' = 16$	$\sum fd'^2 = 56$

$\bar{X}$ = a + $\frac{\sum fd'}{N} \times h$ = 5+ $\frac{16}{41} \times 2 = 5 + 0.78 = 5.78$ years

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

$= \sqrt{\frac{56}{41} - \left(\frac{16}{41}\right)^2} \times 2 = 1.101 \times 2 = 2.202$ years

C.V. (Model T_1) = $\frac{\sigma}{\bar{X}} \times 100 = \frac{2.202}{5.78} \times 100 = 38.09\%$

. **For Model T_2**

Life (in years)	No. of bikes (f)	mid. value (x)	$d' = \frac{x-5}{2}$	fd'	fd'^2
0-2	5	1	-2	-10	20

2-4 4-6 6-8 8-10	7 11 19 9	3 5 7 9	-1 0 1 2	-7 0 19 18	7 0 19 36
	N = $\sum$ f = 51			$\sum fd' = 20$	$\sum fd'^2 = 82$

$$\bar{X} = a + \frac{\sum fd'}{N} \times h = 5 + \frac{20}{51} \times 2 = 5 + 0.7843 = 5.7843 \text{ years}$$

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

$$= \sqrt{\frac{82}{51} - \left(\frac{20}{51}\right)^2} \times 2 = 1.205842 \times 2 = 2.4116 \text{ years}$$

C.V. (Model T_2) = $\frac{\sigma}{\bar{X}} \times 100 = \frac{2.4116}{5.7843} \times 100 = 41.692\%$

Since, C.V. (Model T_1) < C.V. (Model T_2). Therefore, model T_1 of bike has greater uniformity than model T_2

To Find the correct mean and correct standard deviation

[redacted]

For a group of 200 candidates, the mean and standard deviation were found to be 40 and 15. Later on it was discovered that the score 53 was misread as 35. Find the correct mean and standard deviation corresponding to the correct figures.

Solution:

We have given,

n = 200

Mean ($\bar{X}$) = 40

Standard deviation (σ) = 15

Wrong observation (i.e. wrong score) = 35

Corrected observation (i.e.correct score) = 53

Corrected Mean ($\bar{X}$correct) =?

Corrected standard deviation($\sigma_{correct}$) =?

We know,

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

or, $40 = \frac{\sum X}{200}$

or, $\sum X = 200 \times 40$

or, $\sum X = 8000$

Corrected $\sum X = \sum X$ – Wrong observation + Correct observation

$= 8000 - 35 + 53 = 8018$

$\therefore$ Correct mean$(\overline{X}_{correct}) = \frac{\text{Correct } \sum X}{n} = \frac{8018}{200} = 40.09$

Again,

S.D. $(\sigma) = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$

or, $15 = \sqrt{\frac{\sum X^2}{200} - \left(\frac{\sum X}{n}\right)^2}$

or, $15 = \sqrt{\frac{\sum X^2}{200} - (\overline{X})^2}$ $\quad \because \overline{X} = \frac{\sum X}{n}$

or, $15 = \sqrt{\frac{\sum X^2}{200} - (40)^2}$

Squaring both sides

$225 = \frac{\sum X^2}{200} - (40)^2$

or, $225 = \frac{\sum X^2}{200} - 1600$

or, $225 + 1600 = \frac{\sum X^2}{200}$

or, $1825 = \frac{\sum X^2}{200}$

or, $\sum X^2 = 1825 \times 200 = 365000$

$\therefore$ Corrected $\sum X^2 = \sum X^2$ – (Wrong observation)^2 + (Correct observation)^2

$= 365000 - (35)^2 + (53)^2 = 366584$

$\therefore$ Corrected S.D. $(\sigma_{Correct}) = \sqrt{\frac{\text{Corrected} \sum X^2}{n} - \left(\frac{\text{Corrected } \sum X}{n}\right)^2}$

$= \sqrt{\frac{366584}{200} - \left(\frac{8018}{200}\right)^2}$

$= 15.02$

[redacted]

The mean and standard deviation of a set of 100 workers were found to be 40 and 12 respectively. On checking, it was found that two workers were wrongly taken as 23 and 15 instead of 43 and 18. Calculate the correct mean and standard deviation. Also, find correct variance.

Solution:

We have given,

Total no. of observations (n) = 100

Mean ($\bar{X}$) = 40

Standard deviation (σ) = 12

Wrong observations = 23 and 15

Correct observations = 43 and 18

We know,

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

or, $40 = \frac{\sum X}{100}$

or, $\sum X = 100 \times 40$

or, $\sum X = 4000$

Corrected $\sum X$ = $\sum X$ – Wrong observations + Correct observations

= 8000 – 23–15 + 43+18 = 4023

$\therefore$ Correct mean$(\bar{X}_{correct}) = \frac{Correct\ \sum X}{n} = \frac{4023}{100} = 40.23$

Again,

$$\text{S.D. } (\sigma) = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

or, $12 = \sqrt{\frac{\sum X^2}{100} - \left(\frac{\sum X}{n}\right)^2}$

or, $12 = \sqrt{\frac{\sum X^2}{100} - (\bar{X})^2}$ $\quad \because \bar{X} = \frac{\sum X}{n}$

or, $12 = \sqrt{\frac{\sum X^2}{100} - (40)^2}$

Squaring both sides

$144 = \frac{\sum X^2}{100} - (40)^2$

or, $144 = \frac{\sum X^2}{100} - 1600$

or, $144 + 1600 = \frac{\sum X^2}{100}$

or, $1744 = \frac{\sum X^2}{100}$

or, $\sum X^2 = 1744 \times 100 = 174400$

$\therefore$ Corrected $\sum X^2 = \sum X^2 - (\text{Wrong observations})^2 + (\text{Correct observations})^2$

$= 174400 - (23)^2 - (15)^2 + (43)^2 + (18)^2$

$= 174400 - 529 - 225 + 1849 + 324$

$= 175819$

$\therefore$ Corrected S.D. $(\sigma_{Correct}) = \sqrt{\frac{\text{Corrected}\sum X^2}{n} - \left(\frac{\text{Corrected } \sum X}{n}\right)^2}$

$= \sqrt{\frac{175819}{100} - \left(\frac{4023}{100}\right)^2}$

$= 11.82$

Correct variance $(\sigma^2_{\text{correct}}) = (\sigma_{\text{correct}})^2 = (11.82)^2 = 139.737$

$\therefore$ Corrected mean $(\bar{X}) = 40.23$

Corrected standard = 11.82

& Correct variance $(\sigma^2_{\text{correct}}) = 139.737$

Combined standard deviation

Let $\overline{X}_1$ and $\overline{X}_2$ be the means of first and second series, σ_1^2 and σ_2^2 be the variances of first and second series with n_1 and n_2 number of observations respectively. Then the combined standard deviation of these two series taken together is given by

$$\sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1 + n_2}}$$ **Where,** $d_1 = \overline{X}_1 - \overline{X}_{12}$

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

$$\overline{X}_{12} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2}$$

Similarly, the combined standard for three series is given by

$$\sigma_{123} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_3\sigma_3^2 + n_1d_1^2 + n_2d_2^2 + n_3d_3^2}{n_1 + n_2 + n_3}}$$

Where, $d_1 = \overline{X}_1 - \overline{X}_{123}$

$$d_2 = \overline{X}_2 - \overline{X}_{123}$$

$$d_3 = \overline{X}_3 - \overline{X}_{123}$$

$$\overline{X}_{123} = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + n_3\overline{X}_3}{n_1 + n_2 + n_3}$$

[redacted]

A sample size 15 has mean 3.5 and SD 3.0. Another sample of size 22 has mean 4.7 and SD 4.0. If the two samples are pooled together, find the mean and the SD of the combined sample**.**

Solution:

$n_1 = 15$ $\quad \bar{X}_1 = 3.5,$ $\quad \sigma_1 = 3$

$n_2 = 22$ $\quad \bar{X}_2 = 4.7,$ $\quad \sigma_2 = 4$

Combined mean, $\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{15 \times 3.5 + 22 \times 47}{15 + 22} = \frac{52.5 + 103.4}{37} = 4.2135$

Combined standard deviation $\sigma_{12} = \sqrt{\frac{n_1\sigma_1^{\,2} + n_2\sigma_2^{\,2} + n_1d_1^{\,2} + n_2d_2^{\,2}}{n_1 + n_2}}$ . . . . . (i)

Where, $d_1 = (\bar{X}_1 - \bar{X}_{12}) = (3.5 - 4.2135) = -0.7135$

$d_2 = (\bar{X}_2 - \bar{X}_{12}) = (4.7 - 4.2135) = 0.4865$

From (i)

$\therefore \sigma_{12} = \sqrt{\frac{135 + 352 + 7.636 + 5.207}{37}} = \sqrt{\frac{499.943}{37}}$

$= \sqrt{13.512} = 3.676$

[redacted]

A factory produces two types of CFL bulbs A and B . The following results were obtained relating to their life

	Bulb A	Bulb B
No. of bulbs	100	90
Average length of life	900 hours	1000 hours
Variance	121	144

a) Compare the variability of life of two types of CFL bulbs.
b) Calculate the standard deviation of both types of CFL bulbs taken together.

Solution:

a) For Bulb A For Bulb B

$n_1 = 100$ $n_2 = 90$

$\bar{X}_1 = 900$ hours $\bar{X}_2 = 1000$ hours

$\sigma_1 = \sqrt{121} = 11$ $\sigma_2 = \sqrt{144} = 12$

C.V.$(Bulb\ A) = \frac{\sigma_1}{\bar{X}_1} \times 100$ C.V. $(Bulb\ B) = \frac{\sigma_2}{\bar{X}_2} \times 100$

$= \frac{11}{900} \times 100$ $= \frac{12}{1000} \times 100$

$= 1.222\%$ $= 1.2\%$

Since, C.V. (Bulb A) > C.V.(Bulb B). Therefore, the life of type of Bulb A is more variability than type of Bulb B. That is, the life of type of Bulb B is more consistent than type of Bulb A.

b) The standard deviation of both types of CFL bulbs taken together (i.e. combined standard deviation) is

$$\sigma_{12} = \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1+n_2}}$$

$$= \sqrt{\frac{n_1\sigma_1^2 + n_2\sigma_2^2 + n_1d_1^2 + n_2d_2^2}{n_1+n_2}}$$

$$= \sqrt{\frac{100\times121 + 90\times144 + 100(-47.368)^2 + 90\times(52.631)^2}{100+90}}$$

$= 3.716903$

Where, $\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2} = \frac{100\times900 + 90\times1000}{100+90} = 947.3684$

$d_1 = \bar{X}_1 - \bar{X}_{12} = 900 - 947.3684 = -47.3684$

$d_2 = \bar{X}_2 - \bar{X}_{12} = 1000 - 947.3684 = 52.631$

.

Measures of Central Tendency

Introduction

The averages are the measures which condense a huge unwieldy set of numerical data into single numerical values which are representative of the entire distribution.

Averages are the typical values and they lie between the two extreme observations (i.e. the largest and smallest observations) of the distribution which represents entire data set and give us an idea about the concentration of the values in the central part of the distribution. Accordingly, measures of such single values are also known as the 'Measures of Central Tendency"

The objectives of central tendency are

- To get single value that represents the entire data.
- To facilitate comparison.
- To present the salient feature of a mass of complex data.
- To help for computing various other statistical measures such as dispersion, skewness. kurtosis, correlation, regression and various other basic characteristics of a mass of data.
- To trace mathematical relation.
- To help in decision making.

Requisites of a Good Average or Characteristics of an ideal measure of central tendency

Since, central value of a mass of data is a proxy value, it should have some theoretical consideration, which makes the central value more reliable and representative.

According to **Prof. Yule**, following are the requirements for ideal measures of central tendency.

- It should be rigidly defined
- It should be easy to calculate and simple to understand.
- It should be based on all the observations:
- It should be suitable for further mathematical treatment.
- It should be least affected by the extreme observations.
- It should be least affected by fluctuations of sampling.

Various Measures of Central Tendency

The measures of central tendency or measures of location are designed to measure the central value around which most of the values in the data tend to concentrate.

The following are the measures of central tendency or measures of location:

- Arithmetic mean
  (i) Simple Arithmetic Mean$(\overline{X})$
  (ii) Weighted Arithmetic Mean $(\overline{X}_w)$
- Median (Md)
- Mode(Mo)
- Geometric mean$(G.M.)$ and
- Harmonic mean$(H.M.)$

Mean, median and mode are generally considered as major measures while G.M. and H.M. are considered as minor measures and used in special cases.

Note: G.M. and H.M. are beyond of our syllabus.

Arithmetic Mean or Average (A.M.)

The arithmetic mean is the most popular and widely used measure of central tendency. It is also called simply 'the mean' or 'the average'. It is also considered as an "Ideal measure of central tendency" or the best-known measures of central tendency because it satisfies almost all requisites of ideal measure of central tendency given by Prof. Yule.

In the study of social, economical or commercial problems such as production, income, prices, imports, exports etc. to calculate average, A.M. is used.

Arithmetic mean may either be

(i) Simple arithmetic mean or (ii)Weighted arithmetic mean

Simple arithmetic mean

In case of simple arithmetic mean, all the items in the distribution are equally important. It is denoted by $\bar{X}$ (X bar)

Calculation of Arithmetic mean:

Individual Series

(i) Direct method

Mean, $\overline{X} = \frac{\sum X}{n}$

Where, $\sum X$ = the sum of observations

n = the number of observations.

(ii) **Short-cut method or assumed mean method or change of origin method**

$$\overline{X} = a + \frac{\sum d}{n}$$

Where, a = Assumed mean

d = X – a = Deviations of items from the assumed mean

n = Number of items or observations

[redacted]

The following are the daily incomes of five persons in a certain locality.

Persons	A	B	C	D	E
Income (in Rs.)	400	350	450	500	300

Calculate the arithmetic mean.

Solution:

Persons	Incomes(in Rs) (X)
A	400
B	350
C	450

D	500
E	300
Total	$\sum X$= 2000

Arithmetic mean, $\bar{X} = \frac{\sum X}{n}$

$= \frac{2000}{5} = 400$

Hence, the average income is Rs.400

Calculate the average wage by using step deviation method from the following data

Wage ('00' Rs.): 50, 55, 60, 65, 70, and 75

Solution

Taking A= 60

Wage ('00' Rs.) X
50
55
60
65
70
75
$\sum X$ = 375

Here, n = 6

Average wage, $\bar{X} = \frac{\sum X}{n}$

$= \frac{375}{6}$ = Rs. 62.5 (in 00)

= Rs.6250

Hence, average wage is Rs. 6250

A Nepalese Child-care centre at community level is eligible for social services as long as the average age of its children stays below 9. If these data represents the ages of all the children currently attending Child-care Centre. Do they qualify for the grant?

8, 5, 9, 10, 9, 12, 7, 12, 13, 7, 8

Solution:

Age (X): 8, 5, 9, 10, 9, 12, 7, 12, 13, 7, 8

n = 11

$\sum X = 8+5+9+10+ \ldots\ldots\ldots +7+8 = 100$

Average age, $\overline{X} = \frac{\sum X}{n}$

$= \frac{100}{11} = 9.09$

Since, average age, $\overline{X} = 9.09$ which is more than 9, therefore they don't qualify for the grant.

The mean weight of a student in a certain group of students is 119 *lbs.* The individual weights of five of them are 115, 109, 129,117 and 114. What is the weight of the sixth student?

Solution:

Let the weight of sixth student be X_1 lbs.

n = 6

S.No. of student	Weight (in lbs.) X
1	115
2	109
3	129
4	117
5	114
6	X_1
	$\sum X = 583+X_1$

Here, $\overline{X}$ = 119 lbs.

$\overline{X} = \frac{\sum X}{n}$

or, $119 = \frac{584+X_1}{6}$

or, $714 = 583 + X_1$

or, $714 - 583 = X_1$

$\therefore X_1 = 130$ lbs

Discrete Series

(i) Direct method

Mean, $\overline{X} = \frac{\sum fX}{N}$

Where N = Σf = Total frequency

X = value of given variable.

(ii) Short-cut method or assumed mean method or coding method or change of origin method

Mean, $\bar{X} = a + \frac{\sum fd}{N}$

Where, a = Assumed mean

d = X - a = Deviation of the items from the assumed mean

N = Total frequency

[redacted]

Calculate arithmetic mean for the following frequency distribution**.**

X	5	10	15	20	25
f	2	4	7	20	25

Solution:

Calculation of mean

X	f	fX
5	2	10
10	4	40
15	7	105
20	3	60
25	1	25
	N = 17	ΣfX = 240

Now, arithmetic mean, $\bar{X} = \frac{\sum fX}{N}$

$= \frac{240}{17}$

= 14.12

Hence, arithmetic mean, $\overline{X}$ = 14.12

[redacted]

Determine the average mark by changing the origin of data (i.e. short-cut method) from the data given below**.**

Marks	10	20	30	40	50	60	70
No. of students	20	15	12	10	4	8	6

Solution

Calculation of average mark

Let a = 40

Marks	d = X – 40	f	fd
10	– 30	20	– 600
20	– 20	15	– 300
30	– 10	12	– 120
40	0	10	0
50	10	4	40
60	20	8	160
70	30	6	180
		N = 75	Σfd = – 640

Here, average mark $(\bar{X})$ = a + $\frac{\Sigma fd}{N}$ = 40 + $\frac{(-640)}{75}$ = 40 – 8.53 = 31.47

Hence, average mark = 31.47

[redacted]

Given that A.M. is equal to 7.3, find the missing frequency from the following data:

X	5	6	7	8	9
f	4	6	12	?	8

Solution:

Let f_1 be the missing frequency.

X	f	fX
5	4	20
6	6	36
7	12	84
8	f_1	$8f_1$
9	8	72
	N = 30 + f_1	$\sum$ fX = 212+ $8f_1$

Given, A.M., $\bar{X}$ = 7.3

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

or, 7.3 = $\frac{212+8f_1}{30+f_1}$

or, 212 +$8f_1$ = 219 + $7.3f_1$

or, $8f_1$- $7.3f_1$ = 219 -212

or, $0.7f_1$ = 7

or, $f_1 = \frac{7}{0.7}$

$\therefore f_1 = 10$

[redacted]

From the following data, find the missing frequency when mean is 15.38

Size	10	12	14	16	18	20
frequency	3	7	?	20	8	5

Solution:

Let f_1 be the missing frequency corresponding to size 14

Size (X)	Frequency(f)	fX
10	3	30
12	7	84
14	f_1	14 f_1
16	20	320
18	8	144
20	5	100
	N = 43+ f_1	$\sum fX = 678+14\ f_1$

Solution: Given, $\overline{X} = 15.38$

$$\overline{X} = \frac{\sum fX}{N}$$

or, $15.38 = \frac{678+14f_1}{43+f_1}$

or,$678 +14\ f_1 = 661.34+15.38\ f_1$

or, $678-661.34 =15.38\ f_1-14\ f_1$

or, $16.66 = 1.38\ f_1$

or, $f_1 = \frac{16.66}{1.38}$

$\therefore f_1 = 12$

[redacted]

For a certain frequency table which has only been partly reproduced here, the mean was found to be 1.46

No. of accidents	0	1	2	3	4	5	total
No. of days	46	?	?	25	10	5	200

Solution:

Let f_1 & f_2 be the missing frequencies corresponding to number of accidents X =1 and X = 2.

No. of accidents (X)	Frequency (f)	fX
0 1 2 3 4 5	46 f_1 f_2 25 10 5	0 f_1 $2f_2$ 75 40 25
	$N = 86+ f_1+ f_2$	$\sum fX = 140+f_1 + 2f_2$

Here, N = 200

$N = 86+ f_1+ f_2 = 200$

or, $f_1+ f_2 = 200-86$

$\Rightarrow f_2 = 114 - f_1$. . . . . (i)

Also, $\bar{X} = 1.46$

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

or, $1.46 = \frac{140+f_1+2f_2}{200}$

or, $140 + f_1 + 2f_2 = 1.46 \times 200$

or, $f_1 + 2f_2 = 292-140$

or, $f_1 + 2(114 - f_1) = 152$ (from i)

or, $f_1+228 - 2f_1 = 152$

or, $-f_1 = 152-228$

or, $-f_1 = - 76$

or, $f_1 = 76$

Putting the value of f_1 in equation (i)

$f_2 = 114 - f_1$

$= 114-76$

$= 38$

$\therefore$ $f_1 = 76$ & $f_2 = 38$

Continuous Series (Grouped Data)

(i) Direct method

Mean, $\overline{X} = \frac{\sum fX}{N}$

Where, X = midpoint of the class interval.

$= \frac{\text{Lpwer limit+Upper limit}}{2}$

N = Total frequency

Step-deviation method or change of origin and scale method or coding method

The formula for calculating A.M. by this method is given by

Mean $(\bar{X})$ = $a + \frac{\sum fd'}{N} \times h$

Where, $d' = \frac{X-a}{h}$,

a = Assumed mean

h = Class size or class width

X = mid points of class intervals.

N = Σf = Total frequency

Note: For unequal class size h is taken as common factor.

The formula for calculating arithmetic mean (A.M.) in continuous series is same as discrete series but in continuous series X is the mid value of class intervals.

[redacted]

The following is the monthly income distribution of the persons:

Income (000Rs.)	10-20	20-30	30-40	40-50	50-60
No. of persons	3	7	5	4	2

Find average income.

Solution:

Income (000 Rs.)	N0. of persons (f)	Mid.value (X)	f X
10-20	3	15	45
20-30	7	25	175
30-40	5	35	175
40-50	4	45	180
50-60	2	55	110
	N = ∑f = 21		∑fX = 685

Average income $(\bar{X}) = \frac{\Sigma fX}{N}$

$= \frac{685}{21}$

= 32.619(000 Rs.)

= Rs. 32619

[redacted]

Data below represents the wages (Rs.) received by workers in a construction site. Calculate A.M.

Wages	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70
No. of recipients	10	13	18	21	24	28	20	11	8

Solution:

Calculation of average wages using changing origin and scale of data

Let a = 47.5

Wages (Rs.)	Mid. value (X)	d' = $\frac{X-47.5}{5}$	f	fd'
25-30	27.5	– 4	10	– 40
30-35	32.5	– 3	13	– 39
35-40	37.5	– 2	18	– 36
40-45	42.5	– 1	21	– 21
45-50	47.5	0	24	0
50-55	52.5	1	28	28
55-60	57.5	2	20	40
60-65	62.5	3	11	33
65-70	67.5	4	8	32
			N = Σf = 153	Σfd' = – 3

We have,

A.M. $(\bar{X})$ = a + $\frac{\Sigma fd'}{N} \times h$

$= 47.5 + \frac{(-3)}{153} \times 5 = 47.5 - 0.098 = 47.40$

Hence, mean wages is Rs. 47.40

Calculation of A.M. or Average in case of Open end classes

Calculate average income from the given income distribution of 1400 workers of a factory.

Income in Rs.	Below 1500	1500-2000	2000-2500	2500-3000	3000-3500	3500-4000	4000 or above
No. of workers	200	225	275	250	180	150	120

Solution

Here, the given distribution has the open-end classes. Therefore, for calculating average income it is necessary to estimate the lower limit of the first class and upper limit of the last class. The limit is estimated according to the width of the second class and class preceding the last class.

Calculation of Average Income

Income in Rs.	No. of workers (f)	Mid-value (x)	$d' = \frac{X - 2759}{500}$	fd'
1000-1500	200	1250	– 3	– 600
1500-2000	225	1750	– 2	– 450
2000-2500	275	2250	– 1	– 275
2500-3000	250	2750	0	0
3000-3500	180	3250	1	180
3500-4000	150	3750	2	300
4000-4500	120	4250	3	360
	N = Σf = 1400			Σfd' = 485

A = 2759, h =500

$$\therefore \overline{X} = a + \frac{\Sigma fd'}{N} \times h = 2750 - \frac{485}{1400} \times 500$$

$$= 2750 - 173.21 = 2576.79$$

Hence, the average income is Rs. 2576.79.

Calculation of A.M. in case of cumulative frequency distribution (Less than and more than cumulative frequency distribution):

[redacted]

The following data shows the life time in in hours of 400 tube lights. Find the mean life time of the data.

Life time in hrs.	Less than 300	Less than 400	Less than 500	Less than 600	Less than 700	Less than 800	Less than 900	Less than 1000	Less than 1100	Less than 1200
No.of tube lights	0	20	60	116	194	265	324	374	392	400

OR

Life time in hours (Below)	300	400	500	600	700	800	900	1000	1100	1200
No. of tube lights	0	20	60	116	194	265	324	374	392	400

Solution:

Since, the given frequency distribution is in the type of less than cumulative frequency distribution. So, it should be converted into simple frequency distribution.

Life time (in hrs.)	No. of tubes (f)	Mid.value (X)	$d' = \frac{X - 750}{100}$	fd′
300-400	20	350	-4	-80
400-500	40	450	-3	-120
500-600	56	550	-2	-112
600-700	78	650	-1	-78
700-800	71	750	0	0
800-900	59	850	1	59
900-1000	50	950	2	100
1000-1100	18	1050	3	54
1100-1200	8	1150	4	32
	N = $\sum$ f =400			$\sum$ fd′ = -145

Here, a = 750, h= 100

A.M. $(\overline{X}) = a + \frac{\sum fd'}{N} \times h$

$= 750 + \frac{(-154)}{400} \times 100$

$= 713.75$

$\therefore \overline{X} = 713.75$ hours

The following table represents the marks of 100 students.

Marks (More than)	20	30	40	50	60	70	80	90
No.of students	100	95	80	60	55	50	30	15

OR

Marks	Above 20	Above 30	Above 40	Above 50	Above 60	Above 70	Above 80	Above 90
No.of students	100	95	80	60	55	50	30	15

Find the mean marks of all 100 students.

Solution:

Since, the given frequency distribution is in the type of more than cumulative frequency distribution. So, it should be converted into simple frequency distribution.

Marks	No.of students (f)	Mid.value (X)	$d' = \frac{X - 55}{10}$	fd'
20-30	5	25	-3	-15
30-40	15	35	-2	-30
40-50	20	45	-1	-20
50-60	5	55	0	0
60-70	5	65	1	40
70-80	20	75	2	45
80-90	15	85	3	60
90-100	15	95	4	
	N =$\sum$ f =100			$\sum fd' = 85$

Here, a = 55, h = 10

Mean $(\overline{X}) = a + \frac{\sum fd'}{N} \times h$

$= 55 + \frac{(85)}{100} \times 10$

$= 55 + 8.5$

$= 63.5$

∴ mean marks is 63.5

[redacted]

If the arithmetic mean of the following data is 67.45, find the missing frequency.

Marks	60-62	63-65	66-68	69-71	72-74
No. of Students	15	54	-	81	24

Solution:

Let f_1 be the missing frequency corresponding to class 66-68

Marks	No. of students (f)	Mid.value (X)	fX
60-62	15	61	915
63-65	54	64	3456
66-68	f_1	67	67 f_1
69-71	81	70	5670
72-74	24	73	1752
	N =$\sum f = f_1+174$		$\sum fX = 67f_1 + 11793$

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

or, $67.4 = \frac{67f_1 + 11793}{f_1+174}$

or, $67.45f_1+11736.3 = 67f_1 + 11793$

or, $67.45f_1-67\ f_1=11793-11736.3$

or, $0.45\ f_1= 56.7$

or, $f_1 = \frac{56.7}{0.45}$

∴ $f_1 =126$

[redacted]

The following table represents the marks of 50 students.

Marks	0-10	10-20	20-30	30-40	40-50	Total

No. of students	4	f_1	10	f_2	10	50

If the average marks = 30.2 Find missing frequencies.

Solution

Let the number of students securing marks between 10-20 and 30-40 be f_1 and f_2 respectively.

Calculation of missing frequencies

Marks	No. of students (f)	Mid.value (X)	fX
0-10	4	5	20
10-20	f_1	15	$15f_1$
20-30	10	25	250
30-40	f_2	35	$35f_2$
40-50	10	45	450
	$N=\sum f = 24 + f_1 + f_2$		$\Sigma fX = 15f_1 + 35f_2 + 720$

By given condition

$N = 50$

or, $24 + f_1 + f_2 = 50$

or, $f_1 + f_2 = 50-24$

or, $f_2 = 26 - f_1$ . . . . . (i)

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

$$\Rightarrow 30.2 = \frac{15f_1 + 35f_2 + 720}{50}$$

or, $1510 = 15f_1 + 35f_2 + 720$

or, $15f_1 + 35f_2 = 1510 - 720 = 790$

or, $3f_1 + 7f_2 = 158$

or, $3f_1 + 7(26 - f_1) = 158$ (from equation (i))

or, $3f_1 + 182 - 7f_1 = 158$

or, $-4f_1 = 158 - 182$

or, $-4f_1 = -24$

or, $f_1 = 6$

Putting the value of f_1 in equation (i)

$f_2 = 26 - f_1$

$= 26 - 6$

$= 20$

$\therefore$ $f_1 = 6$ & $f_2 = 20$

Hence, the required no. of students securing marks between 10-20 and 30-40 are respectively 6 and 20.

Note: (i) A.M. is more suitable average than others while we are dealing with quantitative measures such as average bonus, average income, average sales, average profit, average production, average height, and average expenditure, average revenue etc.

(i) For A.M., it is not necessary to convert unequal class intervals into equal class intervals. But in case of unequal class intervals, h is taken as common factor from the mid-value of classes.

(ii) It is also not necessary to convert inclusive class interval into exclusive class intervals because mid points remain same whether inclusive or exclusive.

Weighted Arithmetic Mean

While calculating simple arithmetic mean, it is based on the assumption that all the items in the distribution are equally important. But in practice, this may not be so. The relative importance of some items in a distribution are more important than others. So, when the weights are assigned for individual items with their relative importance or priorities (or weights), then the arithmetic mean calculated with respect to their priorities is called weighted arithmetic mean.

Then weighted arithmetic mean, usually denoted by $\bar{X}_w$ and is given by

$$\bar{X}_w = \frac{\sum WX}{\sum W}$$

Where, X = Value of variable in rate (per)

W = given weight or proportion or frequency

An enquiry into the budgets of middle class families in a family gave the following information.

	Food	Rent	Clothing	Fuel	Others
Expenses on	10%	15%	20%	15%	25%
Index no.	90	100	85	65	137

Compute weighted A.M.

Solution: We have, $\bar{X}_w = \frac{\sum WX}{\sum W}$

Calculation of weighted average

Group	Weights (W)	Index No. (x)	WX
Food	10	90	900
Rent	15	100	1500
Clothing	20	85	1700
Fuel	15	65	975
Others	25	137	3425
	ΣW = 85		ΣWX = 8500

∴ Weighted mean ($\overline{X}_w$) = $= \frac{8500}{85} = 100$

An examination was held to decide about the award of a scholarship in Ministry of Education. The marks obtained by three candidates and their respective weights of various subjects are given below:

Weight	Marks			
	Weight	A	B	C
Physics	4	70	80	85
Chemistry	3	90	75	75
Biology	1	50	60	45
Mathematics	2	60	45	65

If the candidate getting the highest marks is to be awarded the scholarship who should get it.

Solution:

Subject	Weight W	Candidate A		Candidate B		Candidate C	
		X_1	WX_1	X_2	WX_2	X_3	WX_3
Physics	4	70	280	45	180	85	340
Chemistry	3	60	180	75	225	75	225
Biology	1	50	50	60	60	45	45
Mathematics	2	90	180	80	160	65	130

	$\sum W = 10$		$\sum WX_1 = 690$		$\sum WX_2 = 625$		$\sum WX_3 = 740$

$$\bar{X}_w(A) = \frac{\sum WX_1}{\sum W}$$

$$= \frac{690}{10}$$

$$= 69$$

$$\bar{X}_w(B) = \frac{\sum WX_2}{\sum W}$$

$$= \frac{625}{10}$$

$$= 62.5$$

$$\bar{X}_w(C) = \frac{\sum W3}{\sum W}$$

$$= \frac{740}{10}$$

$$= 74$$

Since, the weighted mean of student C is the highest. So, C should be awarded the scholarship

Combined mean or Mean of combined Series

Let $\bar{X}_1$ and $\bar{X}_2$ be the arithmetic means of two series of sizes n_1 and n_2 respectively. Then the combined mean of two series of size ($n_1 + n_2$) is denoted by $\bar{X}_{12}$ and is given by

$$\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

Similarly, the combined mean of three series of size ($n_1+n_2+n_3$) is defined by the following formula

$$\bar{X}_{123} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2 + n_3\bar{X}_3}{n_1 + n_2 + n_3}$$ and so on

[redacted]

Particular regarding incomes of two villages are given below.

	Village	
	A	B
No. of people	600	500
Average income	Rs. 175.00	Rs. 186.00

Calculate the combined mean.

Solution:

Village A	Village B	Combined
n_1 = 600 $\overline{X}_1$ =Rs. 175	n_2 = 500 $\overline{X}_2$= Rs.186	$n_1 + n_2$= 600+500 =1100 $\overline{X}_{12}$= ?

Combined mean $(\overline{X}_{12}) = \frac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$

$= \frac{600\times175+500\times186}{600+500}$

= Rs 180

[redacted]

4. The average income of 100 persons in a city A is Rs.3500 and that of 150 persons a city B is Rs.4000. What is the average income of both the city A and city B?

Solution:

City A	City B	Combined
n_1 = 100 $\overline{X}_1$ =Rs.3500	n_2 = 150 $\overline{X}_2$= Rs.4000	$n_1 + n_2$= 250 $\overline{X}_{12}$= ?

The average income of both the city A and city B is

Combined mean $(\overline{X}_{12}) = \frac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$

$= \frac{100\times3500+150\times4000}{100+150}$

= Rs 3800

[redacted]

The mean of monthly salaries paid of 77 employees in the company was Rs. 78. The mean salary of 32 of them was Rs.75 and that of other 25 was Rs. 82. What was the mean salary of the remaining?

Solution:

		Remaining	Combined
n_1 = 32 $\overline{X}_1$ =Rs.75	n_2 = 25 $\overline{X}_2$= Rs.82	n_3 = 20 $\overline{X}_3$= ?	$n_1 + n_2 + n_3$= 77 $\overline{X}_{123}$ = 78

The average income of both the city A and city B is

Combined mean $(\overline{X}_{123}) = \frac{n_1\overline{X}_1+n_2\overline{X}_2+n_3\overline{X}_3}{n_1+n_2+n_3}$

or, 78 = $\frac{32\times75+25\times82+20X_3}{77}$

or, 78 = $\frac{4450+20\overline{X}_3}{77}$

or, $4450 + 20\overline{X}_3 = 6006$

or, $20\overline{X}_3 = 6006\text{-}4450$

or, $20\overline{X}_3 = 1556$

or, $\overline{X}_3 = \frac{1556}{20}$

$\therefore \overline{X}_3 = 77.8$

[redacted]

From the information given below, find

(iv) Which factory pays larger amount as daily wage?

(v) What is the average daily wage for the workers of two factories?

	Factory A	Factory B
No.of wage earners	35	20
Average daily wage	Rs.200	Rs.250

Solution:

(i)

For Factory A

$n_1 = 35$, $\overline{X}_1$ =Rs.200

$\overline{X}_1 = \frac{\sum X_1}{n_1}$

or, $200 = \frac{\sum X_1}{35}$

or, $\sum X_1 = 35 \times 200$ = Rs. 700

Amount paid as daily wage in factory A = Rs. 7000

For factory B:

$n_2 = 20$, $\overline{X}_2$ =Rs.250

$\bar{X}_2 = \frac{\sum X_2}{n_2}$

or, $250 = \frac{\sum X_2}{20}$

or, $\sum X_2 = 20 \times 250$ = Rs. 500

$\therefore$ Amount paid as daily wage in factory B =Rs.5000

Hence, Factory A pays larger amount as daily wage than factory B

(ii)

Combined mean $(\overline{X}_{12}) = \frac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$

$= \frac{35\times200+20\times250}{35+20}$

= Rs 218.181

∴ The combined average daily wage of two factories is Rs.218.181

In a class of 50 students 5 have failed and their average of marks is 10. The total marks secured by the entire class were 281. Find the average marks of those who have passed.

Solution

Given,

Passed	Failed	Combined
n_1 = 50-5 = 45 $\overline{X}_1$ =?	n_2 = 5 $\overline{X}_2$= 10	$n_1 + n_2$= 50 $\sum X_{12}$ = 281 $\overline{X}_{12} = \frac{\sum X_{12}}{50} = \frac{281}{50} = 5.62$

The average income of both the city A and city B is

Combined mean $(\overline{X}_{12}) = \frac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$

or, $5.62 = \frac{45\bar{X}_1+5\times10}{50}$

or, $45\bar{X}_1 + 50 = 5.62\times 50$

or, $45\bar{X}_1$ = 281-50

or, $45\bar{X}_1$ = 231

or, $\bar{X}_1 = \frac{231}{45}$ = 5.13

∴ The average marks of those who have passed = 5.13

The mean age of a combined group of men and women is 30 years. If the mean age of the group of men is 32 and that of the group of women is 27. Find out the percentage of men and women in the group.

Solution:

men	women	Combined
$\%n_1$ = ? $\overline{X}_1$ =32	$\%n_2$ = ? $\overline{X}_2$= 27	$n_1 + n_2$= 100 or, $n_2 = 100 - n_1$ $\overline{X}_{12}$= 30

Combined mean $(\overline{X}_{12}) = \frac{n_1\overline{X}_1+n_2\overline{X}_2}{n_1+n_2}$

or, $30 = \frac{n_1\times32+(100-n_1)\times27}{100}$

or, $3000 = 32n_1 + 2700-27n_1$

or, $3000-2700 = 5n_1$

or, $300 = 5n_1$

or, $\frac{300}{5} = n_1$

or, $60\% = n_1$

$n_2 = 100 - n_1$

$= 100 - 60$

$= 40\ \%$

$\therefore$ % men = 60%

& % women = 40%

Corrected mean

The formula for the calculation of corrected mean is given by

Corrected mean $(\overline{X}_{\text{corrected}}) = \frac{\text{Corrected}\sum X}{\text{correct n}}$

Where, Incorrect $\sum X = n\overline{X} = n \times$ incorrect mean

Correct$\sum X$ = Incorrect$\sum X$ - Incorrect items + correct items

[redacted]

Mean of 100 items was 45. Later on it was found that two items were misread as 60 and 8 instead of 192 and 66. Find the correct mean.

Solution

Here, we have given,

No. of items, n = 100,

Mean ($\overline{X}$) = 45

Misread items = 60 and 8

Correct items = 192 and 66

We have,

$$\overline{X} = \frac{\sum X}{n}$$

Or, $50 = \frac{\sum X}{100}$

Or, $\sum X = 5{,}000$

$\therefore$ Incorrect $\sum X = 5000$

Correct n = 100-1-1+1+1=100

Correct $\sum X = 5{,}000 - 60 - 8 + 192 + 66 = 5190$

Correct $(\overline{X}) = \frac{correct \sum X}{correct.n} = \frac{5190}{100} = 51.9$

$\therefore$ Correct mean $(\overline{X}) = 51.9$

[redacted]

Arithmetic mean of 150 items is 50. Two items 40 and 60 were left out at the time of calculations. What is the correct mean of all the items?

Solution:

Given, mean $(\overline{X}) = 50$

No. of items, n = 150

Left out items = 40 and 60

Required mean of all the items $(\overline{X})$ =?

We have,

$\overline{X} = \frac{\sum X}{n}$

or, $50 = \frac{\Sigma X}{150}$

or, $\sum X = 50 \times 150 = 7500$

Wrong $\sum X = 7500$

$\therefore$ Correct $\sum X = 7500 + 40 + 60 = 7600$

Correct n = 150 + 1 + 1 = 152

Correct $\overline{X} = \frac{correct \sum X}{correct.n} = \frac{7600}{100} = 76$

$\therefore$ Correct $(\overline{X}) = 76$

The mean marks of 100 students were found to be 65. At the time of checking it was found that the three marks 40, 50 and 55 were incorrect. Find the correct mean if the incorrect marks are omitted (or weeded out).

Solution

Given, mean ($\overline{X}$) = 65

No. of students, n = 100

Incorrect marks = 40, 50 & 55

Correct mean after omitting incorrect marks ($\overline{X}$) =?

We have,

$$\overline{X} = \frac{\sum X}{n}$$

or, $65 = \frac{\sum X}{100}$

or, $\sum X = 65 \times 100 = 6500$

$\therefore$ Correct $\sum X$ = 6500 - 40 -50-55 = 6355

i.e. incorrect $\sum X$ = 6500

Correct n = 100 -1 -1-1 = 97

Correct ($\overline{X}$) = $\frac{correct \sum X}{correct.n} = \frac{6355}{97} =$ = 65.515

The mean of 20 items was found to be 10. At the of checking, it was found that one item 8 was incorrect. Calculate the mean when

(i) The wrong item is omitted.
(ii) It is replaced by 12

Solution:

Here, we have given,

No. of items, n = 20,

Incorrect mean ($\overline{X}$) = 10

Incorrect (wrong) item = 8

Now,

$$\overline{X} = \frac{\sum X}{n}$$

or, $\sum X = n\overline{X}$

$\sum X = 20 \times 10 = 200$

Incorrect $\sum X = 200$

Then,

(i) When, incorrect item 8 is omitted

Correct n = 20-1=19

Correct $\sum X = 200\text{-}8 = 192$

$$\therefore \text{Correct } (\overline{X}) = \frac{correct \sum X}{correct.n} = \frac{192}{19} = 10.1$$

(ii) When the incorrect item is replaced by 12

Correct n = 20-1+1=20

Correct $\sum X = 200\text{-}8+12 = 204$

$$\therefore \text{Correct } (\overline{X}) = \frac{correct \sum X}{correct.n} = \frac{204}{20} = 10.2$$

Median or Positional average (Md)

The variate value which divides the total number of observations into two equal parts is called the median. It is denoted by Md.

(i) Md is suitable measure of central tendency (or average) for the qualitative characteristics such as knowledge, intelligent, beauty, honesty, talent, good, bad, defective, etc.

(ii) It is also more appropriate (or suitable) average (or measure of central tendency) for the open ended classified data.

Calculation of median depends upon the given series

For Individual series

At first, arranging the given set of observations (data) in ascending order of magnitude.

Median (Md.) = Value of $\left(\frac{n+1}{2}\right)^{th}$ item, Where n = no. of observations

[redacted]

Find the median from the following set of observations:

30, 40, 25, 18, 27, 26, 35

Solution:

At first, arranging the given data say in ascending order of magnitude: 18, 25, 26, 27, 30, 35, 40.

n = 7,

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

Median = value of $\left(\frac{7+1}{2}\right)^{th}$ item

= value of 4^{th} item

$\therefore$ Median = 27

[redacted]

Find median from the following data:

48, 30, 40, 35, 27, 29, 38, 45

Solution:

The given data in the ascending order: 27, 29, 30, 35, 38, 40, 45, 48

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

= Value of $\left(\frac{8+1}{2}\right)^{th}$ items

= Value of $(4.5)^{th}$ items

= Value of $\left(\frac{4^{th}+5^{th}}{2}\right)$ items

$= \frac{35+38}{2} = \frac{73}{2} = 36.5$

$\therefore$ Md. = 36.5

In discrete series

- At first, arrange the given data in ascending order of their magnitudes.
- Obtain the less than cumulative frequency (c.f.)
- Use the formula, Median = value of $\left(\frac{N+1}{2}\right)^{th}$ item ,

N = Σf =Total frequency

[redacted]

The following data gives the daily wages of 80 workers in a firm. Calculate median wage.

Daily wages (in '00'Rs.)	2	7	5	10	15
Number of workers	20	15	12	15	18

Solution: Calculation of median

Daily wages ('00' Rs.) (X)	No. of workers (f)	Less than c.f.
2	20	20
7	15	35
5	12	47
10	15	62
15	18	80
	N = 80	

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

= value of $\left(\frac{80+1}{2}\right)^{th}$ item

= value of 40.5^{th} item

= Rs. 8 (in 00)

Since, the cumulative frequency just greater than 40.5 is 47 and its corresponding value of 'X' is 8.

For continuous series (Grouped frequency distribution)

The following steps are to be used in finding the median in case of continuous series:

- At first, arrange the given data in ascending order of their magnitudes.
- Prepare the less than cumulative frequency distribution.
- Find $\frac{N}{2}$
- See cumulative frequency equal to or just greater than the value of $\frac{N}{2}$ and note the corresponding frequency.
- The corresponding class contains the median value and is called the median class.

Then, Median is computed by applying the following formula,

$$M_d = L + \frac{\frac{N}{2} - c.f.}{f} \times h$$

Where, N = Total frequency

$\frac{N}{2}$ = position of the median class.

L = Lower limit of median class

f = frequency of median class

h = difference of class interval of median class or class size of median class.

c.f. = Less than cumulative frequency preceding the median class.

Note: (i) The classes should be exclusive type

(iii) Median can also be calculated for the distribution having unequal class interval.

i.e. For calculation of M_d. It is not necessary to be equal class size.

The following data gives the weekly wages in rupees of 24 workers of a firm.

Wages per Week (in '000' RS.)	10-14	15-19	20-24	25-29	30-34
No. of Workers	4	7	8	3	2

Compute Median wage.

Solution

Since, the given class intervals are in inclusive type. So, for median we should first convert the given inclusive class interval into exclusive class interval before calculating median by using correction factor.

Correction factor = $\frac{15-14}{2} = 0.5$

Calculation of median

Wages per Week ('000'Rs.)	Workers (f)	Less than c.f.
9.5-14.5	4	4
14.5-19.5	7	11
19.5-24.5	8	19
24.5-29.5	3	22
29.5-34.5	2	24
	N = 24	

$\frac{N}{2} = \frac{24}{2} = 12$. The c.f. just greater than 12 is 19, whose corresponding class is 19.5 - 24.5 i.e. median lies in the class 19.5 - 24.5.

Where, L = 19.5, f = 8, h = 5, c.f. = 11

$$M_d = L + \frac{\frac{N}{2} - c.f.}{f} \times h$$

$M_d = 19.5 + \frac{12 - 11}{8} \times 5 = 20.125$

$\therefore M_d$ = Rs. 20.125 (in 000 Rs.)

Find out the appropriate measure of central tendency from the following distribution and support for your choice of measure.

Monthly Income (Rs.)	Below 100	100-199	200-299	300-399	400-499	500 & Above
No. of Families	50	550	1105	1205	1235	1250

Solution:

Since, the given distribution has open end classes (i.e. The lower limit of first class and upper limit of last class interval are in the form below and over). Therefore, M_d is the appropriate or suitable measure of central tendency.

We have, $M_d = L + \frac{\frac{N}{2} - \text{c.f.}}{f} \times h$

Correction factor = $\frac{200-199}{2} = 0.5$

Here, the classes are in the form of inclusive, so at first it should be converted into exclusive form.

Monthly Income (Rs.)	No. of Families (f)	less than c.f.
Below 99.5	50	50
99.5-199.5	500	550
199.5-299.5	555	1105
299.5-399.5	100	1205
399.5-499.5	30	1235
499.5 and above	15	1250
	N = 1250	

Here, $\frac{N}{2} = \frac{1250}{2} = 625$. The c.f. just greater than 625 is 1105.

So, Md lies between 199.5-299.5

Where, L = 199.5, h = 100, f = 555, c.f. = 550

Now, Median (Md) = $199.5 + \frac{625 - 550}{555} \times 100 = 213.01$

$\therefore$ Median (Md) = Rs. 213.01

The following table gives the weekly wages in rupees of workers in certain commercial organization. The frequency of the class interval 50-54 is missing.

Weekly wages (Rs.)	40-44	45-49	50-54	55-59	60-64
No. of workers	31	58	60	?	27

It is known that the median of the above frequency distribution is Rs. 53.5. Find the missing frequency.

Solution:

Since, the given class intervals are in inclusive type. So, for median we should first convert the given inclusive class interval into exclusive class interval before calculating median by using correction factor.

$$\text{Correction factor} = \frac{45-44}{2} = 0.5$$

Let the missing frequency be 'f_1'.

Calculation of missing frequency

Weekly wages (Rs.)	No. of workers (f)	Less than c.f.
39.5-44.5	31	31
44.5-49.5	58	89
49.5-54.5	60	149
54.5-59.5	f_1	$149 + f_1$
59.5-64.5	27	$176 + f_1$
	$N = 176 + f_1$	

Since, median = 53.5, which lies in the weekly wages group 49.5-54.5, so, median class is 49.5-54.5.

Where, L = 49.5, h = 5, f = 60, c.f. = 89

We have, $M_d = L + \frac{\frac{N}{2} - \text{c.f.}}{f} \times h$

or, $53.5 = 49.5 + \frac{\frac{176+f_1}{2} - 89}{60} \times 5$

or, $53.5 = 49.5 + \frac{176+f_1-178}{120} \times 5$

or, $53.5 = 49.5 + \frac{f_1-2}{24}$

or, $53.5 - 49.5 = \frac{f_1 - 2}{24}$

or, $f_1 - 2 = 4 \times 24$ or, $f_1 = 96 + 2 = 98$

$\therefore$ The required missing frequency = 98.

[redacted]

The following table represents the marks of 100 students

Marks	0-20	20-40	40-60	60-80	80-100
No. of students	14	?	27	?	15

If the median mark is 48 find missing frequencies and the mean marks of all 100 students.

Solution:

Let f_1 and f_2 be the missing frequency of 20-40 and 60-80 respectively.

Calculation of Missing Frequencies

Marks	No. of students (f)	Less than c.f.
0-20	14	14
20-40	f_1	$14 + f_1$
40-60	27	$41 + f_1$
60-80	f_2	$41 + f_1 + f_2$
80-100	15	$56 + f_1 + f_2$
	$N = 56 + f_1 + f_2$	

From given condition,

$N = 100$

or, $56 + f_1 + f_2 = 100$

or, $f_1 + f_2 = 100 - 56$

or, $f_1 + f_2 = 44$

or, $f_2 = 44 - f_1$ . . . . . .(i)

Again,

Since, Median (Md) = 48 lies in class 40 - 60. So, median class is 40 - 60

Where,

L = 40, h = 20, f = 47, c.f. = $14 + f_1$

We know that,

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

or, $$48 = 40 + \frac{\frac{100}{2} - (14 + f_1)}{27} \times 20$$

or, $$48 - 40 = \frac{(50 - 14 - f_1)}{27} \times 20$$

or, $8 \times 27 = (36 - f_1) \times 20$

or, $216 = 720 - 20f_1$

or, $20f_1 = 720 - 216 = 504$

or, $f_1 = \frac{504}{20} = 25.2 \approx 25$

From equation (i)

$25 + f_2 = 44$

or, $f_2 = 44 - 25 = 19$

∴ The required missing frequencies are respectively 25 and 19.

Calculation of Mean

Marks	No. of students (f)	Mid-value (X)	fX
0-20	14	10	140
20-40	25	30	750
40-60	27	50	1350
60-80	19	70	1330
80-100	15	90	1350
	N = 100		ΣfX = 4920

Mean ($\overline{X}$) = $\frac{\Sigma fX}{N} = \frac{4920}{100} = 49.2$

Hence, mean mark of all students is 49.2.

Calculate the median income from the following income distribution of 100 persons.

Income in Rs. (Less than)	100	150	200	275	350	400	500
No. of persons	5	15	30	50	75	90	100

OR

Income in Rs.	Below 100	Below 150	Below 200	Below 275	Below 350	Below 400	Below 500
No. of persons	5	15	30	50	75	90	100

Solution:

Since, the given frequency distribution of income is in the less than cumulative frequency distribution form. So, it should be converted into simple frequency distribution.

Income (in Rs.)	f	Less than c.f.
Below100	5	5
100-150	15-5=10	15
150-200	30-15=15	30
200-275	50-30=20	50
275-350	75-50 =25	75
350-400	90-75 =15	90
400-500	100-90 =10	100
	N = 100	

Here, $\frac{N}{2} = \frac{100}{2} = 50$. The cumulative frequency just greater than or equal to 50 is 50. So median lies in the class interval 200-275.

Here, L = 200, h = 75, f = 20, c.f. = 30

We have,

$$M_d = L + \frac{\frac{N}{2} - c.f.}{f} \times h$$

$$= 200 + \frac{50-30}{20} \times 75 = 275$$

Hence, Median income is Rs.275

Following is the distribution of marks in accountancy obtained by 50 students. Find median

Marks (more than):	0	10	20	30	40	50
No. of students	50	46	40	20	10	3

OR

Marks	Above 0	Above 10	Above 20	Above 30	Above 40	Above 50
No. of students	50	46	40	20	10	3

Solution:

Since, the given frequency distribution of marks is in the more than cumulative frequency distribution form. So, at first it should be converted into simple frequency distribution.

Calculation of median

Marks	No. of Students (f)	Less than c.f.
0-10	50 – 46 = 4	4
10-20	46 – 40 = 6	10
20-30	40 – 20 = 20	30
30-40	20 – 10 = 10	40
40-50	10 – 3 = 7	47
50 & above	3	50
	N = 50	

Here, $\frac{N}{2} = \frac{50}{2} = 25$. The cumulative frequency just greater than 25 is 30. So median lies in the class interval 20-30.

Where, L = 20, h = 10, f = 20, c.f. = 10

We have,

$$M_d. = L + \frac{\frac{N}{2} - c.f.}{f} \times h = 20 + \frac{25 - 10}{20} \times 10 = 27.5$$

$\therefore$ Median = 27.5

The Partition Values

The values which divide the total number of observations into a number of equal parts are called partition values. Thus, median may also be regarded as a particular partition value because it divides the given data into two equal parts.

Depending upon the equal number of parts, the important amongst these partition values are

(i) Quartiles
(ii) Deciles
(iii) percentiles

Note: For partition values, classes should be exclusive but not necessary to be equal class size.

Quartiles: Quartiles are those variate values, which divide the total observations into four equal parts and each part equal to 25%. Hence, in such cases there are three quartiles (points) Q_1, Q_2 and Q_3 such that $Q_1 < Q_2 < Q_3$. To divide into four equal parts three quartiles namely Q_1, Q_2 and Q_3 are used. Q_1 is called the lower quartile and Q_3 is the upper quartile.

Individual series

After arranging the given data in ascending order of magnitudes,

Quartiles can be obtained by the following formula

Q_i = value of $\frac{i(n+1)^{th}}{4}$ item Where, i = 1, 2, 3

n = No. of observations

[redacted]

Find out the lower (or first quartile) and upper quartile (or third quartile) from the given data.

18, 20, 15, 16, 25, 19, 12, 22, 14

Solution

Here, number of observations i.e. n = 9,

At First, the given data are arranged in ascending order.

12, 14, 15, 16, 18, 19, 20, 22, 25

Lower quartile (Q_1) = value of $\frac{(n+1)^{th}}{4}$ item

= value of $\frac{(9+1)^{th}}{4}$ item

= value of 2.5th item

= value of 2nd item + 0.5 ($3^{rd} - 2^{nd}$) item

$= 14 + 0.5\ (15 - 14)$

$\therefore\ Q_1 = 14.5$

Similarly, upper quartile (Q_3) = value of $3\ \frac{(n+1)^{th}}{4}$ item

$= \text{value of } 3\ \frac{(9+1)^{th}}{4} \text{ item}$

$= \text{value of } 7.5\text{th item}$

$= \text{value of } 7\text{th item} + 0.5(8^{th} - 7^{th}) \text{ item}$

$= 20 + 0.5\ (22 - 20) = 21$

$\therefore Q_3 = 21$

Discrete series

Q_i = value of $\frac{i(N+1)^{th}}{4}$ item , i = 1, 2 & 3

$N = \sum f =$ Total frequency

[redacted]

Find first and third quartiles from the given data.

X	1	2	3	4	5	6	7
f	2	5	7	10	4	3	2

Solution:

Calculation of quartiles

X	f	Less than c.f.
1	2	2
2	5	7
3	7	14
4	10	24
5	4	28
6	3	31
7	2	33
	$N = \sum f = 33$	

Q_1 = Value of $\frac{(N+1)^{th}}{4}$ item

= value of $\left(\frac{33+1}{4}\right)^{th}$ item

= value of 8.5^{th} item. The c.f. just greater than 8.5 is 14 and its corresponding value of X is 3

$\therefore Q_1 = 3$

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

= value of 3 $\left(\frac{33+1}{4}\right)^{th}$ item

= value of 25.5^{th} item. The c.f. just greater than 25.5 is 28 and its corresponding value of X is 5

$\therefore Q_3 = 5$

Continuous series or Grouped frequency distribution

Quartiles are given by the relation:

$$Q_i = L + \frac{\frac{iN}{4} - \text{c.f.}}{f} \times h$$

where i=1, 2, 3

When i = 1, Q_1 = first (or lower) quartile = $L + \frac{\frac{N}{4} - \text{c.f.}}{f} \times h$

When i = 2, Q_2 = second (or middle or median) quartile

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

When i = 3, Q_3 = third (or upper) quartile = $L + \frac{\frac{3N}{4} - \text{c.f.}}{f} \times h$

$\frac{iN}{4}$ = the size for i^{th} quartile's class

L = lower limit of i^{th} quartile's class

f = frequency of i^{th} quartile's class

h = difference of class interval of ith quartile's class

c.f. = preceding c.f. of i^{th} quartile's class.

Find first and third quartiles from the data given below.

Income (in '000'Rs.)	20-30	30-40	40-50	50-60	60-70

No. of households	10	15	25	16	14

Solution:

Income ('000' Rs.)	No. of households (f)	less than c.f.
20-30	10	10
30-40	15	25
40-50	25	50
50-60	16	66
60-70	14	80
	N = $\sum$ f = 80	

For first quartile (Q_1)

$\frac{N}{4} = \frac{80}{4} = 20$

Q_1 lies in class 30-40 , L = 30, f = 15, c.f. = 10, h = 10

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

$= 30 + \frac{20-10}{15} \times 10$

= 36.66667(in 000)

= Rs.36666.67

For third quartile (Q_3)

$\frac{3N}{4} = \frac{3\times80}{4} = 60$

Q_3 lies in class 50-60, L = 50, f = 16, c.f. = 50, h = 10

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

$= 50 + \frac{60-50}{16} \times 10$

= 56.25 (in 000)

= Rs.56250

From the given following data, calculate the number of workers getting wage between first and third quartile.

Wage in Rs.(less than)	10	20	30	40	50
No. of workers	45	85	160	75	35

OR

Wage (in Rs.)	Below10	Below 20	Below 30	Below 40	Below 50

No. of workers	45	85	160	75	35

Solution:

Since, the given frequency distribution of income is in the less than cumulative frequency distribution form. So, it should be converted into simple frequency distribution.

Calculation for Quartiles

Wage in Rs.	No. of workers (f)	c.f.
0-10	45	45
10-20	85	130
20-30	160	290
30-40	75	365
40-50	35	400
	N = 400	

For Q_1: $\frac{N}{4} = \frac{400}{4} = 100$, the c.f. just greater than 100 is 190.

So, Q_1 lies in the class 10-20, L = 10, c.f. = 45, f = 85 and h = 10.

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

$$= 10 + \frac{100 - 45}{85} \times 100 = \text{Rs. } 16.47$$

For Q_3: $\frac{3N}{4} = \frac{3 \times 400}{4} = 300$, the c.f. just greater than 300 is 365.

So, Q_3 lies in the class 30-40, L = 30, c.f. = 290, f = 75 and h = 10

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

$$= 30 + \frac{300 - 290}{75} \times 10$$

$$= \text{Rs. } 31.33$$

$\therefore$ The number of workers getting wage between Rs. 16.47 to Rs. 31.33.

$$= \frac{20 - 16.67}{10} \times 85 + 16 + \frac{31.33 - 30}{10} \times 75$$

$$= 55.98 \approx 56$$

Deciles:

The variate values which divide the total number of observations into ten equal parts are called deciles and each part equal to 10%. Hence, there are in all nine deciles denoted by D_1, D_2, ...,D9 such that $D_1 < D_2 < D_3 < \ldots\ldots\ldots < D_9$. D_5 the fifth decile divides the series into two halves and hence it is the same as median.

Individual series

After arranging the given data in ascending order of magnitudes,

D_j = value of $\frac{j(n+1)^{th}}{10}$ item.

Where j = 1, 2, 3. . ., 9

n = No. of observations

Discrete series

D_j = value of $\frac{j(N+1)^{th}}{10}$ item. , i= 1, 2 ,3 . . . . . ,9

Where, $N = \sum f$ = Total frequency

Continuous series or Grouped frequency distribution

$$D_j = L + \frac{\frac{jN}{10} - c.f.}{f} \times h$$

Where, j =1, 2, 3. . . . ., 9

Where,

$\frac{jN}{10}$= the size for j^{th} decile's class

L = lower limit of j^{th} decile's class

f = frequency of j^{th} decile's class

h = difference of class interval of jth decile's class

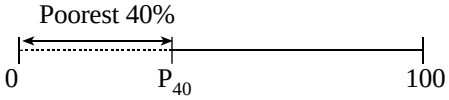
c.f. = preceding c.f. of j^{th} decile's class.

Percentiles:

The variate values which divide the total number of observations into 100 equal parts are called percentiles. There are in all 99 percentiles and are denoted by P_1, P_2, ...,P_{99} respectively such that $P_1 < P_2 < P_3 < \ldots\ldots\ldots < P_{99}$. P_{50}^{th} percentile is same as median.

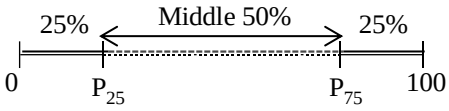
Case I: To find the highest value (maximum value) of % failed, lowest earner, poorest, flattest, shortest etc.

i.e. The highest income of the poorest 40% of the people is given by 40^{th} percentile i.e. P_{40}.



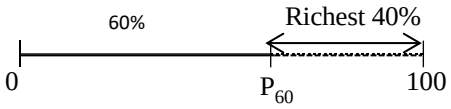
Case II: To find the limits (Range) of middle %

i.e. The limits of income of middle 50% of families is given by the 25^{th} and 75^{th} percentiles. i.e. P_{25} and P_{75}.



Case III: To find the lowest value (or minimum value) of % top, pass, richest, highest earner, longest, tallest etc.

i.e. The lowest income of the richest 40% of the people is given by 60^{th} percentile i.e. P_{60}.



Individual series

After arranging the given data in ascending order of magnitudes,

Percentiles can be obtained by the following formula

P_k = value of $\frac{k(n+1)^{th}}{100}$ item.

Where k = 1, 2, 3, ..., ,99

n = No. of observations

Discrete series

P_k = value of $\frac{K(N+1)^{th}}{100}$ item. N = $\sum$ f = Total frequency

k= 1, 2, 3 . . . . . ,99

Continuous series or Grouped frequency distribution

$$P_K = L + \frac{\frac{KN}{100} - c.f.}{f} \times h$$

Where, K =1, 2,3, ...,99

$\frac{kN}{100}$ = the size for K^{th} percentile's class

L = lower limit of k^{th} percentile's class

f = frequency of k^{th} percentile's class

h = difference of class interval of k^{th} percentile's class

c.f. = preceding c.f. of k^{th} percentile's class.

ˆ Median = Q_2 = D_5 = P_{50}

Find Q_1, D_3 and P_{65} from the given data: 8, 6, 5, 4, 10, 15, 3, 16

Solution

Here, the number of observation, i.e. n = 8

First, the data are arranged in ascending order: 3, 4, 5, 6, 8, 10, 15, 16.

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

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

$$= 2.25^{th} \text{ item}$$

$$= 2^{nd} \text{ item} + 0.25\,(3^{rd} - 2^{nd}) \text{ item}$$

$$\therefore \quad Q_1 = 4 + 0.25\,(5 - 4) = 4.25$$

$$\text{Similarly, } D_3 = \text{Value of } \left(\frac{3\,(n+1)^{th}}{10}\right) \text{ item}$$

$$= \text{value of } \frac{3(8+1)^{th}}{10} \text{ item}$$

$$= \text{value of } 2.7^{th} \text{ item}$$

$$= \text{value of } 2^{nd} \text{ item} + 0.7\,(3^{rd} - 2^{nd}) \text{ item}$$

$$\therefore \quad D_3 = 4 + 0.7\,(5 - 4) = 4.7$$

$$P_{65} = \text{Value of } \left(\frac{65\,(n+1)^{th}}{100}\right) \text{ item}$$

$$= \text{value of } \frac{65(8+1)^{th}}{100} \text{ item}$$

$$= \text{value of } 5.85^{th} \text{ item}$$

$$= \text{value of } 5^{th} \text{ item} + 0.85\,(6\text{th item} - 5^{th} \text{ item})$$

$$= 8 + 0.85\,(10 - 8) = 9.7$$

$$\therefore \quad P_{65} = 9.7$$

Find upper quartile and upper decile i.e. D_9 from the given data. Also obtain P_{77}.

X	1	2	3	4	5	6	7	8	9	10	11
f	2	5	8	10	12	8	6	4	3	2	1

Solution:

Calculation of partition values

X	f	Less than c.f.
1	2	2
2	5	7
3	8	15
4	10	25
5	12	37
6	8	45
7	6	51
8	4	55
9	3	58
10	2	60
11	1	61
	N = 61	

For Q_3 :

Q_3 = value of $\frac{3(N+1)^{th}}{4}$ item.

= value of $\frac{3(61+1)^{th}}{4}$ item.

= value of 46.5th item. The value in c.f. just greater than 46.5 is 51 its corresponding value of Variable X is 7

So upper quartile i.e. $Q_3 = 7$.

Similarly, For **D_9**

D_9 = value of $\frac{9(N+1)^{th}}{10}$ item.

= value of $\frac{9(61+1)^{th}}{10}$ item.

= Value of 55.8^{th} item

The value of c.f. just greater than 55.8 is 58 its corresponding value of Variable X is 9.

So, upper decile i.e. $D_9 = 9$.

For P_{77}

P_{77} = value of $\frac{77(N+1)^{th}}{100}$ item.

= value of $\frac{77(61+1)^{th}}{100}$ item.

= Value of 47.74^{th} item

The value of c.f. just greater than 47.74 is 51 and its corresponding value of Variable X is 7.

$\therefore P_{77} = 7$.

The following table shows the frequency distribution of wages (in Rs.) of 50 employees of Darwin Inc.

Wages(Rs.)	50-70	70-90	90-110	110-130	130-150
No. of Employees	6	11	15	10	8

Write reference to the above table,

(i) Determine the percentage of employees having earnings more than Rs. 85.

(ii) Find the number of employees having earnings less than Rs.125.

(iii) Find the number of employees having earnings between Rs.65 and Rs.140.

Solution:

Wages(Rs.)	50-70	70-90	90-110	110-130	130-150	Total
No. of Employees	6	11	15	10	8	50

Frequency (No. of employees) = $\frac{\text{Difference}}{\text{class size}} \times$ corresponding frequency of the class

(i) The number of workers having earnings more than Rs. 85 = $\frac{(90-85)}{20} \times 11+15+10+8$

$= 2.75+15+10+8$

$= 35.75 \approx 36$

$\therefore$ The percentage of workers having earnings more than Rs. 85 = $\frac{35.75}{50} \times 100$

$= 71.5\%$

(ii) The number of workers having earnings less than Rs.125 = $6+11+15 + \frac{(125-110)}{20} \times 10$

$= 6+11+15+ 7.5$

$= 39.5 \approx 40$

(iii) The number of employees having earnings between Rs.65 and Rs.140

$= \frac{(70-65)}{20} \times 6 + 11 + 15+ 10+ \frac{(140-130)}{20} \times 8$

$= 1.5+11+15+10+4$

$= 41.5 \approx 42$

The data given below represents the daily wages (in Rs.) received by 55 workers in a construction site.

Wages (In Rs.)	200-250	250-300	300-400	400-450	450-650	650-700
No. of workers	5	10	20	15	3	2

Find (i) lower quartile (ii) upper quartile (iii) 7th decile (iv) 3rd decile (v) 60th percentile & (vi) 80th percentile.

Note: For partition values, classes should be exclusive but not necessary to be equal class size.

Solution:

Wages (in Rs.)	No. of workers (f)	Less than c.f.
200-250	5	5
250-300	10	15
300-400	20	35
400-450	15	50
450-650	3	53
650-700	2	55
	$N = \sum f = 55$	

(i) For upper quartile (Q_1)

$\frac{N}{4} = \frac{55}{4} = 13.75$

Q_1 lies in class 250-300 , L = 250, f = 10, c.f. = 5, h = 50

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

$= 250 + \frac{13.75-5}{10} \times 50$

= Rs. 293.75

(ii) For third quartile (Q_3)

$\frac{3N}{4} = \frac{3\times55}{4} = 41.25$

Q_3 lies in class 400- 450 , L = 400, f = 15, c.f. = 35, h = 50

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

$= 400 + \frac{41.25-35}{15} \times 50$

= Rs. 420.83

(iii) 7th decile (D_7)

$\frac{7N}{10} = \frac{7\times55}{10} = 38.5$

D_7 lies in class 300- 400 , L = 300, f = 15, c.f. = 35, h = 50

$D_7 = L + \frac{\frac{7N}{10} - c.f.}{f} \times h$

$= 400 + \frac{38.5-35}{15} \times 50$

= Rs. 411.6667

(iv) 3rd decile (D_3)

$\frac{3N}{10} = \frac{3\times55}{10} = 16.5$

D_3 lies in class 300- 400 , L = 300, f = 20, c.f. = 15, h = 100

$$D_3 = L + \frac{\frac{3N}{10} - \text{c.f.}}{f} \times h$$

$$= 300 + \frac{16.5-15}{20} \times 100$$

= Rs. 307.5

(v) 60^{th} percentile (P_{60})

$$\frac{60N}{100} = \frac{60\times55}{100} = 33$$

P_{60} lies in class 300- 400, L = 300, f = 20, c.f. = 15, h = 100

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

$$= 300 + \frac{33-15}{20} \times 100$$

= Rs. 390

(vi) 80^{th} percentile. (P_{80})

$$\frac{80N}{100} = \frac{80\times55}{100} = 44$$

P_{80} lies in class lies in class 400- 450 , L = 400, f = 15, c.f. = 35, h = 50

$$P_{80} = L + \frac{\frac{60N}{100} - \text{c.f.}}{f} \times h$$

$$= 400 + \frac{44-35}{15} \times 50$$

= Rs. 430

[redacted]

Find the 1^{st} quartile and the limits of marks of middle 50% students from the following data. Also interpret the value of 1^{st} quartile.

Marks	10-20	20-30	30-40	40-50	50-60
Frequency	8	6	7	12	7

Solution:

Marks	frequency (f)	Less than c.f.
10-20	8	8
20-30	6	14
30-40	7	21
40-50	12	33
50-60	7	40
	$N = \sum f = 40$	

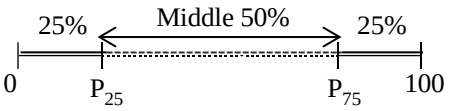
For Q_1: $\frac{N}{4} = \frac{40}{4} = 10$, the c.f. just greater than 10 is 14.

So, Q_1 lies in the class 20-30, L = 20, c.f. = 8, f = 6 and h = 10.

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

$$= 20 + \frac{10-8}{6} \times 10$$

$$= 23.33 \text{ marks}$$



The limits of marks of middle 50% of students is given by the 25th and 75th percentiles. i.e. P_{25} and P_{75}.

For P_{25}

$$\frac{25N}{100} = \frac{25\times40}{100} = 10$$

P_{25} lies in class 20-30, L = 20, c.f. = 8, f = 6 and h = 10.

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

$$= 20 + \frac{10-8}{6} \times 10$$

$$= 23.33 \text{ marks}$$

For P_{75}.

$$\frac{75N}{100} = \frac{75\times400}{100} = 30$$

P_{75} lies in class 40-50 , L = 40, f = 12, c.f. = 21, h = 10

$$P_{75} = L + \frac{\frac{75N}{100} - \text{c.f.}}{f} \times h$$

$$= 40 + \frac{30-21}{12} \times 10$$

$$= 47.5 \text{ marks}$$

The limits of marks of middle 50% of marks are

Lower limit, P_{25} = 23.33 marks

Upper limit, P_{75} = 47.5 marks

$Q_1 = P_{25}$ = 23.33 marks

Interpretation of the value of 1st quartile (Q_1):

Q_1 is called the lower quartile and it indicates that 25% marks are below the value of Q_1 i.e. 75% marks are above the value of Q_1

Monthly income (in 000 Rs.) distribution of 400 families of a certain town are given below:

Income (in 000 Rs.) Below	80	120	160	200	240	280
No. of families	10	35	165	310	380	400

OR

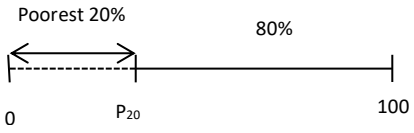
Income (in 000Rs.)	Below 80	Below 120	Below 160	Below 200	Below 240	Below 280
No. of families	10	35	165	310	380	400

a) Find the highest income of the poorest 20% of the families.
b) Find the limits of incomes of middle 50% families.
c) Find the lowest income of the richest 20 % of the families.

Solution:

Income (000in Rs.)	No. of families (f)	Less than c.f.
Below 80	10	10
80-120	25	35
120-160	130	165
160-200	145	310
200-240	70	380
240-280	20	400
	$N = \sum f = 400$	

a)



The highest income of the poorest 20% of the families is given by P_{20}

$\frac{20N}{100} = \frac{20\times400}{100} = 80$

P_{20} lies in class 120- 160, L = 120, f = 130, c.f. = 35, h = 40

$P_{20} = L + \frac{\frac{20N}{100} - c.f.}{f} \times h$

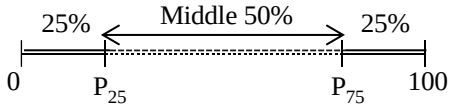
$= 120 + \frac{80-35}{130} \times 40$

= Rs. 133.8462 (000)

= Rs 133846.2

The highest income of poorest 20% of the families is Rs 133846.2

b)



The limits of income of middle 50% of families is given by the 25^{th} and 75^{th} percentiles. i.e. P_{25} and P_{75}.

For P_{25}

$$\frac{25N}{100} = \frac{25\times400}{100} = 100$$

P_{25} lies in class 120- 160, L = 120, f = 130, c.f. = 35, h = 40

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

$$= 120 + \frac{100-35}{130} \times 40$$

= Rs. 140 (000)

= Rs 140000

For P_{75}.

$$\frac{75N}{100} = \frac{75\times400}{100} = 300$$

P_{75} lies in class 160- 200, L = 160, f = 145, c.f. = 165, h = 40

$$P_{75} = L + \frac{\frac{75N}{100} - c.f.}{f} \times h$$

$$= 160 + \frac{300-165}{145} \times 40$$

= Rs. 197.2414 (000)

= Rs 197241.4

The limits of income of middle 50% of families are

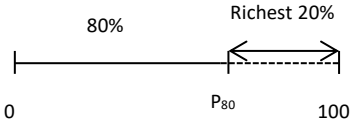
Lower limit, P_{25} = Rs 140000

Upper limit, P_{75} = Rs 197241.4

Range = Rs 197241.4- Rs 140000

= Rs. 57241.4

c) Find the lowest income of the richest 20 % of the families.



The lowest income of the richest 20 % of the families is given by P_{80}

For P_{80}.

$$\frac{80N}{100} = \frac{80\times400}{100} = 320$$

P_{80} lies in class 200- 240, L = 200, f = 70, c.f. = 310, h = 40

$$P_{80} = L + \frac{\frac{80N}{100} - \text{c.f.}}{f} \times h$$

$$= 200 + \frac{320-310}{70} \times 40$$

= Rs. 205.7143 (000)

= Rs 205714.3

∴ The lowest income of the richest 20 % of the families is Rs 205714.3

[redacted]

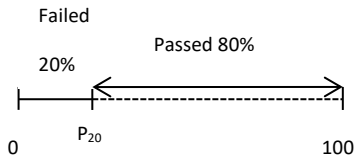
From the following distribution of marks of 500 students of a campus, find the minimum pass mark if only 20% of the students had failed and also find the minimum marks obtained by the top 25% of the students**.**

Marks	0-20	20-40	40-50	50-60	60-80	80-100
No. of students	50	100	150	90	60	50

Solution: Calculation of partition values

Marks	No. of students (f)	Less than c.f.
0-20	50	50
20-40	100	150
40-50	150	300
50-60	90	390
60-80	60	450
80-100	50	500

Total	N = 500	



If 20% students had failed, it means 80% students had passed. Then, the minimum pass mark is given by P_{20}.

$P_{20} = D_2 = ?$

Now, $\frac{20N}{100} = \frac{20 \times 500}{100} = 100$. The c.f. just greater than 100 is 150.

So, P_{20} lies in class 20 - 40. L = 20, f =100, c.f. = 50, h= 20

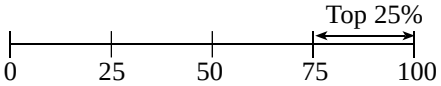
$$P_{20} = L + \frac{\frac{20N}{100} - \text{c.f.}}{f} \times h$$

$$P_{20} = 20 + \frac{100 - 50}{100} \times 20 = 30$$

∴ The required minimum pass mark = 30

For the top 25%:

The minimum marks obtained by top 25% of the students is given by P_{75}.



$P_{75} = Q_3$?

For P_{75}:

$\frac{75N}{100} = \frac{3 \times 500}{4}$ = 375. The c.f. just greater than 375 is 390. So P_{75} lies in class 50-60.

L = 50, f = 90, c.f. = 300, h = 10

$P_{75} = L + \frac{\frac{75N}{100} - \text{c.f.}}{f} \times h$

$= 50 + \frac{375-300}{90} \times 10$

$=58.33 \approx 58$

Therefore, the minimum marks obtained by the top 25% students = 58.

[redacted]

The marks distribution of 150 students in a class test as follows.

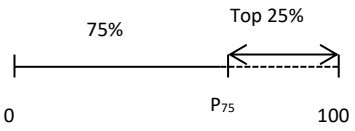
Marks (less than)	10	20	30	40	50	60	70
No. of students	8	28	50	100	120	142	145

OR

Marks	Below10	Below 20	Below 30	Below 40	Below 50	Below 60	Below 70
No. of students	8	28	50	100	120	142	145

If the campus has a policy to run a special class for the top 25% of the students what is the lowest mark obtained by those top 25% of the students.

Solution:



The lowest mark obtained by those top 25% of the students is given by P_{75} i.e. Q_3

Marks	No. of students (f)	Less than c.f.
Less than 10	8	8
10-20	20	28
20-30	22	50
30-40	50	100

40-50	25	125
50-60	17	142
60-70	3	145
	N = 145	

$$\frac{75N}{100} = \frac{3\times145}{4} = \frac{3\times145}{4} = 108.75.$$

The c.f. just greater than 108.75 is 125. So, Q_3 or P_{75} lies in class 40-50,

Where, L = 40, h = 10, f = 25, c.f. = 100.

$$P_{75} = L + \frac{\frac{75N}{100} - \text{c.f.}}{f} \times h$$

$$= 40 + \frac{108.75-100}{25} \times 10$$

$$= 43.5$$

∴The lowest marks obtained by those top 25% students is 43.5

The table given below represents the daily wage distribution of 130 workers. Find out the range of income of the middle 60% workers.

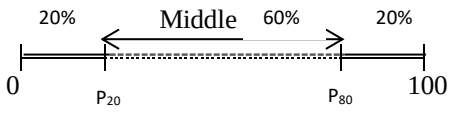
Wage (Rs. per week)	More than 70	More than 85	More than 100	More than 115	More than 130	More than 145	More than 160	More than 175
No. of workers	130	122	109	79	44	26	14	5

OR

Wage (Rs. per week) (More than)	70	85	100	115	130	145	160	175
No. of workers	130	122	109	79	44	26	14	5

Solution:

The limits of income of the middle 60% workers are given by P20 and P_{80}



Calculation of partition values

Wage (Rs. per week)	No. of workers (f)	Less than c.f.
70-85	8	8
85-100	13	21
100-115	30	51
115-130	35	86
130-145	18	104
145-160	12	116
160-175	9	125
More than 175	5	130
	N = 130	

Size for P_{20}:

$\frac{20N}{100} = \frac{20 \times 130}{100} = 26$. The c.f. just greater than 26 is 51.

So, P_{20} lies in class 100 – 115. L =100, f = 30, c.f. =21, h = 15

$$P_{20} = L + \frac{\frac{20N}{100} - \text{c.f.}}{f} \times h$$

$$P_{20} = 100 + \frac{26 - 21}{30} \times 15$$

= Rs. 102.50

Similarly,

For P_{80},

$\frac{80N}{100} = \frac{80 \times 130}{100} = 104.$

The c.f. just equal to 104. So, P_{80} lies in class 130 – 145.

L =130, f = 18, c.f. = 86, h= 15

$$P_{80} = L + \frac{\frac{80N}{100} - \text{c.f.}}{f} \times h$$

$$\therefore P_{80} = 130 + \frac{104 - 86}{18} \times 15$$

$$= 130 + \frac{18 + 15}{18}$$

$$= 130 + 15 = 145.$$

Lower limit, P_{20} = Rs 102.50

Upper limit, P_{80} = Rs 145

Hence, the range of income of the middle 60% workers = $P_{80} - P_{20}$

$$= 145 - 102.50$$

$$= \text{Rs. } 42.5.$$

Mode or Modal value or Most repeated value (Mo)

Mode is that variate value which repeats maximum number of times.

There are many situations in which A.M. and median (Md.) fail to reveal the characteristics of data such as most common stock, most common wage, most common height, most common size of shoe, size of T-shirts and other ready-made garments we have in mind mode but not the A.M. or Median

Calculation of mode

For Individual series:

Mode = Value of variable X which repeats maximum number of times

Case I: **If the distribution is regular and unimodal (i.e. only one maximum frequency**).

The mode for various distributions is given below.

For discrete series.

Mode = Value of variable (X) corresponding to maximum frequency

For continuous series (Grouped frequency distribution)

In case of continuous series, if the distribution is regular and unimodal, mode is calculated by using the following formula.

$$\text{Mode (Mo)} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

f_1 = maximum or modal class frequency.

f_0 = preceding frequency of modal class.

f_2 = following frequency of modal class.

L = lower limit of modal class.

h = difference of class interval of modal class or class size of modal class.

Note: For Mode

(i) It is necessary to be equal class size as well as exclusive class intervsls

Case II: When the distribution is regular and bimodal or multimodal, the mode can be determined by using Karl Pearson's empirical relation

$M_o = 3Md - 2\bar{X}$

This empirical relation is used for all series and mean ($\overline{X}$) and median (Md) are calculated by using the formulae according to their series.

Case III: When the distribution is irregular, mode is determined by using grouping method.

Note: Case III is beyond of our syllabus.

Relationship between Arithmetic Mean, Median and Mode

If the distribution is symmetrical then the arithmetic mean, median and mode coincide
i.e. the mean = median = mode.
However, for a moderately asymmetrical (non-symmetrical or skewed) distribution, mean, median and mode follows the following relationship.

Mode = 3 Median – 2 Mean

For Individual series:

Mode = Value of variable X which repeats maximum number of times

[redacted]

Find the mode from the data of marks obtained by 10 students.
Marks: 10, 27, 24, 12, 27, 20, 18, 15, 30, 27
Solution
Here, the value 27 repeats maximum number of times i.e. 3 times.
Hence, Mode = 27 marks

For discrete series.

Mode = Value of variable X corresponding to maximum frequency

[redacted]

Find the modal size of shoes from the data given below:

Size of shoes	5	6	7	8	9	10	11
No. of persons	10	20	25	40	22	15	6

Solution:
Since, the given frequency distribution is regular and unimodal.
Mode = Value of variable corresponding to maximum frequency
= 8
∴ Modal size of shoes is 8.

For continuous series (Grouped frequency distribution)

In case of continuous series, if the distribution is regular and unimodal, mode is calculated by using the following formula.

$$\text{Mode (Mo)} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where,

f_1 = maximum or modal class frequency.

f_0 = preceding frequency of modal class.

f_2 = following frequency of modal class.

L = lower limit of modal class.

h = difference of class interval of modal class or class size of modal class.

Note: For Mode

(i) It is necessary to be equal class size as well as exclusive class intervsls

[redacted]

Find the mode from the following frequency distribution

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No. of students	4	12	15	20	32	14	13

Solution

Since, the given frequency distribution is regular and unimodal and maximum frequency is 32.

So, modal class is 40-50

L=40, h =10, f_1=32, f_0= 20, f_2=14

Mode (Mo) = $L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

$= 40 + \frac{32-20}{2\times32-20-14} \times 10$

$= 40 + \frac{12}{30} \times 10$

= 44 marks

[redacted]

Find the modal value from the given frequency distribution.

Value (mid-point)	5	10	15	20	25	30
Frequency	3	7	12	16	13	3

Solution

Since, the mid values of the distribution are given.So, at first we need to construct the class intervals.

Class size (h) = difference between two successive mid-values

=10-5

=5

$\therefore \frac{h}{2} = 2.5$

Subtract $\frac{h}{2} = 2.5$ from the first middle value for lower limit of first class interval and add 2.5 to the same mid value for the upper limit of first class interval and so on. Other class intervals are constructed in the similar fashion as shown in the calculation table below.

Calculation of mode

Class	f
2.5 - 7.5	3
7.5 - 12.5	7
12.5 - 17.5	12
17.5 - 22.5	16
22.5 - 27.5	13
27.5 - 32.5	3

Here, the given frequency distribution is regular and unimodal and maximum frequency is 16, so mode lies between 17.5 - 22.5 i.e. modal classe is 17.5 - 22.5.

Where, $f_1 = 16$, $f0 = 12$, $f_2 = 13$, $L = 17.5$, $h = 5$

We have, mode $= L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

$= 17.5 + \frac{16 - 12}{2 \times 16 - 12 - 13} \times 5$

$= 17.5 + \frac{4}{32 - 25} = 20.36$

$\therefore$ Mode = 20.36

Unequal Class Interval

In this case, first of all class intervals should be converted into equal class interval, then express other unequal class intervals according to that suitable reference regular equal class interval is also divided proportionately. To fix the idea about unequal class interval, following examples can be considered.

[redacted]

Find out the mode or most repeated value from the following frequency distribution.

Class	5-10	10-15	15-20	20-25	25-35	35-45	45-50
Frequency	5	7	10	18	16	4	1

Solution:

Since, the given frequency distribution has unequal class width. In order to calculate mode, let us first convert the given unequal class intervals into the suitable class intervals having class width 5 for each class intervals.

Calculation of mode

Class	Frequency (f)
5-10	5
10-15	7
15-20	10
20-25	18
25-30	8
30-35	8
35-40	2
40-45	2
45-50	1

Since, the given distribution is regular and unimodal

And maximum frequency (f_1) = 18. So, modal class is 20-25

L = 20, h =5, f_o = 10, f_2 = 8

Here, Mode = $L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

$= 20 + \frac{18 - 10}{18 - 10 - 8} \times 5$

$= 20 + \frac{8}{18} \times 5 = 20 + 2.22 = 22.22$

[redacted]

If the mode for the following distribution is Rs 24, find the missing frequency corresponding to the class 30-40.

Expenditure per day (in Rs.)	0-10	10-20	20-30	30-40	40-50
No.of persons	14	23	27	?	15

Solution

Let f′ be the missing frequency corresponding to class 20-30

Since Mo = 24 lies in the class 20-30. So, modal class is 20- 30.

L = 20, f_1= 27, f_0=23, $f_2 = f'$, h =10

Mode = $L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

or, $24 = 20 + \frac{27-23}{2\times27-23-f'} \times 10$

or, $24\text{-}20 = \frac{40}{31-f'}$

or, $4 = \frac{40}{31-f'}$

or, $40 = 124 - 4f'$

or, $4f' = 84$

$\therefore f' = 21$

Hence, the required missing frequency is 21.

[redacted]

Modal marks for a group of 47students is 27. Five students got marks between 0 to 10, 15 students got marks between 20-30 and 7 students got marks between 40-50. Find the number of students getting marks between 10-20 and 30-40, if the maximum marks in the test were 50**.**

Solution:

Given, N = 47, Mode = 27

Let the number of students getting marks between 10-20 and 30-40 are respectively f_1 and f_2.

Calculation of missing frequencies

Marks	No. of students
0-10	5
10-20	f_1
20-30	15
30-40	f_2
40-50	7
	N = 47

Here, given mode = 27 which lies in the class 20-30, so, modal class is 20-30.

Here, f_1 = 15, $f_0 = f_1$, $f_2 = f_2$, L = 20, h = 10

We have,

$$\text{Mode} = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

or, $27 = 20 + \frac{15 - f_1}{2 \times 15 - f_1 - f_2} \times 10$

or, $7(30 - f_1 - f_2) = 150 - 10f_1$

or, $210 - 7f_1 - 7f_2 = 150 - 10f_1$

or, $3f_1 - 7f_2 = 150 - 210$

or, $7f_2 - 3f_1 = 60$ ...(i)

Again, we have, total no. of students = 47

$\therefore$ $5 + f_1 + 15 + f_2 + 7 = 47$

or, $f_1 + f_2 = 20$ ...(ii)

Solving (i) and (ii),

We get,

$f_2 = 12$

and $f_1 = 8$

$\therefore$ The required number of students obtaining marks between 10-20 and 30-40 are respectively 8 and 12.

Case Analysis:

A factory has two machines producing and packing particular types of Ice cream in suitable boxes. These boxes are specially designed to maintain cool temperature so that they comfortably reach the destinations, namely Ice Cream parlours. The process times to provide pack and ship the Ice Cream for these two units are given below:

Machine 1: 11, 11, 33, 22, 23, 43, 17, 17, 11, 23, 13, 15, 15, 16, 9, 21, 15, 19, 15, 18

Machine 2: 19, 23, 33, 25, 52, 31, 29, 26, 27, 20, 13, 25, 18, 25, 12, 5, 29, 11, 13, 19

Compute the Measure of central tendency (Arithmetic mean, Median and mode) for the two machines and interprets the result. Which shows better performance-Machine 1 or Machine 2?

Solution:

At first, arranging the given data in ascending order:

For Machine 1

Time (X): 9, 11, 11, 11, 13, 15, 15, 15, 15, 16, 17, 17, 18, 19, 21, 22, 23, 23, 33, 43

n = 20

$\sum X$ = 9+11+11+. . . . . +23+33+43 = 367

$\sum X^2 = 9^2 + 11^2 + 11^2 + \ldots\ldots + 23^2 + 33^2 + 43^2 = 7953$

Arithmetic mean $(\bar{X}) = \frac{\sum X}{n}$

$= \frac{367}{20}$

$= 18.35$

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

= Value of $\left(\frac{20+1}{2}\right)^{th}$ items

= Value of $(10.5)^{th}$ items

= Value of $\left(\frac{10^{th}+11^{th}}{2}\right)$ items

$= \frac{16+17}{2}$

$\therefore$ Md. = 16.5

Mode (M_o) = Value of variable X which repeats maximum number of times

= 15

$\because$ 15 repeats maximum number of times (i.e. 4 times)

For Machine 2

Time (X): 5, 11, 12, 13, 13, 18, 19, 19, 20, 23, 25, 25, 25, 26, 27, 29, 29, 31, 33, 52

n = 20

$\sum X$ = 5+11+12+. . . . . +31+33+52 = 455

$\sum X^2 = 5^2 + 11^2 + 12^2 + \ldots\ldots + 31^2 + 33^2 + 52^2 = 12319$

Arithmetic mean $(\bar{X}) = \frac{\sum X}{n}$

$= \frac{455}{20}$

$= 22.75$

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

= Value of $\left(\frac{20+1}{2}\right)^{th}$ items

= Value of $(10.5)^{th}$ items

= Value of $\left(\frac{10^{th}+11^{th}}{2}\right)$ items

$= \frac{23+25}{2}$

$\therefore$ Md. = 24

Mode (M_o) = Value of variable X which repeats maximum number of times

= 25

$\because$ 25 repeats maximum number of times (i.e. 3 times)

Since, for both machines 1 & 2 , $\overline{X} < Md < M_o$; therefore the distribution of the process times to provide pack and ship the Ice Cream for these two machines 1 and 2 are positively skewed.

To test better performance-Machines

For machine 1:

$\overline{X} = 18.35$

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

$= \sqrt{\frac{7953}{20} - \left(\frac{367}{20}\right)^2}$

$= 7.805$

Coefficient of variation, C.V. (Machine 1) $= \frac{\sigma}{\overline{X}} \times 100$

$= \frac{7.805}{18.35} \times 100$

$= 42.534\%$

For machine 2:

$\overline{X} = 22.75$

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

$= \sqrt{\frac{12319}{20} - \left(\frac{455}{20}\right)^2}$

$= 9.919$

Coefficient of variation, C.V. (Machine 2) $= \frac{\sigma}{\bar{X}} \times 100$

$= \frac{9.919}{22.75} \times 100$

$= 43.60\%$

Since, C.V. (Machine 1) < C.V. (Machine 2), therefore, machine 1 shows better performance than Machine 2.

Case Analysis:

The number of checks cashed each day at the five branches of NIC bank during the past month has the following distribution.

Class	0-200	200-400	400-600	600-800	800-1000
Frequency	10	13	17	42	18

The director of operations for the bank knows that a standard deviation in check cashing of more than 200 checks per day creates staffing and organizational problems at the branches because of the uneven workload.

i. Calculate mean, mode and standard deviation of the data?
ii. Should the director worry about staffing next month?

Solution:

Let a = 500

Class	Frequency (f)	mid.value (X)	$d' = \frac{X-500}{200}$	f d'	fd'^2
0-200	10	100	-2	-20	40
200-400	13	300	-1	-13	13
400-600	17	500	0	0	0
600-800	42	700	1	42	42
800-1000	18	900	2	36	72
	N = $\sum$ f = 100			$\sum$ fd′ = 45	$\sum f d'^2$ = 167

(i) Mean $(\bar{X})$ = a + $\frac{\sum f d\prime}{N} \times h$

$= 500 + \frac{45}{100} \times 200$

= 590 checks

For Mode (M_0)

Since, the given frequency distribution is regular and unimodal and maximum frequency is 42.
So, modal class is 600-800
L= 600, h = 200, f_1= 42, f_0= 17, f_2=18

Mode (Mo) = $L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$

$= 600 + \frac{42-17}{2\times42-17-18} \times 200$

$= 600 + \frac{25}{49} \times 200$

= 702.041 checks

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

$= \sqrt{\frac{167}{100} - \left(\frac{45}{100}\right)^2} \times 200$

$= \sqrt{1.4675} \times 200$

$= 1.2114 \times 200$

= 242.28 checks

Case Analysis:

Banks are permitted to sell a form of the life insurance called saving bank life insurance (SBLI), the approval process consists of underwriting, which include a review of the application, a medical information bureau check, possible request for additional medical and medical exams and a policy compilation stage during which the policy pages are generated and sent to the bank for delivery. The ability to deliver approved polices to customers in a timely manner is critical to the profitability of this service to the bank. During a period of one month, a random sample of 30 approved polices was selected' and the following total processing times in days were recorded; the data are:

73, 19 , 16, 64, 28, 28, 31, 90, 60, 56, 31, 56, 22, 18, 45, 48, 17, 17, 17, 17, 17, 17, 91, 92, 63, 50, 51, 69, 16, 17.

(i) Compute mean, mode, standard deviation.
(ii) Find median, first quartile and third quartile.

Solution:

Arranging the given data in ascending order:

Time (in days) X : 16, 16, 17, 17, 17, 17, 17, 17, 17, 18, 19, 22, 28, 28, 31, 31, 45, 48, 50, 51, 56, 56, 60, 63, 64, 69, 73, 90, 91, 92,

n = 30

$\sum X$ = 16+16+17+. . . . . +90+91+92 = 1236

$\sum X^2 = 16^2 + 16^2 + 17^2 + \ldots\ldots + 90^2 + 91^2 + 92^2 = 69496$

(i) Arithmetic mean $(\bar{X}) = \frac{\sum X}{n}$

$= \frac{1236}{300}$

= 41.2 days

Mode (M_o) = Value of variable X which repeats maximum number of times

= 17 days

∵ 17 repeats maximum number of times (i.e. 7 times)

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

$= \sqrt{\frac{69496}{30} - \left(\frac{1236}{30}\right)^2}$

= 24.881 days

(ii)

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

= Value of $\left(\frac{30+1}{2}\right)^{th}$ item

= Value of $(15.5)^{th}$ item

= Value of $\left(\frac{15^{th}+16^{th}}{2}\right)$ item

$= \frac{31+31}{2}$

∴ Md. = 31 days

First quartile (Q_1) = Value of $\left(\frac{n+1}{4}\right)^{th}$ item

= Value of $\left(\frac{30+1}{4}\right)^{th}$ item

= Value of $(7.75)^{th}$ item

= Value of 7^{th} item + 0.75 (8^{th} item-7^{th} item)

= 17 + 0.75 (17-17)

= 17 + 0.75× 0

= 17 days

Third quartile (Q_3) = Value of $3\left(\frac{n+1}{4}\right)^{th}$ item

= Value of $3\left(\frac{30+1}{4}\right)^{th}$ item

= Value of $(23.25)^{th}$ item

= Value of 23^{th} item + 0.25 (24^{th} item-23^{th} item)

= 60 + 0.25 (63-60)

= 60 + 0.25× 3

= 60.75 days

Simple Correlation and Regression Analysis

Correlation Analysis

Correlation analysis is the statistical tool which is used to describe the degree of relationship between two or more than two variables. Thus, correlation is a statistical tool or device which is used to measure the degree of association between two or more variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of relationship between two or more variables. To measure the degree of association between such variables, one more relative measure is needed and is known as correlation coefficient. It is generally denoted by 'r'. It is pure number and independent of units of measurement.

The correlation analysis provides only a quantitative measure. It does not necessarily indicate a cause and effect relationship Two variables are said to be correlated when the change in the value of one variable is accompanied with the change in the value of other variable. For example, change in volume of sales is associated with the change in advertisement expenditure, the change in the sales of woolen garment is related to the change in the temperature, change in the supply of a commodity is related to the change in the price of commodity etc.

Therefore, two or more variables are said to be correlated if change in the value of one variable is accompanied by the change in the value of other variable(s).

"*Correlation is defined as a statistical measure which is used to study the degree of relationship between* two *or more than two variables"*

Definition

Some important definitions are given below:

"Correlation is an analysis of co-variation between two or more variables" **A.M. Tuttle**

"*The measure of relationship between two or more variables is termed as correlation"*

Sir Francis Galton.

Importance

The study of correlation has an immense important in practical life due to the following reasons:

- Related to quantity of fertilizer used, types of soil, quality of seeds, amount of rainfall and so on. Correlation helps in quantifying precisely the degree and direction of such relationships.
- Correlation analysis contributes to the understanding of economic behaviour, aids in locating the critically important variables on which others depend, may reveal to the economist the connections by which disturbances spread and suggest to him the paths through which stabilizing forces may become effective. (W.A. Neiswanger)
- In economic theory and business studies we come across several types of variables which show some kind of relationship. For example, there exists a relationship between price, supply and quantity demanded; advertising expenditure and sales promotion measures; yield of crop
- The concepts of regression and ratio of variation are based on measure of correlation.
- Measures of correlation give us the more reliable prediction of variables. According to Tippett, "the effect of correlation is to reduce the range of uncertainty of our prediction".
- In most researches in social sciences, correlation analysis helps in arriving at very important conclusions.

8.3 Types of Correlation

Correlation may be classified as follows:

- Positive and Negative Correlation

- Linear and Non-linear Correlation
- Simple, Partial and Multiple Correlations

Positive and Negative Correlation

Two variables are said to be positive correlation if the values of the both variables deviate in the same direction i.e.correlation is said to be positive or direct if the increase (decrease) in the values of one variable results on an average, in a corresponding increase (decrease) in the values of other variable. e.g. price and supply of commodities, income and expenditure of a family etc.

On the other hand, correlation is said to be negative or inverse if the variables deviate in the opposite direction, i.e. if the increase (decrease) in the values of one variable results on the average, in a corresponding decrease (increase) in the values of other variables. Examples of such correlation are: the day temperature and sale of woolen dresses, price and demand of commodities etc.

The following examples illustrate the concept of positive and negative correlation.

1. Positive Correlation

i.	Increasing ↑	X:	17	20	25	30	34
	Increasing ↑	Y:	8	12	15	18	22
ii.	Decreasing ↓	X:	60	51	40	35	30
	Decreasing ↓	Y:	18	17	10	7	5

2. Negative Correlation

i.	Increasing ↑	X:	17	20	25	30	34
	Decreasing ↓	Y:	22	18	15	12	8
ii.	Decreasing ↓	X:	60	51	40	35	30
	Increasing ↑	Y	5	7	10	17	18

It may be noted here that the change (increasing or decreasing) in the values of both variables is not proportional or fixed.

Linear and Non-Linear Correlation

The correlation between two variables is said to be linear if corresponding to a unit change in one variable, there is a constant change in the other variable over the entire range of the values. Following example illustrates a linear correlation between two variables X and Y.

X:	1	2	3	4	5
Y:	5	7	9	11	13

Thus, for a unit change in the value of X, there is constant change 2 in the corresponding values of Y. when these pairs of values X and Y are plotted on a graph paper; the line joining these points would be a straight line.

The correlation between two variables is said to be non linear or curvilinear if corresponding to a unit change in one variable, the other variable does not change at a constant rate over the entire range of the values. The following example illustrates a non-linear correlation between two variables X and Y.

X:	1	2	3	4	5
Y:	7	8	15	20	23

We find the values of Y do not change at a constant ratio, if we plot the above values and graph paper we will get curve linear instead of a straight line.

Simple, Partial and Multiple Correlation

1. **Simple correlation:** The degree of relationship between only two variables is called simple correlation. e.g. (i) A study on the yield of crop with respect to only amount of fertilizer, (ii) sales revenue with respect to amount of money spent on advertisement.

2. **Partial correlation:** The degree of relationship (association) between the dependent variable and only one particular independent by keeping the effects of all other independent variables constant is called partial correlation. For example; the relationship between two variables: yield of crops production and amount of irrigation by keeping the fertilizer, quality of seeds etc constant is the study of partial correlation.

3. **Multiple correlation:** The degree of association (or relationship) between one dependent variable and joint (or combined) effect of independent variables is called multiple correlation coefficient. For example; the relationship between dependent variable yield of crops and combined effect of independent variables fertilizer, irrigation and seed quality is the multiple correlation.

Methods of Studying Simple Correlation

The commonly used methods for studying correlation (linear relationship) between two variables are:

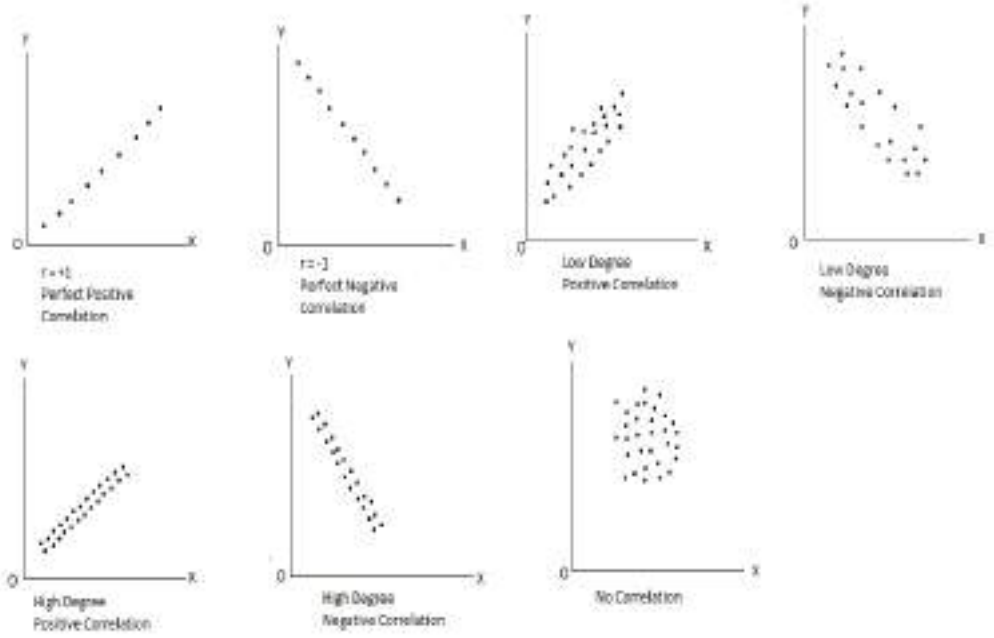
- Scatter Diagram Method or Graphic method.
- Karl Pearson's Coefficient of Correlation (Covariance method)
- Bivariate Correlation Method (Two way frequency table)
- Spearman's Rank Correlation Method

Scatter Diagram (Graphic method)

Scatter diagram is one of the simplest diagrammatic representations of bivariate distribution consisting of dots or points. It is a graphical method of studying the correlation between two variables. In this method, the points are represented by dots by keeping values of one variable on X-axis and the values of other variable on Y-axis in the XY-plane. Then, the diagram of dots so obtained is called scatter diagram. Scatter diagram is used to observe the existence of correlation betwween two variables. The following are the different types of pattern in scatter diagram of a bivariate data

- The pattern of points on a scatter diagram reveals on upward trend rising from bottom left to top right. The variables in this case shows positive correlation between them.
- The pattern of points on a scatter diagram reveals a downward trend falling from top left to bottom right. The variables in this case show the negative correlation between them.
- When we can not trace any trend, there is no correlation between them.

The following are some scatter diagrams depicting different types of correlation between two variables.



Observations on Scatter Diagram

- If the points on the scatter diagram show a very little spread, then the fair good amount of correlation can be expected between the two variables if the points on scatter diagram show a widely spread, then the poor correlation may be expected.
- The scatter diagram provides a rough measure of correlation only. It gives the direction and degree of correlation between the two variables.
- It enables us to locate the line of best fit.

Merits:

- It is the simplest method of measuring correlation.
- It is least affected by the extreme observations.
- It can be easily understood by non –statistician.
- If helps us to detect abnormal variates in the data.

Demerits:

- It gives only rough idea.
- It can not be numerically expressed.
- It is not amenable to algebraic treatment.

Karl Pearson's Coefficient of Correlation (Covariance Method)

Karl Pearson's developed a widely used mathematical formula to measure the degree (intensity) of linear relationship between two variables is called Karl Pearson's correlation coefficient. It is also called product moment correlation coefficient or simple correlation coefficient. This method is based on the linear relationship between two variables (Series). Karl Pearson's correlation coefficient is specially useful when data are quantitatively measured.

According to him, correlation coefficient between two variables X & Y is denoted by r(X, Y) or r_{XY} or simply r and is defined as the ratio of the covariance between them to the product of their corresponding standard deviations.

∴ Coefficient of correlation, $r = \frac{COV(X.Y)}{\sqrt{V(X)}\ \sqrt{V(Y)}} = \frac{COV(X.Y)}{\sigma_X \sigma_Y}$ ..... (i)

Where,

Cov (X, Y) = Covariance between two variables X & Y

$= \frac{1}{n}\sum(X - \bar{X})\ (Y - \bar{Y})$

Covariance measures the simultaneous changes between two variables

V(X) = Variance of X = $\sigma_X{}^2 = \frac{1}{n}\sum(X - \bar{X})^2$

Standard deviation of x = $\sigma_X = \sqrt{\frac{1}{n}\sum(X - \bar{X})^2}$

V(Y) = Variance of Y = $\sigma_Y{}^2 = \frac{1}{n}\sum(Y - \bar{Y})^2$

Standard deviation of Y = $\sigma_Y = \sqrt{\frac{1}{n}\sum(Y - \bar{Y})^2}$

Substituting Cov(X, Y), σ_X and σ_Y in (i) we have,

$$r = \frac{\frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\frac{1}{n}\sum(X-\bar{X})^2}\sqrt{\frac{1}{n}\sum(Y-\bar{Y})^2}}$$

or, $$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}} \quad \text{..... (ii)}$$

On simplification, we get

$$r = \frac{n\sum XY-(\sum X)(\sum Y)}{\sqrt{n\sum X^2-(\sum X)^2}\sqrt{n\sum Y^2-(\sum Y)^2}} \quad \text{..... (iii)}$$

This method is also known as direct method.

Different formulae for calculating Karl Pearson's Coefficient of Correlation

1. We have,

$$r = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum(X-\bar{X})^2}\sqrt{\sum(Y-\bar{Y})^2}} = \frac{\sum xy}{\sqrt{\sum x^2}\sqrt{y^2}} \text{ ..... (iv)}$$

This method is known as Actual mean method.

Where, x and y denote the deviations of X and Y from their arithmetic means $\bar{X}$ and $\bar{Y}$ respectively, i.e., $x = X - \bar{X}$, $y = Y - \bar{Y}$.

2. When actual means of X and Y are in fraction, the calculation of Pearson's correlation coefficient can be simplified by taking deviations of X and Y values from their assumed means a and b respectively. That is U = X – a and V = Y – a, when a and b are assumed means of X and Y values. The formula (iii) becomes as given below and known as short cut method.

Short cut method or Assumed mean method (i.e. Change of origin)

$$r = \frac{n\sum UV-(\sum U)(\sum V)}{\sqrt{n\sum U^2-(\sum U)^2}\sqrt{n\sum V^2-(\sum V)^2}} \quad \text{..... (v)}$$

Where, U = X – a, V = Y - b

a, b = Assumed means

3. Step deviation method (i.e. Change of origin & scale)

$$r = \frac{n\sum U'V'-(\sum U')(\sum V')}{\sqrt{n\sum U'^2-(\sum U')^2}\sqrt{n\sum V'^2-(\sum V')^2}} \quad \text{..... (vi)}$$

Where $U' = \frac{X-a}{h}$, $V' = \frac{Y-b}{k}$

h = common factor (or class size) for variable X, n = No. of pair observations.

k = common factor (or class size) for variable Y

Interpretation of Correlation Coefficient

Coefficient of correlation lies between – 1 to + 1 ($-1 \leq r \leq +1$)

SThe degrees of correlation coefficient according to Karl Pearson's formula's are as follows:

Degree of Correlation	**Direction**	
	Positive	**Negative**
Perfect correlation	+1	–1
Very high degree of correlation	+ 0.9 or more	–0.9 to – 0.99
High degree of correlation	+ 0.75 to + 0.9	– 0.75 to – 0.9
Moderate degree of correlation	+ 0.50 to + 0.75	– 0.50 to – 0.75
Low degree of correlation	+ 0.25 to + 0.50	– 0.25 to – 0.50
Very low degree of correlation	less than + 0.25	less than – 0.25

No correlation	0	0

Worked Out Examples

Example 1

Following are the marks in the statistics and Economics of six students. Calculate correlation coefficient between them.

Marks in Statistucs (X)	26	24	24	27	25	23
Marks in Economics (Y)	13	12	14	16	15	11

Solution:

Marks in Statistics (X)	Marks in Economics (Y)	XY	X^2	Y^2
26	13	338	676	169
24	12	288	576	144
24	14	336	576	196
27	16	432	729	256
25	15	375	625	225
23	11	253	529	121
$\Sigma X = 149$	$\Sigma Y = 81$	$\Sigma XY = 2022$	$\Sigma X^2 = 3711$	$\Sigma Y^2 = 1111$

The Karl Pearson's correlation coefficient is

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{6\times 2022 - 149\times 81}{\sqrt{6\times 3711 - (149)^2}\sqrt{6\times 1111 - (81)^2}}$$

$$= \frac{63}{\sqrt{65}\sqrt{105}}$$

$$= 0.672$$

There is high degree positive correlation between two variables marks in Statistics (X) and marks in Economics (Y)

Example 2

Calculate the co-variance from the following:

x: 7 9 11 13 15 17

f: 2 3 5 9 6 5

Solution:

Let f = Y

X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
7	2	-5	-3	15
9	3	-3	-2	6
11	5	-1	0	0
13	9	1	4	4
15	6	3	1	3
17	5	5	0	0

$\sum X = 72$	$\sum Y = 30$			$\sum(X-\bar{X})(Y-\bar{Y}) = 28$

Mean $(\bar{X}) = \frac{\sum X}{n} = \frac{72}{6} = 12$, Mean $(\bar{Y}) = \frac{\sum Y}{n} = \frac{30}{6} = 5$

∴ Cov (X, Y) = Covariance between two variables X & Y

$= \frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y})$

$= \frac{28}{6}$

= 4. 667

Example 3

Calculate the co-variance from the following.

X	10-20	20-30	30-40	40-50	50-60	60-70
f	7	3	5	9	6	5

Solution:

Let f = Y

Class	mid. value(X)	Y	$(X-\bar{X})$	$(Y-\bar{Y})$	$(X-\bar{X})(Y-\bar{Y})$
10-20	15	7	-25	-1.167	29.175
20-30	25	3	-15	-2.833	42.495
30-40	35	5	-5	-0.833	4.165
40-50	45	9	5	3.167	15.835
50-60	55	6	15	0.167	2.505
60-70	65	5	25	-0.833	-20.825
	$\sum X = 240$	$\sum Y = 35$			$\sum(X-\bar{X})(Y-\bar{Y}) =$ 73.35

Mean $(\bar{X}) = \frac{\sum X}{n} = \frac{240}{6} = 40$, Mean $(\bar{Y}) = \frac{\sum Y}{n} = \frac{35}{6} = 5.833$

∴ Cov (X, Y) = Covariance between two variables X & Y

$= \frac{1}{n}\sum(X-\bar{X})(Y-\bar{Y})$

$= \frac{73.35}{6}$

= 12.225

Example 4

Find the Karl Pearson's correlation coefficient for the following data:

X:	2	4	6	8	10
Y:	6	8	9	10	12

Use: a) Direct method

b) Deviation from assumed mean method

Solution: a) Direct method

X	Y	XY	X^2	Y^2
2	6	12	4	36
4	8	32	15	64
6	9	54	36	81

8	10	80	64	100
10	12	120	100	144
$\Sigma X = 30$	$\Sigma Y = 45$	$\Sigma XY = 298$	$\Sigma X^2 = 220$	$\Sigma Y^2 = 425$

The Karl Pearson's correlation coefficient is

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}} = \frac{5(298) - (30)(45)}{\sqrt{5(220) - (30)^2}\sqrt{5(425) - (45)^2}} = 0.99.$$

There is very high degree positive correlation between two variables X & Y.

OR

b) Deviation from assumed mean method

$a = 4, \quad b = 10$

X	Y	$U = X - 4$	$V = Y - 10$	UV	U^2	V^2
2	6	–2	–4	8	4	16
4	8	0	–2	0	0	4
6	9	2	–1	–2	4	1
8	10	4	0	0	16	0
10	12	6	2	12	36	4
		$\Sigma U = 10$	$\Sigma V = -5$	$\Sigma UV = 18$	$\Sigma U^2 = 60$	$\Sigma V^2 = 25$

$$r = \frac{n\sum UV - (\sum U)(\sum V)}{\sqrt{n\sum U^2 - (\sum U)^2}\sqrt{n\sum V^2 - (\sum V)^2}}$$

$$= \frac{5(18) - (10)(-5)}{\sqrt{5(60) - (10)^2}\sqrt{5(25) - (-5)^2}} = 0.99$$

There is very high degree positive correlation between X & Y.

Example 5

Compute the product moment correlation coefficient or Karl Pearson's correlation coefficient for the following:

Height of father	67	65	68	64	66
Height of son	69	65	69	65	67

Solution:

Let X be a variable of father's height and Y be the variable of son's height

Computation of Correlation Coefficient

Let $a = 66, \quad b = 67$

Height of father (X)	Heigh of son (Y)	$U = X - 66$	$V = Y - 67$	U^2	V^2	UV
67	69	1	+ 2	1	4	2
65	65	–1	–2	1	4	2
68	69	2	2	4	4	4
64	65	–2	–2	4	4	4
66	67	0	0	0	0	0
		$\Sigma U = 0$	$\Sigma V = 0$	$\Sigma U^2 = 10$	$\Sigma V^2 = 16$	$\Sigma UV = 12$

Karl Pearson's correlation coefficient is

$$r = \frac{n\sum UV - (\sum U)(\sum V)}{\sqrt{n\sum U^2 - (\sum U)^2}\sqrt{n\sum V^2 - (\sum V)^2}}$$

$$= \frac{5\times12 - 0\times0}{\sqrt{5\times10 - (0)^2}\sqrt{5\times16 - (0)^2}} = 0.948$$

There is very high degree positive correlation between X & Y.

Example 6

Calculate the Karl Pearson's correlation coefficient for the following data of sales and expenses in thousands of rupees of 5 firms.

Sales	43	41	36	34	50
Expenses	12	24	15	21	19

Solution:

Let X be the variable of sales and Y be the variable of expenses in '000' Rs.

Computation of Correlation Coefficient

Sales (X)	Deviations from assume mean U = X – 41	U^2	Expenses (Y)	Deviations from assume mean V = Y – 19	V^2	Product of deviation U.V
43	2	4	12	– 7	49	– 14
41	0	0	24	5	25	0
36	–5	25	15	–4	16	20
34	–7	49	21	2	4	–14
50	9	81	19	0	0	0
	$\Sigma U = -5$	$\Sigma U^2 = 159$		$\Sigma V = -4$	$\Sigma V^2 = 94$	$\Sigma UV = -8$

Therefore, The Karl Pearson's correlation coefficient between sales (X) and expenses (Y) is given by

$$r = \frac{n\sum UV - (\sum U)(\sum V)}{\sqrt{n\sum U^2 - (\sum U)^2}\sqrt{n\sum V^2 - (\sum V)^2}}$$

$$r = \frac{5 \times (-8) - (-5) \times (-4)}{\sqrt{5 \times 159 - (-5)^2} \cdot \sqrt{5 \times 94 - (-4)^2}}$$

$$= \frac{-40 - 20}{\sqrt{795 - 25} \cdot \sqrt{470 - 16}} = \frac{-60}{591.253}$$

$$= -0.10.$$

There is very low degree negative correlation between sales and expenses.

Example 7

Compute the coefficient of correlation from the following data:

No. of items (X):	200	270	340	360	400	300
No. of defective items (Y):	150	160	165	180	195	155

Solution: Computation of Correlation Coefficient

X	Y	$U' = \frac{X - 300}{10}$	$V' = \frac{Y - 165}{5}$	$(U')^2$	$(V')^2$	U'.V'
200	150	–10	–3	100	9	30
270	160	–3	–1	9	1	3
340	165	4	0	16	0	0
360	180	6	3	36	9	18
400	195	10	6	100	36	60
300	155	0	–2	0	4	0
		$\Sigma U' = 7$	$\Sigma V' = 3$	$\Sigma U'^2 = 261$	$\Sigma V'^2 = 59$	$\Sigma U'V' = 111$

The Karl Pearson's correlation coefficient between no. of items (X) and no. of defective items (Y) is given by

$$r = \frac{n\sum U'V' - (\sum U')(\sum V')}{\sqrt{n\sum U'^2 - (\sum U')^2}\sqrt{n\sum V'^2 - (\sum V')^2}}$$

$$= \frac{6 \times 11 - 7 \times 3}{\sqrt{6 \times 261 - 7^2} \cdot \sqrt{6 \times 59 - 3^2}}$$

$$= \frac{666 - 21}{\sqrt{1566 - 49} \cdot \sqrt{354 - 9}}$$

$$= \frac{645}{\sqrt{1517} \cdot \sqrt{345}} = \frac{645}{723.44} = 0.89.$$

There is high degree positive correlation between no .of items and no. of defective items.

Example 8

Findout Karl Pearson's coefficient of correlation between the values of X and Y from the following data:

X	100	110	115	116	120	125	130	135
Y	18	18	17	16	16	25	13	10

Solution: Computation of Correlation Coefficient

X	Y	U = X – 120	V = Y – 16	U^2	V^2	U.V
100	18	–20	2	400	4	–40
110	18	–10	2	100	4	–20
115	17	–5	1	25	1	–5
116	16	–4	0	16	0	0
120	16	0	0	0	0	0
125	25	5	9	25	81	45
130	13	10	–3	100	9	–30
135	10	15	–6	225	36	–90
		$\Sigma U = -9$	$\Sigma V = 5$	$\Sigma U^2 = 891$	$\Sigma V^2 = 135$	$\Sigma U.V = -140$

Correlation coefficient between two variables X and Y is given by

$$r = \frac{n\Sigma UV - (\Sigma U)(\Sigma V)}{\sqrt{n\Sigma U^2 - (\Sigma U)^2}\sqrt{n\Sigma V^2 - (\Sigma V)^2}}$$

$$= \frac{8 \text{ x } (-140) - (-9) \text{ x } 5}{\sqrt{8 \text{ x } 891 - (-9)^2} \cdot \sqrt{8 \text{ x } 135 - (-5)^2}}$$

$$= \frac{-1120 + 45}{\sqrt{7128 - 81} \cdot \sqrt{1080 - 25}} = \frac{-1075}{\sqrt{7047} \cdot \sqrt{1055}} = \frac{-1075}{2726.64} = -0.3942.$$

Hence, there is low degree negative correlation between X & Y.

Example 9

Calculate coefficient of correlation from the following data given below.

No. of orange (X):	10	15	20	25	30	40
No. of bad oranges (Y):	2	4	5	6	8	11

Solution:

Computation of Correlation Coefficient

X	Y	$U' = \frac{X - 20}{5}$	$V' = \frac{Y - 6}{1}$	U'^2	V'^2	U'V'
10	2	–2	–4	4	16	8
15	4	–1	–2	1	4	2
20	5	0	–1	0	1	0
25	6	1	0	1	0	0
30	8	2	2	4	4	4

40	11	4	5	16	25	20
		$\Sigma U' = 4$	$\Sigma V' = 0$	$\Sigma U'^2 = 26$	$\Sigma V'^2 = 50$	$\Sigma U'V' = 34$

Karl Pearson's correlation coefficient between two variables no. of orange (X) and no. of bad orange (Y) is

$$r = \frac{n\Sigma U'V' - (\Sigma U')(\Sigma V')}{\sqrt{n\Sigma U'^2 - (\Sigma U')^2}\sqrt{n\Sigma V'^2 - (\Sigma V')^2}}$$

$$= \frac{6 \times 34 - 4 \times 0}{\sqrt{6 \times 26 - 4^2} \cdot \sqrt{6 \times 50 - 0^2}}$$

$$= \frac{204 - 0}{\sqrt{156 - 16} \cdot \sqrt{300 - 0}} = \frac{204}{204.94} = 0.995.$$

Hence, there is very high degree positive correlation between no. of orange and no. of bad oranges.

Example 10

The following are the monthly figures of advertising expenditure and sales of a firm. It is generally found that advertising expenditure has its impact on sales generally after 2 months. Allowing for this time-lag calculate coefficient of correlation.

Month	Advertising Expenditure (Rs. Lakhs)	Sales (Rs. Lakhs)
Jan	50	1200
Feb	60	1500
Mar	70	1600
Apr	90	2s000
May	120	2200
Jun	150	2500
Jul	140	2400
Aug	160	2600
Sep	170	2800
Oct	190	2900
Nov	200	3100
Dec	150	3900

Solution:

Allow for a time-lag of 2 months i.e. link advertising expenditure of January with sales for March and so on.

Let x and Y be two variables of advertising expenditure (in lakhs) and sales (in lakhs) respectively.

Computation of Correlation Coefficient

Months	X	Y	$U' = \frac{X - 120}{10}$	$V' = \frac{Y - 2600}{100}$	U'^2	V'^2	U'.V'
Jan	50	1600	–7	–10	49	100	70
Feb	60	2000	–6	–6	36	36	36
Mar	70	2200	–5	–4	15	16	20
Apr	90	2500	–3	–1	9	1	3
May	120	2400	0	–2	0	4	0
Jun	150	2600	3	0	9	0	0
Jul	140	2800	2	2	4	4	4
Aug	160	2900	4	3	16	9	12
Sep	170	3100	5	5	25	25	25
Oct	190	3900	7	13	49	169	91
					$\Sigma U'^2 = 222$	$\Sigma V'^2 = 364$	$\Sigma U'.V' = 261$

The correlation coefficient is

$$r = \frac{n\sum U'V' - (\sum U')(\sum V')}{\sqrt{n\sum U'^2 - (\sum U')^2}\sqrt{n\sum V'^2 - (\sum V')^2}}$$

$$= \frac{10 \times 261 - 0 \times 0}{\sqrt{10 \times 222 - 0} \cdot \sqrt{10 \times 364 - 0}}$$

$$= \frac{2610}{\sqrt{2220} \cdot \sqrt{3640}} = \frac{2610}{2842.67} = 0.92$$

There is high degree positive correlation between advertising expenditure and sales.

Example 11

If sample size n is 50, variance of X is 9, S.D. of Y is 4, then covariance is 9.8, find Karl Pearson's correlation coefficient.

Solution:

We have,

n = 50, variance of X = $\sigma_x^2 = 9$, S.D. of X = $\sigma_x = 3$, S.D. (Y) = $\sigma_y = 4$

Cov (X, Y) = 9.8

Correlation coefficient (r) =?

Now,

$$r = \frac{Cov(X,Y)}{\sigma_x \cdot \sigma_y} = \frac{9.8}{3 \times 4} = 0.82$$

There is high degree positive correlation between two variables X & Y.

Example 12

Compute the correlation coefficient of correlation. From the following results are obtained between two variables.

	Variable X	Variable Y
Number of sets	7	7
Arithmetic mean	4	8
Sum of squares of deviations from arithmetic mean	28	76

Summation of products of deviation of variables X and Y from this respective mean 46

Solution: In the usual notations, we have given

n = 7, $\overline{X} = 4$, $\overline{Y} = 8$,

$\Sigma(X - \overline{X})^2 = \Sigma x^2 = 28$,

$\Sigma(Y - \overline{Y})^2 = \Sigma y^2 = 76$

$\Sigma(X - \overline{X})\,\Sigma(Y - \overline{Y}) = \Sigma xy = 46$

Then, Karl Pearson's correlation coefficient between x-series and y-series is given by:

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2} \cdot \sqrt{\Sigma y^2}} = \frac{46}{\sqrt{28} \cdot \sqrt{76}} = 0.997$$

There is very high degree positive correlation between two variables X& Y.

Example 13

If r = 0.5, $\Sigma xy = 120$, $\sigma_y = 8$ and $\Sigma x^2 = 90$, find the number of items. Where x and y are deviation from their respective means**.**

Solution:

We have given,

r = 0.5, $\Sigma xy = 120$, $\sigma_y = 8$, $\Sigma x^2 = 90$ $\because x = X - \overline{X}$ & $y = Y - \overline{Y}$

$$\sqrt{\frac{1}{n}\Sigma(Y - \overline{Y})^2} = \sigma_y$$

$$\Rightarrow \quad \sqrt{\frac{1}{n}\Sigma(Y-\overline{Y})^2} = 8 \quad \Rightarrow \sqrt{\frac{1}{n}\Sigma y^2} = 8 \Rightarrow \frac{1}{n}.\Sigma y^2 = 64$$

$$\Rightarrow \quad \Sigma y^2 = 64n$$

Number of items (n) =?

Then Karl Pearson's correlation coefficient between variables X and Y is given by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}.\sqrt{\Sigma y^2}} \quad \Rightarrow 0.5 = \frac{120}{\sqrt{90}.\sqrt{64n}} \quad \Rightarrow (0.5)^2 = \frac{(120)^2}{90 \times 64 \times n}$$

$$\Rightarrow \quad n = \frac{(120)^2}{90 \times 64 \times (0.5)^2} = \frac{14400}{5760 \times 0.25} = 10.$$

Example 14

Find the Karl Pearson's correlation coefficient between two variables X and Y from 12 pairs of observations,

$\Sigma X = 30, \Sigma Y = 5, \Sigma X^2 = 670, \Sigma Y^2 = 285, \Sigma XY = 383$

Solution:

We have given,

Paris of observation (n) = 12

$\Sigma X = 30, \Sigma Y = 5, \Sigma X^2 = 670, \Sigma Y^2 = 285, \Sigma XY = 385$

Correlation coefficient between two variables X and Y is

$$r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{12 \times 385 - 30 \times 5}{\sqrt{12 \times 670 - 30^2}.\sqrt{12 \times 285 - 5^2}}$$

$$= \frac{4620 - 150}{\sqrt{7140}.\sqrt{3395}} = \frac{4470}{4923.44} = 0.91$$

There is Very high degree positive correlation between X & Y.

Example 15

Compute the Karl Pearson's coefficient of correlation from the following data by the Karl Pearson's method

Price A	25	28	35	20	22	30	31	22
Price B	35	39	48	29	30	38	40	32

Also

(a) Calculate its probable error

(b) Interpret if the value of r is significant or not

(c) Determine the limits within which the population correlation coefficient may be expected to lie.

Solution:

Let X be the price of tea and Y be the price of coffee in Rupees.

Computation of Correlation Coefficient

X	Y	U = X – 28	V = Y – 38	U^2	V^2	UV
25	35	–3	–3	9	9	9
28	39	0	1	0	1	0
35	48	7	10	49	100	70
20	29	–8	–9	64	81	72
22	30	–6	–8	36	64	48
30	38	2	0	4	0	0
31	40	3	2	9	4	6

22	32	– 6	–6	36	36	36
		$\Sigma U = -11$	$\Sigma V = -13$	$\Sigma U^2 = 207$	$\Sigma V^2 = 295$	$\Sigma UV = 241$

Karl Pearson's correlation of coefficient is

$$r = \frac{n\sum U'V' - (\sum U')(\sum V')}{\sqrt{n\sum U'^2 - (\sum U')^2}\sqrt{n\sum V'^2 - (\sum V')^2}}$$

$$= \frac{8 \times 241 - (-11) \times (-13)}{\sqrt{8 \times 207 - (-11)^2} \cdot \sqrt{8 \times 295 - (-13)^2}}$$

$$= \frac{1928 - 143}{\sqrt{1535} \cdot \sqrt{2191}} = 0.9733$$

a) Probable error of correlation coefficient is given by

$$P.E.(r) = 0.6745 \times \frac{1 - r^3}{\sqrt{n}} = 0.6745 \times \frac{1 - (0.9733)^2}{\sqrt{8}} = 0.0125$$

b) Significance of r

$6 \times P.E.(r) = 6 \times 0.0125 = 0.0753$

Since , r is much greater than $6 \times P.E.(r)$, the value of r is highly significant.

c) Limit of population correlation coefficient

$r \pm 6 \times P.E.(r)$

$= 0.9733 \pm 0.0753 = (0.9733 - 0.0753, 0.9733 + 0.0753)$

$= (0.8980, 1.048) = (0.8980, 1.0)$

Example 16

A computer while calculating the correlation coefficient between two variants X and Y from 25 pairs of observation obtained the following information's:

$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508$

It was however, discovered later at the time of checking that it had copied down two pairs of wrong observations as

X	Y
6	14
8	6

While the correct values were

X	Y
8	12
6	8

Obtain correct value of the correlation coefficient between x and y.

Solution: We have given,

$n = 25, \Sigma X = 125, \Sigma X^2 = 650, \Sigma Y = 100, \Sigma Y^2 = 460, \Sigma XY = 508$

Now,

Corrected $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$

$$\text{Corrected } r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2}\sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2} \cdot \sqrt{25 \times 436 - (100)^2}}$$

$$= \frac{500}{\sqrt{325} \cdot \sqrt{900}} = \frac{500}{750} = 0.67$$

There is moderate positive correlation between X & Y.

Regression Analysis

Introduction

The literal or dictionary meaning of the word "Regression" is "Stepping back" or "Returning back" towards the average value. The concept of regression was first developed by a British biometrician Sir Francis Galton in 1877 (1822-1911) in report of his research on heredity. He made a study and found that the off-springs having either short or tall parents tend to "regress" or "step back" towards the average height of general population. But the term "regression" as now used in Statistics is only a convenient term without having any reference to biometry. Nowadays, in Statistics the concept of regression analysis is applicable to all those fields where two or more variables have tendency to move back to the average behaviour.

Regression analysis is a mathematical measure (method) of the average relationship between two or more of variables in terms of original units of data. In other words, the statistical method (device) that helps to formulate an algebraic relationship between two or more variables in the form of equation to predict or estimate value of one variable, given another variable in terms of original units of data, is called regression analysis. Therefore, the cause and effect relationship is clearly indicated through regression analysis. Thus, regression is a statistical tool (device) which is used to estimate or predict one variable's value from the given value of other variables. For example, the yield of a crop depends on the amount of rainfall, quantity of fertilizer used, quality of seeds, types of soil etc. Linear regression is the most basic and commonly used predictive analysis.

Definition: *Regression analysis is a statistical technique of studying the functional relationship between one dependent variable and independent variables to predict the average value of dependent variable with help of known values of independent variables in terms of original units of data.*

There are two types of variables involved in regression analysis:

(i) Dependent variable
(ii) Independent variable

Dependent Variable: The variable whose value is to be predicted or estimated is called dependent variable. The dependent variable is also called explained variable or response variable or outcome variable or Regressand variable or predictand variable.

Independent Variable: The variable which is used for prediction is called independent variable. The variable which influences the value of dependent variable is known as independent variable. In other words, an independent variable is a variable that is manipulated to determine the value of a dependent variable. The independent variable is also called explanatory variable or predictor or regressor or covariate.

Galton studied the average relationship between these two variables graphically and called the line of regression.

Uses of Regression Analysis

Regression Analysis is the most useful Statistical tool which can be used in both natural as well as social sciences. Some uses are given below:

- In social and economic field, it is used for projection of population, birth rates, death rates, marital status, planning etc.
- In the business field, it is widely used. For example, businessmen are interested to predict the future production, consumption, investment, prices, profit, sales etc.
- It can be used to estimate unknown value of a variable (dependent variable) from the given value of other variable (independent variable) which are interrelated.

- Regression analysis helps to explore cause and effect relationship between variables.
- It also helps to determine correlation coefficient between the variables under study.

Objectives of Regression Analysis

The main objectives of regression analysis are as follows:

(i) To predict or estimate the value of dependent variable corresponding to given values of independent variables.

(ii) To measure the effect of independent variables on the dependent variable i.e. to measure the cause and effect relationship between the variables.

(iii) To measure the coefficient of determination or the proportion of variation in the dependent variable which is explained by the independent variable(s).

Types of Regression

(i) **Simple Regression**: Regression analysis between only two variables (i.e. one is dependent and one is independent variable) is known as simple regression. For example, the yield of crop depends on the amount of rainfall; profit depends on the sales etc.

(ii) **Multiple Regression:** Regression analysis between more than two variables (i.e. one is dependent variable and others are independent variables) is known as multiple regression. For example, the yield of crop depends on the amount of rainfall, types of seed, quality of fertilizers used, irrigation, soil quality etc.

Simple Regression Analysis

The regression analysis confirmed to the study of only two variables in which one is dependent and other is independent at a time is called simple regression. In other words, if there are only two variables under consideration, then the regression is called simple regression. For examples, the study of regression between income and expenditure, profit and sales of a product, the yield of a crop and the amount of rainfall etc.

Lines (or Equations of regression)

Galton studied the average relationship between these two variables graphically and called the line of regression.

Line of regression is the line which gives the best estimate of one variable for any given value of the other variable. In case of simple regression only two variables X & Y are studied. If X and Y are two variables, the algebraic expressions of regression equations are called regression equations (lines)

Depending on the number of variables, there are two regression lines (Equations)

(i) Regression equation of Y on X

(ii) Regression equation of X on Y

Regression Equation of Y on X:

The regression equation of Y on X which describes the variation in the values of dependent variable Y for given changes in independent variable X. So, the line of regression of Y on X is the line which gives the best estimate for the value of Y for any specified value of X. Thus, the regression equation of Y on X is used to estimate the value of Y when the value of X is given.

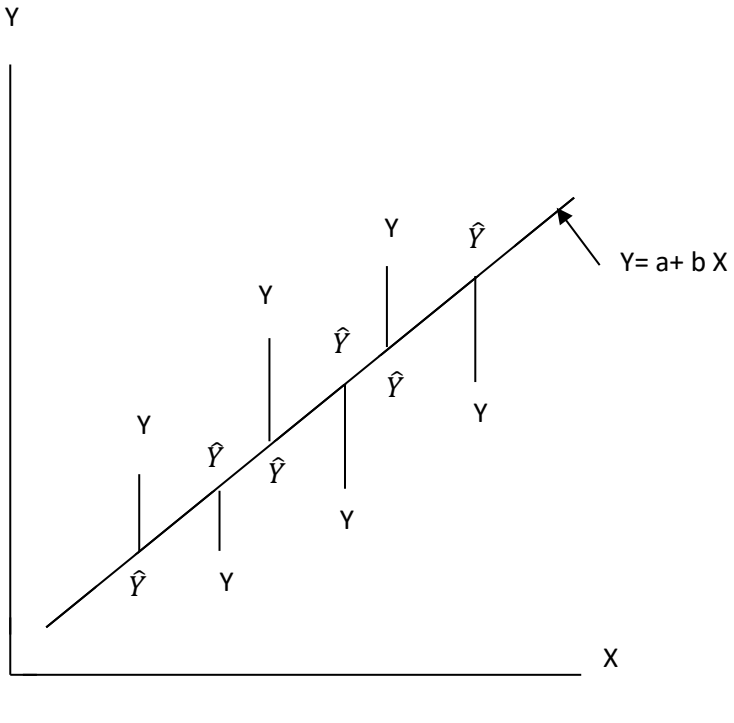
Since, the line of regression is the line of best fit. So, the term 'best fit' is interpreted in accordance with the principle of least square which consists in minimizing the sum of squares of the residuals or the errors of estimates i.e. the deviations between the given observed values of the variable and their Corresponding estimated values as given by the line of best fit

The sum of squares of the errors either parallel to Y-axis or parallel to X-axis may be minimized. The equation of the line of regression of Y on X is obtained by minimizing the sum of squares of errors parallel to Y- axis**.**

Let regression equation of Y on X be

$Y = a + bX$ ..... (i)

Where



Y = dependent variable (To find or estimate)

X = independent variable (Given or known)

a = Average value of Y when X = 0

= Intercept of the line (Y - intercept)

b = b_{YX} = Regression coefficient of Y on X.

= slope of line

= Rate of change in Y due to unit change in X or it measures the average change in the value of Y as per unit change in the value of X.

Applying principle of least square method, the normal equations for estimating a & b are

$\sum Y = n\,a + b\sum X$ ..... (ii)

$\sum XY = a\sum X + b\sum X^2$ ..... (iii)

Where n = no. of pairs of observations.

By solving equations (ii) and (iii), we get estimates of a and b. Substituting theses values in equation (i), we get the required equation of line or regression equation of Y on X i.e.

$\hat{Y} = \hat{a} + \hat{b}\,X$ ..... (iv)

This is called regression equation of Y on X. It is also called line of best fit or prediction equation or estimating equation or linear estimated equation or least square regression equation of Y on X.

Alternatively

The regression equation of Y on X is

$\mathbf{Y - \bar{Y} = b_{YX}\,(X - \bar{X})}$ .....(i)

Where

Y = dependent variable (to find or estimate)

X = independent variable (known)

$\bar{Y}$ = Average (mean) value of variable Y

$\bar{X}$ = Average (mean) value of variable X

b_{YX} = regression coefficient of Y on X.

= slope of line

= rate of change in Y due to unit change in X

Calculation of regression coefficient of Y on X (i.e. b_{YX}):

(i) Actual mean method

$b_{YX} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2}$

This method is appropriate when $\bar{X}$ & $\bar{Y}$ are whole numbers.

(ii) Direct method

$b_{YX} = \frac{n\sum XY-(\sum X)(\sum Y)}{n\sum X^2-(\sum X)^2}$

(iii) Short cut method or Assumed mean method (Change of origin)

$b_{YX} = \frac{n\sum UV-(\sum U)(\sum V)}{n\sum U^2-(\sum U)^2}$ where $U = X - A$ & $V = Y - B$

(iv) Step deviation method (i.e. Change of origin and scale)

$b_{YX} = \frac{n\sum U'V'-(\sum U')(\sum V')}{n\sum U'^2-(\sum U')^2} \times \frac{k}{h}$ where $U' = \frac{X-A}{h}$ & $V' = \frac{Y-B}{k}$

h and k are common factors or class sizes.

(v) If standard deviation (S.D.) and correlation coefficient (r) are given, then the regression coefficient

$b_{YX} = r\frac{\sigma_y}{\sigma_x}$ and $b_{xy} = r\frac{\sigma_x}{\sigma_y}$

where,

σ_x = S.D. of variable X.

σ_y = S.D. of variable Y.

r = correlation coefficient between two variable X and Y.

Example 17

From the following information find the regression equation of Y on X.

$\sum X = 21, \sum Y = 20, \sum X^2 = 91, \sum XY = 74$, n = 7. Also, estimate the value of Y on X = 3

Solution:

The regression equation of Y on X is $Y - \bar{Y} = b_{YX}(X - \bar{X})$ .....(i)

Where, $\bar{X} = \frac{\sum X}{n} = \frac{21}{7} = 3$, $\bar{Y} = \frac{\sum Y}{n} = \frac{20}{7} = 6.666$

$b_{YX} = \frac{n\sum XY-(\sum X)(\sum Y)}{n\sum X^2-(\sum X)^2}$

$= \frac{7\times74-21\times20}{7\times91-(21)^2} = 0.5$

From equation (i)

$Y - \bar{Y} = b_{YX}(X - \bar{X})$

or, $Y-6.666 = 0.5(X-3)$

or, $Y = 6.666 + 0.5X - 1.5$

or, $\hat{Y} = 5.166 + 0.5X$ is the required estimated regression equation of Y on X.

When X = 3

Y = ?

$\hat{Y} = 5.166 + 0.5X$

$= 5.166 + 0.5 \times 3$

$\hat{Y} = 6.666$

Hence, the estimated value of Y is 6.6666 when X = 3.

OR

Alternative method

The regression equation of Y on X is $\mathbf{Y = a + bX}$ . . . . . .(i)

Where, $\bar{X} = \frac{\sum X}{n} = \frac{21}{7} = 3$, $\bar{Y} = \frac{\sum Y}{n} = \frac{20}{7} = 6.666$

$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$

$= \frac{7 \times 74 - 21 \times 20}{7 \times 91 - (21)^2}$

$b = 0.5$

& $a = \bar{Y} - b\bar{X} = 6.666 - 0.5 \times 3$

$= 5.166$

From equation (i)

$Y = a + bX$

or, $Y = 5.166 + 0.5X$

∴ $\hat{Y} = 5.166 + 0.5X$ is the required estimated regression equation of Y on X.

When X = 3

Y = ?

$\hat{Y} = 5.166 + 0.5X$

$= 5.166 + 0.5 \times 3$

$\hat{Y} = 6.666$

Hence, the estimated value of Y is 6.6666 when X = 3.

Example 18

Estimate the marks in JAVA when the marks in Statistics is 65 by using following data:

Marks in Statistics	57	58	59	59	60	61	62	64
Marks in JAVA	77	78	75	78	82	8,2	79	81

Solution:

Here,

n = 6

Let X = Marks in Statistics

Y = Marks in JAVA (To find or estimate)

Let the regression equation of Y on X be

$Y = a + b X$. . . . . (i)

Calculation of sum of values

Marks in Statistics (X)	Marks in JAVA (Y)	XY	X^2
57	77	4389	5929
58	78	4524	6084
59	75	4425	5625
59	78	4602	6084
60	82	4920	6724
61	82	5002	6724
62	79	4898	6241
64	81	5184	6561
$\sum X = 480$	$\sum Y = 632$	$\sum XY = 37944$	$\sum X^2 = 49972$

$\overline{X} = \frac{\Sigma X}{n} = \frac{480}{8} = 60$

$\overline{Y} = \frac{\Sigma Y}{n} = \frac{632}{8} = 79$

Regression coefficient of Y on X:

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{8\times 37944 - 480\times 632}{8\times 49972 - (632)^2} = \frac{192}{288} = 0.667$$

$b = 0.667$

$\& \; a = \overline{Y} - b\,\overline{X} = 79 - (0.667) \times 60 = 38.98$

Substituting the value of a & b in equation (i), we get

$Y = a + b\,X$

$\hat{Y} = 38.98 + 0.667X$ is the required regression equation of Y on X .

When marks in Statistics X = 65, then

$\hat{Y} = 38.98 + 0.667X$

$= 38.98 + 0.667 \times 65$

$= 82.335$

∴ The estimated marks in JAVA 82.354 when marks in Statistics X = 65

OR

Alternative method

Solution:

n = 6

Let X = Marks in Statistics

Y = Marks in JAVA (To find or estimate)

Then, the regression line of Y on X is given line

$Y - \overline{Y} = b_{yx}\,(X - \overline{X})$ ..... (i)

Calculation of sum of values

Marks in Statistics (X)	Marks in JAVA (Y)	XY	X^2
57	77	4389	5929
58	78	4524	6084
59	75	4425	5625
59	78	4602	6084
60	82	4920	6724
61	82	5002	6724
62	79	4898	6241
64	81	5184	6561
$\sum X = 480$	$\sum Y = 632$	$\sum XY = 37944$	$\sum X^2 = 49972$

$\overline{X} = \frac{\Sigma X}{n} = \frac{480}{8} = 60$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{632}{8} = 79$$

Regression coefficient of Y on X:

$$b_{YX} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$$

$$= \frac{8\times 37944 - 480\times 632}{8\times 49972 - (632)^2} = \frac{192}{288} = 0.667$$

$b_{YX} = 0.667$

From equation (i)

The regression equation of Y on X is

$$Y - \overline{Y} = b_{yx}(X - \overline{X})$$

$\Rightarrow Y - 79 = 0.667 (X - 60)$

$\Rightarrow Y = 79 + 0.667X - 40.02$

$\Rightarrow Y = 38.98 + 0.667X$

$\Rightarrow \hat{Y} = 38.98 + 0.667\ X$ is the required regression equation of Y on X.

When marks in Statistics X = 65, then

$\hat{Y} = 38.98 + 0.667X$

$= 38.98 + 0.667 \times 65$

$= 82.335$

∴ The estimated marks in JAVA 82.335 when marks in Statistics X = 65

Example 19

The following table showing age of cars of a certain make and annual maintenance costs, contain the regression equation for cost related to age:

Age of cars (years):	2	4	6	8	10	12	14	16	18	20
Annual maintenance cost (000 Rs.):	3	5	7	9	11	13	15	18	23	25

Also, calculate the maintenance costs for the car 25 years old.

Solution:

Let X = Age of cars (in years)

Y = Annual maintenance cost (000 Rs.)

(To find or estimate)

The regression equation of Y on X is $\mathbf{Y - \overline{Y} = b_{YX}(X - \overline{X})}$ . . . . . .(i)

Age of cars (in years) X	Annual maintenance cost (000 Rs.)Y	XY	X^2
2	3	6	4
4	5	20	16
6	7	42	36
8	9	72	64
10	11	110	100
12	13	156	144
14	15	210	196
16	18	288	256
18	23	414	324
20	25	500	400
$\sum X = 110$	$\sum Y = 129$	$\sum XY = 1818$	$\sum X^2 = 1540$

$\overline{X} = \frac{\sum X}{n} = \frac{110}{10} = 11, \quad \overline{Y} = \frac{\sum Y}{n} = \frac{129}{10} = 12.9$

$b_{YX} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$

$= \frac{10\times1818 - 110\times129}{10\times1540 - (110)^2} = 1.209$

From equation (i)

$\mathbf{Y - \overline{Y} = b_{YX}(X - \overline{X})}$

or, $Y - 12.9 = 1.209\ (X - 11)$

or, $Y = 12.9 + 1.209\ X - 13.299$

or, $\hat{Y} = -0.399 + 1.209\ X$ is the required estimated regression equation of Y on X.

When age of cars X = 25 years

Annual maintenance cost (Y) = ?

$\hat{Y} = -0.399 + 1.209\ X$

$= -0.399 + 1.209 \times 25$

$\hat{Y} = 29.826$ (in 000 Rs.)

= Rs. 29826

Hence, the estimated Annual maintenance cost, Y is Rs. 29826 when age of cars, X = 25 years

OR

Alternative method

Let X = Age of cars (in years)

Y = Annual maintenance cost (000 Rs.)

(To find or estimate)

The regression equation of Y on X is $\mathbf{Y = a + bX}$ . . . . . .(i)

Where, $\overline{X} = \frac{\sum X}{n} = \frac{110}{10} = 11, \qquad \overline{Y} = \frac{\sum Y}{n} = \frac{129}{10} = 12.9$

$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$

$= \frac{10\times1818 - 110\times129}{10\times1540 - (110)^2}$

$b = 1.209$

& a = $\overline{Y} - b\,\overline{X} = 12.9 - 1.209 \times 11$

= - 0.399

From equation (i)

Y = a + bX

or, Y = - 0.399 + 1.209 X

∴ $\hat{Y}$ = - 0.399 + 1.209 X is the required estimated regression equation of Y on X.

When age of cars X = 25 years

Annual maintenance cost (Y) = ?

$\hat{Y}$ = - 0.399 + 1.209 X

= - 0.399 + 1.209× 25

$\hat{Y}$ = 29.826 (in 000 Rs.)

= Rs. 29826

Hence, the estimated Annual maintenance cost, Y is Rs. 29826 when age of cars, X = 25 years

Example 20

Compute the appropriate regression equation for the following data**:**

X (Independent variable)	2	4	5	6	8	11
Y (Dependent variable)	18	12	10	8	7	6

Find Y when X = 12

Solution: Here, Y = dependent variable (To find or estimate)

X = independent variable

n = 6

Let the regression equation of Y on X be

Y = a + b X ..... (i)

Calculation of sum of values

X	Y	XY	X^2
2	18	36	4
4	12	48	16
5	10	50	25
6	8	48	36
8	7	56	64
11	5	55	121
$\sum X = 36$	$\sum Y = 60$	$\sum XY = 293$	$\sum X^2 = 266$

$\overline{X} = \frac{\Sigma X}{n} = \frac{36}{6} = 6$

$\overline{Y} = \frac{\Sigma Y}{n} = \frac{60}{6} = 10$

Regression coefficient of Y on X:

$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2}$

$= \frac{6\times 293 - 36\times 60}{6\times 266 - (36)^2} = \frac{1758 - 2160}{1596 - 1296} = \frac{-402}{300}$

b = –1.34

& a = $\overline{Y} - b\,\overline{X} = 10 - (-1.34) \times 6$

= 18.04

Substituting the value of a & b in equation (i), we get

$Y = a + b X$

$\hat{Y} = 18.04 - 1.34X$ is the required regression equation of Y on X .

when X = 12, then

$\hat{Y} = 18.04 - 1.34 \times 12$

$= 18.04 - 16.08$

$= 1.96$

∴ The estimated values of Y is 1.96 when X = 12

OR

Alternative method

Solution:

Here, Y = dependent variable (To find or estimate)

X = independent variable

Then, the regression line of Y on X is given line

$Y - \overline{Y} = b_{yx} (X - \overline{X})$ ..... (i)

X	Y	XY	X^2
2	18	36	4
4	12	48	16
5	10	50	25
6	8	48	36
8	7	56	64
11	5	55	121
$\Sigma X = 36$	$\Sigma Y = 60$	$\Sigma XY = 293$	$\Sigma X^2 = 266$

Now,

$$\overline{X} = \frac{\Sigma X}{n} = \frac{36}{6} = 6$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{60}{6} = 10$$

Regression coefficient of Y on X:

$$b_{YX} = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{n\Sigma X^2 - (\Sigma X)^2}$$

$$= \frac{6 \times 293 - 36 \times 60}{6 \times 266 - (36)^2} = \frac{1758 - 2160}{1596 - 1296} = \frac{-402}{300} = -1.34$$

From equation (i)

The regression equation of Y on X is

$Y - \overline{Y} = b_{yx} (X - \overline{X})$

$\Rightarrow Y - 10 = -1.34 (X - 6) \Rightarrow Y = -1.34 X + 1.34 \times 6 + 10$

$\Rightarrow Y = -1.34X + 8.04 + 10$

$\Rightarrow \hat{Y} = 18.04 - 1.34X$ is the required regression equation of Y on X

when X = 12

Y =?

$\hat{Y} = 18.04 - 1.34 \times 12 = 1.96$

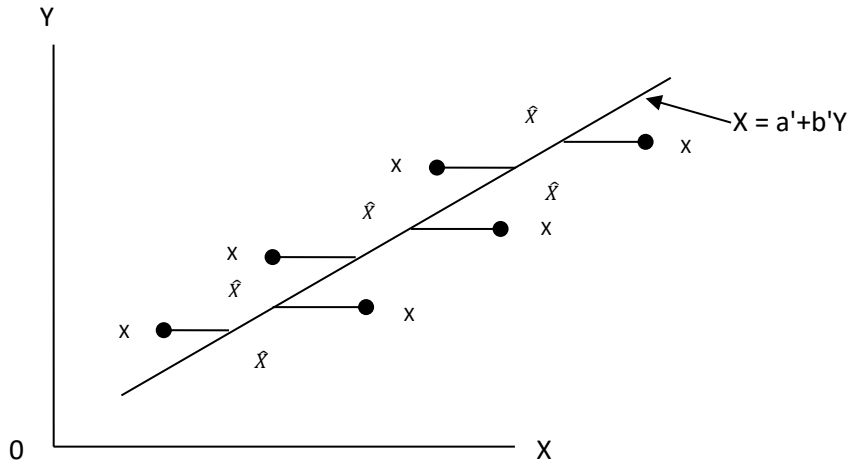
∴ The estimated values of Y is 1.96 when X = 12

Regression Equation of X on Y:

The regression equation of X on Y which describes the variation in the values of dependent variable X for given changes in independent variable Y. So, the line of regression of X on Y is the line which gives the best estimate for the value of X for any specified value of Y. Thus, the regression equation of X on Y is used to estimate the value of X when the value of Y is given.

Since, the equation of the line of regression of X on Y is obtained by minimizing the sum of squares of errors parallel to X- axis

Then, the required equation of the line of regression of X on Y becomes



Let regression equation of X on Y be

$X = a' + b' Y$ ..... (i)

Where

X = dependent variable (To find or estimate)

Y = independent variable (Given or known)

a' = value of X when Y = 0

= Intercept of the line (X - intercept)

$b' = b_{XY}$ = regression coefficient of X on Y.

= slope of line

= rate of change in X due to unit change in Y or it measures the average change in the value of X as per unit change in the value of Y.

Applying principle of least square method, the normal equations for estimating a' & b' are

$\Sigma X = n a' + b'\Sigma Y$ ..... (ii)

$\Sigma XY = a'\Sigma Y + b'\Sigma Y$..... (iii)

Where n = no. of pairs of observations.

By solving equations (ii) and (iii), we get estimates of a' and b'. Substituting theses values in equation (i), we get the required equation of regression line or regression of X on Y i.e.

$\hat{X} = \widehat{a'} + \widehat{b'} Y$ ..... (iv)

This is called regression equation of X on Y. It is also called line best fit of X on Y or prediction equation or estimating equation or linear equation or least square regression equation of X on Y or estimated regression equation of X on Y.

Alternatively

The regression equation of X on Y is

$$X - \bar{X} = b_{XY}(Y - \bar{Y}) \quad \ldots\ldots\ldots (i)$$

Where

X = dependent variable

Y = independent variable

$\bar{Y}$ = Average (mean) value of variable Y

$\bar{X}$ = Average (mean) value of variable X

b_{XY} = regression coefficient of X on Y.

= slope of line

= rate of change in X due to unit change in Y

Calculation of regression coefficient of X on Y (i.e. b_{XY}):

(i) Actual mean method

$$b_{XY} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(Y-\bar{Y})^2}$$

This method is appropriate when $\bar{X}$ & $\bar{Y}$ are whole numbers.

(i) Direct method

$$b_{XY} = \frac{n\sum XY-(\sum X)(\sum Y)}{n\sum Y^2-(\sum Y)^2}$$

(ii) Short cut method or Assumed mean method (Change of origin)

$$b_{XY} = \frac{n\sum UV-(\sum U)(\sum V)}{n\sum V^2-(\sum V)^2} \quad \text{where } U = X - A \text{ \& } V = Y - B$$

(iii) Step deviation method (i.e. Change of origin and scale)

$$b_{XY} = \frac{n\sum U'V'-(\sum U')(\sum V')}{n\sum V'^2-(\sum V')^2} \times \frac{h}{k} \quad \text{where } U' = \frac{X-A}{h} \text{ \& } V' = \frac{Y-B}{k}$$

(iv) If standard deviation (S.D.) and correlation coefficient (r) are given, then the regression coefficients are given by

(v) $b_{XY} = r\frac{\sigma_x}{\sigma_y}$

where,

σ_x = S.D. of variable X.

σ_y = S.D. of variable Y.

r = correlation coefficient between two variable X and Y.

Example 21

Find the regression equation of X on Y from the following data:

X	11	14	13	18	12
Y	40	43	38	45	37

Also estimate the value of X when Y = 40.

Solution: Here, X = dependent variable (To find or estimate)

Y = independent variable

n = 5

Let the regression equation of Y on X be

$X = a + b\,Y$ ..... (i)

Calculation of sum of values

X	Y	XY	Y^2
11	40	440	1600
14	43	602	1849
13	38	494	1444
18	45	810	2025
12	37	444	1369
$\sum X = 68$	$\sum Y = 203$	$\sum XY = 2790$	$\sum Y^2 = 8287$

$\overline{X} = \frac{\Sigma X}{n} = \frac{68}{5} = 13.6$

$\overline{Y} = \frac{\Sigma Y}{n} = \frac{203}{5} = 40.6$

Regression coefficient of Y on X:

$$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$$

$$= \frac{5\times 2790 - 68\times 203}{5\times 8287 - (203)^2} = \frac{146}{226} = 0.646$$

b = 0.646

& $a = \bar{X} - b\,\bar{Y} = 13.6 - 0.646 \times 40.6$

= -12.627

Substituting the value of a & b in equation (i), we get

X = a + b Y

$\hat{X}$ = - 12.627 + 0.646Y is the required regression equation of X on Y.

When Y = 40, then

$\hat{X}$ = - 12.627 + 0.646Y

= -12.627 + 0.646× 40

= 13.213

∴ The estimated values of X is 13.213 when Y = 40

OR

Alternative method

Solution:

Here, X = dependent variable (To find or estimate)

Y = independent variable

Then, the regression line of X on Y is

$X - \bar{X} = b_{XY}(Y - \bar{Y})$..... (i)

Calculation of sum of values

X	Y	XY	Y^2
11	40	440	1600
14	43	602	1849
13	38	494	1444
18	45	810	2025
12	37	444	1369
$\sum X = 68$	$\sum Y = 203$	$\sum XY = 2790$	$\sum Y^2 = 8287$

$\overline{X} = \frac{\Sigma X}{n} = \frac{68}{5} = 13.6$

$\overline{Y} = \frac{\Sigma Y}{n} = \frac{203}{5} = 40.6$

Regression coefficient of X on Y:

$b_{XY} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$

$= \frac{5\times 2790 - 68\times 203}{5\times 8287 - (203)^2} = \frac{146}{226} = 0.646$

From equation (i)

The regression equation of X on Y is

$X - \overline{X} = b_{XY}(Y - \overline{Y})$

$\Rightarrow X - 13.6 = 0.646(Y - 40.6)$

$\Rightarrow X = 13.6 + 0.646Y - 26.227$

$\Rightarrow \hat{X} = 12.627 + 0.646Y$ is the required regression equation of X on Y

When Y = 40, then

$\hat{X} = -12.627 + 0.646Y$

$= -12.627 + 0.646\times 40$

$= 13.213$

∴ The estimated values of X is 13.213 when Y = 40

Example 22

The following table gives the normal weight of a baby during the first six months of life:

Weight in lbs. (X)	5	7	8	10	12
Age in months (Y)	0	2	3	5	6

Estimate the weight of a baby at the age of 4 months.

Solution: Here, X = Weight (in lbs.) (To estimate or find)

Y = Age (in months)

n = 5

Let the regression equation of X on Y be

$X = a + bY$ ..... (i)

Calculation of sum of values

X	Y	XY	Y^2
5	0	0	0
7	2	14	4
8	3	24	9
10	5	50	25
12	6	72	36
$\sum X = 42$	$\sum Y = 16$	$\sum XY = 160$	$\sum Y^2 = 74$

$\overline{X} = \frac{\Sigma X}{n} = \frac{42}{5} = 8.4$

$\overline{Y} = \frac{\Sigma Y}{n} = \frac{16}{5} = 3.2$

Regression coefficient of Y on X is

$b = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum Y^2 - (\sum Y)^2}$

$= \frac{5\times 160 - 42\times 16}{5\times 74 - (16)^2} = \frac{128}{114} = 1.123$

$b = 1.123$

$$\& \ a = \overline{X} - b\,\overline{Y} = 8.4 - 1.123 \times 3.2$$
$$= 4.8064$$

Substituting the values of a & b in equation (i), we get

$X = a + b\,Y$

$\hat{X} = 4.8064 + 1.123\,Y$ is the required regression equation of X on Y.

When Y = 4 months, then

$\hat{X} = 4.8064 + 1.123 \times 4$

$= 9.2984$ lbs

∴ The estimated weight of baby at the age of 4 months is 9.2984 lbs.

OR

Alternative method

Solution:

Here, X = Weight (in lbs.) (To estimate or find)

Y = Age (in months)

n = 5

Let the regression equation of X on Y be

$X - \overline{X} = b_{XY}\,(Y - \overline{Y})$..... (i)

Calculation of sum of values

X	Y	XY	Y^2
5	0	0	0
7	2	14	4
8	3	24	9
10	5	50	25
12	6	72	36
$\sum X = 42$	$\sum Y = 16$	$\sum XY = 160$	$\sum Y^2 = 74$

$$\overline{X} = \frac{\Sigma X}{n} = \frac{42}{5} = 8.4$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{16}{5} = 3.2$$

Regression coefficient of Y on X is

$$\mathbf{b_{XY}} = \frac{\mathbf{n\sum XY - (\sum X)(\sum Y)}}{\mathbf{n\sum Y^2 - (\sum Y)^2}}$$

$$= \frac{\mathbf{5 \times 160 - 42 \times 16}}{\mathbf{5 \times 74 - (16)^2}} = \frac{\mathbf{128}}{\mathbf{114}} = 1.123$$

From equation (i)

The regression equation of X on Y is

$X - \overline{X} = b_{XY}\,(Y - \overline{Y})$

or, $X - 8.4 = 1.123\,(Y - 3.2)$

or, $X = 8.4 + 1.123Y - 3.593$

$\Rightarrow \hat{X} = 4.8064 + 1.123Y$ is the required regression equation of X on Y

When Y = 4 months, then

$\hat{X} = 4.8064 + 1.123 \times 4$

$= 9.2984$ lbs

∴ The estimated weight of baby at the age of 4 months is 9.2984 lbs.

OR

The following table gives the normal weight of a baby during the first six months of life:

Weight in lbs. (X)	5	7	8	10	12
Age in months (Y)	0	2	3	5	6

Estimate the weight of a baby at the age of 4 months.

Solution: Here, X = Weight (in lbs.)

Y = Age (in months)

n = 5

Let the regression equation of X on Y be

$X = a + b\,Y$ ..... (i)

By using the principle of least square method, the normal equations for estimating a and b are

$\sum X = n\,a + b\sum Y$ ..... (ii)

$\sum XY = a\sum Y + b\sum Y^2$ ..... (iii)

Calculation of sum of values

X	Y	XY	Y^2
5	0	0	0
7	2	14	4
8	3	24	9
10	5	50	25
12	6	72	36
$\sum X = 42$	$\sum Y = 16$	$\sum XY = 160$	$\sum Y^2 = 74$

Substituting the sum of values in normal equations (ii) and (iii), we get

$42 = 5a + b\,16$

or, $5a + 16\,b = 42$ . . . . . .(iv)

& $160 = a\,16 + b\,74$

or, $16a + 74\,b = 160$ . . . . .(v)

Solving equations (iv) and (v)

$[5a + 16\,b = 42] \times 16$

$[16a + 74\,b = 160] \times 5$

$-\quad -\quad -$

$-114\,b = -128$

$b = \frac{128}{114} = 1.123$

Substituting the value of b in equation (iv), we get

$5a + 16 \times 1.123 = 42$

or, $5a = 42 - 17.968$

or, $a = \frac{24.032}{5} = 4.8064$

From equation (i)

Hence, the required regression equation of X on Y is

$\hat{X} = 4.8964 + 1.123X$

when Y = 4 months, then

$\hat{X} = 4.8964 + 1.123 \times 4$

$= 4.8964 + 4.492$

$= 9.2984$ lbs.

∴ The estimated weight of baby at the age of 4 months is 9.2984 lbs.

Difference between correlation and regression

Correlation	Regression
1. Correlation analysis is the statistical tool (statistical measure) which is used to study or describe the degree of relationship to which the variables are linearly related.	1. Regression analysis is the mathematical measure of the average relationship between two or more variables in terms of original units of data whether the variables are linearly related or non-linearly related.
2. It does not necessarily imply cause and effect relationship between the two variables under study.	2. It necessarily shows (indicates) the cause and effect relationship between the variables such that the cause is taken as independent variable and effect is taken as dependent variable.
3. Correlation coefficient i.e. r_{XY} is a relative measure of linear relationship between two variables and is independent of the units of measurement.	3. Regression coefficients bxy & byx are absolute measures. so, regression coefficients have units of measurement of the variable.
4. Correlation coefficients are symmetric i.e. $r_{XY} = r_{YX}$	4. Regression coefficients are not symmetric. i.e. bxy $\neq$ byx
5. Correlation analysis is confined only to the study of linear relationship between the variables	5. Regression analysis studies linear as well as non-linear (curvilinear) relationship between the variables.
6. Correlation coefficients is a pure number lying between -1 & +1.	6. But regression coefficients are absolute measures and if one of the regression coefficient is greater than unity (one), the other must be less than unity. i.e. i.e. if $b_{yx} . > 1$ then $b_{xy} < 1$
7. Correlation coefficient is independent of change of origin and scale.	7. Whereas regression coefficients are independent of change of origin but not of scale.
8. Correlation coefficient is independent of units of the variables	8. Regression coefficient is expressed in the units of dependent variable.

Introduction to probability

Introduction and Importance of Probability

Probability is one of the important branches of Statistics that is concerned with random phenomena. Everyday concept of "likelihood", "predictability" and "certainty" are formalized by the part of statistics called probability. Probability is a statistical method used to study of the chance of occurrence and non occurrence of different phenomena. It is the synonym of the words most likely, chance, probably etc. Hence, Probability defined as 'uncertainty is numerically expressed.
The theory of probability was developed in the middle of the seventeenth century. It was originated from the problem related to gambling such as tossing a coin, throwing a die, drawing cards from a pack of cards etc. Therefore, in the early days it was known as the theory of games and chance. But nowadays, there is hardly any displine (field) where "probability" has not been used. The theory of probability is used in almost all displine like physical science, natural science, biological science, medical science, engineering, social sciences, business, industry etc. It is extensively used in quantitative analysis of business and economics problems.The concept of probability has many real life applications such as when playing cards, you can determine the likelihood of obtaining a particular hand, when selling insurance you can use probability ot assess the risk of an individual, in a production plant you can use probability to detearmine the possibility of making a defective product etc. It is an essential tool in statistical inference and backbone of probability sampling and forms the basis of the "Decision Theory", viz. decision making in the case of uncertainty with calculated risks.

Business firms, factory managers, stock market policy makers often have been facing the problems regarding the chance of selling the goods, chance of receiving better demand, chance of getting defective or non-defective product and chance of deterioration and prosperity of stock market etc.The theory of probability can be used to measure the risk or errors in forecasting and to minimize such risks. It helps a business manager in reducing the elements of uncertainty associated with the decision making with respect to cost, profit, production, sales, pricing, capital investment etc. Insurance companies make an extensive use of probability. The word probability is a chance or possibility which is widely used in daily life also.

e.g. What is the chance of heavy rain in this morning?

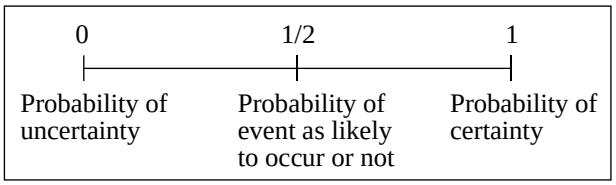
What is the chance of increasing sales when the price of product is decreased?

How likely is it that a given project will be completed in time? and so on.

Definition:

Probability is a numerical measure (with a value lying between 0 and 1) of the likelihood or chance that a particular event will occur or not.

Probability is simply a number lies between zero and one. That is; $0 \leq P \leq 1$. It is generally denoted by 'P'. If there is absolute possibility of occurrence of an event, then its probability is equal to one. This probability is also called the probability of certainty. If there is complete impossibility that an event will occur, then its probability is zero and is called the probability of uncertainty. This shows that the range of probability is in between zero and one. This range of probability is shown as follows.



Probability theory provides us with a mechanism for measuring and analysing uncertainties associated with future events.

Basic Terminology in Probability

Experiment: When the process (phenomenon) performed which results different possible outcomes is known as experiment. For example, tossing a coin, attempting in an examination, throwing a dice etc.

Random Experiment

An experiment is called a random experiment if it performed a large number of times under essentially homogeneous conditions; the result is not unique but may be any one of the various possible outcomes. Tossing an unbiased coin and rolling a uniform die are some examples of random experiment.

Trial and Event

Performing a random experiment is called a trial and outcome or combination of outcomes (i.e. results) of random experiment is called an event. For example; flipping an unbiased coin is a trial and getting either head or tail is an event.

Sample Space

The set of all possible outcomes of a random experiment is called sample space. Each outcome is thus considered as a sample point in the sample space. It is usually denoted by S. For example, in tossing an unbiased coin once, there are two possible outcomes head (H) and tail (T). So, the sample space is given by S = {H, T}

Exhaustive Cases

The total number of possible outcomes of a random experiment is called the exhaustive cases for the experiment. In a toss of single coin, we can get head (H) or tail (T). Hence, exhaustive number of cases is 2

Similarly, in rolling a six faced die, the exhaustive number of cases = 6.

Favourable Cases or Events

The number of outcomes of a random experiment which result in the happening of an event are termed as the cases favourable to the event. In other words, the number of outcomes which are in favour of happening of an event are called favourable cases. For example; in drawing a card from a pack of 52 playing cards, the number of cases favourable to drawing a queen is 4. Similary , in a toss of two coins, the number of cases favourable to the event 'exactly one head ' is 2, viz., HT , TH and for 'two heads' is one viz., HH.

Equally Likely Events

Events or cases are said to be equally likely or equally probable if none of them is expected to occur in preference to other. In other words, two or more events are said to be equally likely if all of them have equal chance of occurrence. In tossing an unbiased coin, head and tail are equally likely, throwing a unbiased die (the faces 1, 2, 3, 4, 5, 6) are equally likely.

Mutually Exclusive Events

Two or more events are said to be mutually exclusive if the happening of any one of them excludes the happening of all others in the same experiments i.e. if two or more than two events cannot occur simultaneously at the same time in the same trial. For example in a single tossing of a coin, we may get either head or tail but not both. Thus, the events head and tail in a single tossing of a coin are mutually exclusive. Similarly, in the

throw of a die, the six faces numbered 1, 2, 3, 4, 5 & 6 are mutually exclusive. Thus, no two or more of them can happen simultaneously.

Independent Events

Events are said to be independent if the occurance of one event does not affect the occurance of the other events and vice versa. In tossing of a coin, the occurrence of head in first tossing is independent of the occurrence of head in the second tossing. Similarly, drawing of balls from an urn gives independent events if the draws are made with replacement.

Dependent Events

Events are said to be dependent if the occurance of one event affects the occurance of the other events and vice versa. For example, if a pack of cards is playing without replacemement, the occurrence of a king in first darw affects the occurrence of other cards in the second draw.

Approaches of Probability

The chance of occurrence or non-occurrence of any event in a random experiment is called probability. There are four approaches to define the probability.

- Mathematical or classical or priori approach
- Statistical or empirical or relative frequency approach
- Subjective approach
- Axiomatic approach

Mathematical or Classical or Priori, approach of Probability

Let n be the exhaustive, mutually exclusive and equally likely cases (or outcomes) out of which 'm' are favourable cases to the happening of an event A. Then the probability of happening (occurrence) of an event A is given by

$$P(A) = \frac{\text{Favourable number of cases to A}}{\text{Exhaustive number of cases}} = \frac{m}{n}$$

or, $P(A) = \frac{m}{n}$ . . . . . .(1)

If (n – m) cases are favourable to the non-occurrence of an event A, then its probability is given by

$$P(A^c) \text{ or } P(\bar{A}) = q = \frac{\text{Favourable number of cases for not happening of event A}}{\text{Exhaustive number of cases}}$$

$$1 - \frac{m}{n} = \frac{n-m}{n}$$

or, $P(\bar{A}) = q = 1 - \frac{m}{n}$

or, $P(\bar{A}) = q = 1 - p(A)$

Hence, $P(A) + P(\bar{A}) = 1$ . . . . . (2)

or, $p + q = 1$, $q = 1 - p$

This shows that total probability is equal to 1.

Hence, probability of occurrence of event A + Probability of non-occurrence of event A = 1.

i.e. 100%

Note that probability lies between zero and 1 i.e. $0 \leq p \leq 1$.

Limitations of Classical Approach of Probability

This approach breaks down in the following cases

- If n, the exhaustive number of cases (outcomes) of the random experiment is infinite.
- If the various outcomes of the random experiment are not equally likely.
- If the actual value of n is not known.

Statistical or Empirical or Relative Frequency Approach

This approach of probability is based on statistical data. If an experiment is performed repeatedly a large number of times under essentially homogeneous and identical conditions, then the limiting value of the ratio of the number of times the event occurs to the total number of trials of the experiment as the number trials becomes indefinitely large, is called the probability of happening of the event. It being assumed that the limit is finite and unique.

Suppose an event A occurs m times in n repeatitions of a random experiment. Then the ratio $\frac{m}{n}$ gives the relative frequency of the event A. In the limiting case when n becomes sufficiently large, then a number which is called the probability of A.

Symbolically,

$$P(A) = \lim_{n \to \infty} \frac{m}{n}$$

Limitations:

- The experimental conditons may not remain essentially homogeneous and identical in a large number of repeatitions of the experiment.
- The relative frequency $\frac{m}{n}$, may not attain a unique value, no matter however large n may be.
- The value of $\frac{m}{n}$ is not unique, it changes as $n \to \infty$, So, exact probability is impossible.

Subjective Approach

The subjective probability approach is purely individualistic in nature. Therfore, this approach of probability is completely based on the personal beliefs, feelings, experience, judgement, personal discretion of a person. Since, different persons may assign different probabilities one cannot arrive at objective conclusions using probabilities assigned by this subjective method. This method of assigning probability is generally used by top level authorities on the basis of their discretion.

Axiomatic Probability

The modern theory of probability is based on the axiomatic approach introduced by the Russian mathematician A.N. Kolmogorov in 1930. The aximatic definition of probability includes both classical and empirical definition of probability and at the same time is the free from their drawbacks. It is also called elementary properties of probability. It states the following facts:

Definiton: Given a sample space of a random experiment, the probability of occurrence of any event 'E' is defined as a set function $P(E)$ satisfying the following axioms:

Axiom 1: Axiom of non-negativity

$P(E)$ is defined, is real and non-negative) i.e. $P(E) \geq 0$

Axiom 2: Axiom of certainty

The sum of the probabilities is unity i.e. $P(S) = 1$, where S = Sample space

Axiom 3: Axiom of additivity

If $E_1, E_2, E_3., \ldots\ldots, E_n$ is any finite or infinite sequence of disjoint events of sample space S, then

$P(U_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i)$ or $P(U_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Note: When probability is not given directly, then it can be categorised into two cases:

Case (i): When one item is selected at a time.

Case (ii): When more than one item is selected at a time.

Examples related to

Case (i): When one item is selected at a time.

Example 4

A card is drawn at random from a well-shuffled pack of 52 cards. What is the probability of getting?

(a) a red card,

(b) a black king

Solution:

There are 52 cards in a pack.

Total number of cases (n) = 52

(a) P (a red card) =?

Since, there are 26 red cards.

Favourable number of cases (m) = 26

Required probability of getting a red card is

P (a red card) = $\frac{m}{n}$

$= \frac{26}{52} = \frac{1}{2}$

(b) P (a black king) =?

There are 2 black kings

Favourable number of cases (m) = 2

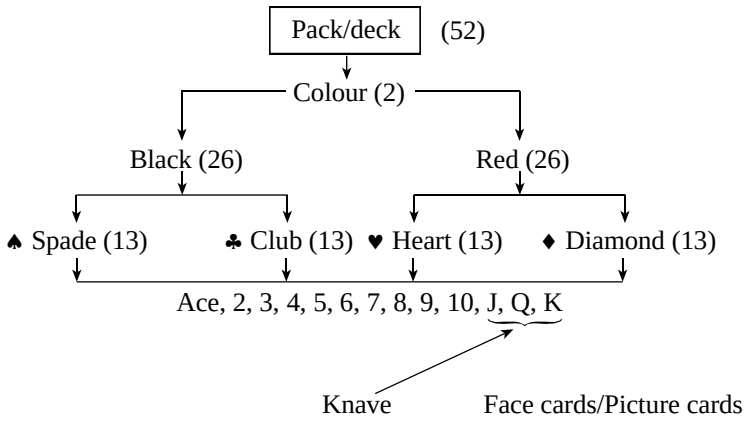
Required probability of getting a black king is

P (a black king) = $\frac{2}{52} = \frac{1}{26}$

Example 5

A card is drawn from a pack of 52 cards at random what is the probability that it is (i) Red, (ii) Spade, (iii) Face card, (iv) an ace, (v) red king, (vi) Knave of heart, (vii) King or queen, (viii) Heart or club, (ix) a red 2 or black 8 or a queen, (x) Spade or ace, (xi) Heart or face card, (xii) Red or face card.

Solution:



Since, there are 52 cards in a pack.

n = Total number of cases (or outcomes) = 52

Since, one card is drawn at random

(i) P (a red card) =?

Let 'E' denotes the event of drawing red cards

Favourable number of case (m) = 26

$\therefore$ P (a red) = P (E) = $\frac{m}{n} = \frac{26}{52} = \frac{1}{2}$

(ii) P (a spade) =?

Let E denotes the event of drawing spade cards

Favourable number of cases (m) = 13

$\therefore$ P (a spade) = P (E) = $\frac{m}{n} = \frac{13}{54} = \frac{1}{4}$

Similarly,

(iii) P (a face card) =?

Let 'E' denotes the event of drawing a face card

Favourable number of cases (m) = 12

(3 face cards in each suit)

P (a face card) = P (E) = $\frac{m}{n} = \frac{12}{52} = \frac{3}{13}$

(iv) P (an ace) =?

Let 'E' denotes the event of drawing an ace

Favourable number of outcomes (m) = 4

P (an ace) =$\frac{m}{n} = \frac{4}{52} = \frac{1}{13}$

(v) P (a red king) =?

Favourable number of cases (m) = number of red kings = 2

P (a red king) = $\frac{2}{52} = \frac{1}{26}$

(vi) P (a knave of heart) =?

Favourable number of cases (m) = number of knave of heart = 1

P (a knave of heart) = $\frac{1}{52}$

(vii) P (a king or queen) =?

Favourable number of cases (m) = number of kings or queen

= 4 + 4 = 8

P (king or queen) = $\frac{8}{52} = \frac{2}{13}$

(viii) P (Heart or club) =?

m = 13 + 13 = 26

P (Heart or club) = $\frac{26}{52} = \frac{1}{2}$

(ix) P (a red 2 or black 8 or a queen) =?

m = number of red 2 or black 8 or a queen

= 2 + 2 + 4 = 8

P (a red 2 or black 8 or a queen) = $\frac{8}{52} = \frac{2}{13}$

(x) P (a spade or ace) =?

m = number of spade or ace

= 13 + 4 – 1 (∵ spade ace is common in both)

= 16

P (a spade or ace) = $\frac{16}{52} = \frac{4}{13}$

(xi) P (a heart or face card) =?

Favourable number of cases (m) = number of heart or face cards

= 13 + 12 – 3

= 22 (∵ 3 face cards of heart are common)

P (a heart or face card) = $\frac{22}{52} = \frac{11}{26}$

(xii) P (a red or face card)

Favourable number of cases (m) = number of red or face cards

= 26 + 12 – 6 (∵ 6 red cards are common)

P (a red or face card) = $\frac{32}{52} = \frac{8}{13}$

Example 6

Twenty balls are numbered from 1 to 20. If one ball is drawn at random, what is the probability that the ball drawn is multiple of 4 or 7?

Solution:

Total number of cases (n) = 20
Favourable number of cases (m) = The number of cases which are multiple of 4 or 7

i.e. {4, 7, 8, 12, 14, 16, and 20}

i.e. m = 5+2 = 7

Required probability that the ball drawn is multiple of 4 or 7 = $\frac{m}{n} = \frac{7}{20}$

Example 7

Two fair dice are thrown at random. What is the probability that the face turn up show (i) a sum 7 (ii) a sum of 8 or 9 (iii) sum less than 5 (iv) number 6 in the first die (v) odd number in the second die (vi) same faces (vii) different faces.

Solution:

Since, two dice are thrown

n = Total number of cases = 6 × 6 = 36

The sample spaces (total outcomes) are presented below:

		Faces in second die					
		1	2	3	4	5	6
Faces in first die	1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
	2	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
	3	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
	4	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
	5	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
	6	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

(i) P (a sum 7) =?

m = Favourable number of outcomes

= number of cases of getting a sum 7 in both dice

i.e. (1, 6) (2, 5) (3, 4) (4, 3) (5, 2) (6, 1)

m = 6

$$P(\text{a sum 7}) = \frac{m}{n} = \frac{6}{36} = \frac{1}{6}$$

(ii) P (a sum of 8 or 9) =?

m = Favourable number of outcomes

= number of cases getting a sum of 8 or 9

i.e. (2, 6) (3, 5) (4, 4) (5, 3) (6, 2) (3, 6) (4, 5) (5, 4) (6, 3)

m = 9

$$P(\text{a sum is 8 or 9}) = \frac{9}{36} = \frac{1}{4}$$

(iii) P (sum less than 5) =?

m = Favourable number of outcomes

= number of cases getting a sum of less than 5

i.e. (1, 1) (1, 2), (2, 1), (1, 3), (2, 2) (3, 1)

m = 6

$$P(\text{sum less than 5}) = \frac{6}{36} = \frac{1}{6}$$

(iv) P (number 6 in the first die) = ?

m = Favourable number of outcomes

= number of cases getting a 6 in 1st die

i.e. (6, 1), (6, 2), (6, 3), (6, 4), (6, 5), (6, 6)

m = 6

$\therefore$ P (6 in first die) = $\frac{6}{36} = \frac{1}{6}$

(v) P (odd number in the second die)

m = Favourable number of outcomes

= number of cases of odd number in the second die

i.e. (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (1, 3), (2, 3)

(3, 3), (4, 3), (5, 3), (5, 3), (6, 3), (1, 5), (2, 5), (3, 5),

(4, 5) (5, 5), (6, 5)

m = 18

P (Odd number in the second die) = $\frac{18}{36} = \frac{1}{2}$

(vi) P (same faces)

m = Favourable number of outcomes (cases)

= number of cases of getting same faces in both dice

i.e. (1, 1) (2, 2) (3, 3) (4, 4) (5, 5) (6, 6)

m = 6

P (same faces in both dice) = $\frac{6}{36} = \frac{1}{6}$

(vii) P (different faces) =?

m = Favourable number of outcomes

= Number of cases of different faces in both dice

i.e. (1, 1) (2, 1) (1, 3) (3, 1) .... , (1, 6) (6, 1)

m = 30

P (different faces) = $\frac{m}{n} = \frac{30}{36} = \frac{5}{6}$

Since, P (different faces) + P (same faces) = 1

$\therefore$ P (different faces) = 1 – P (same faces)

$$= 1 - \frac{6}{36} = \frac{30}{36} = \frac{5}{6}$$

Example 8

What is the chance that a leap year selected randomly consists of 53 Sundays?

Solution:

In a leap year, there are 366 days i.e. 52 complete weeks and 2 days more. These 2 days may be either (i) Sunday and Monday or (ii) Monday and Tuesday or (iii) Tuesday and Wednesday or (iv) Wednesday and Thursday or (v) Thursday and Friday or (vi) Friday and Saturday or (vii) Saturday and Sunday.

Total number of cases (n) = 7

P (53 Sundays) =?

Favourable number of cases (m) = number of cases consisting Sunday

$$= 2$$

∴ Required probability that a leap year selected randomly consists of 53 Sundays i.e.

$$P(A) = \frac{m}{n} = \frac{2}{7}$$

Example 9

What is the chance that a non-leap year selected randomly consists of 53 Sundays?

Solution:

In a leap year, there are 365 days i.e. 52 complete weeks and 1 day more. The 1 day may be either (i) Sunday or (ii) Monday or (iii) Tuesday or (iv) Wednesday or (v) Thursday or (vi) Friday or (vii) Saturday.

Total number of cases (n) = 7

P (53 Sundays) =?

Favourable number of cases (m) = number of cases consisting of Sunday

$$= 1$$

∴ Required probability that a non-leap year selected randomly consists of 53 Sundays i.e.

$$P(A) = \frac{m}{n} = \frac{1}{7}$$

Example 11

In a group of equal number of men and women, 20% of men and 30% of women are unemployed. If a person selected at random, what is the probability the selected person is an employed?

Solution:

Since, the number of men and women are equal and their unemployment percentages are given. Let the population of men and women each be 100 Therefore, the given information can be presented in the following tabular form:

Nature / Sex	Unemployed	Employed	Total
Men	20	80	100
Women	30	70	100
Total	50	150	200

Let A be the event of selecting employed person.

Total number of cases (n) = total number of persons = 200

P (person is an employed) = P (A) =?

Favourable number of cases (m) = number of employed persons = 150

Then,

$P(A) = \frac{m}{n} = \frac{150}{200} = \frac{3}{4} = 0.75$

Examples related to

Case (ii): When more than one item is selected at a time.

Example 12

A bag contains 8 red, 4 white and 5 black coloured balls. Three balls are drawn randomly from a bag. Find the probability that (i) all are red (ii) 2 is red and 1 white (iii) 2 are red and 1 other (iv) all colour balls.

Solution:

Exhaustive (total) number of cases (n) = number of cases of selection of 3 balls out of 17 balls

$$= {}^{17}C_3 = \frac{17 \times 16 \times 15}{1 \times 2 \times 3} = 680.$$

(i) Favourable number of cases of drawing all 3 are red balls (m) = ${}^{8}C_3$

$$= \frac{1 \times 7 \times 6}{1 \times 2 \times 3} = 56$$

P (all are red) = $\frac{m}{n} = \frac{56}{680} = 0.082$

(ii) Favourable cases for 2 red and 1 white (m) = ${}^{8}C_2 \times {}^{4}C_1$

$$= \frac{8 \times 7}{1 \times 2} \times \frac{4}{1} = 112$$

P (2 is red and 1 white) $= \frac{m}{n} = \frac{112}{680} = 0.1647$

(iii) Favourable cases for 2 are red out of three drawn balls i.e. 2 are red and 1 other

(m) = ${}^{8}C_2 \times {}^{9}C_1 = \frac{8 \times 7}{1 \times 2} \times \frac{9}{1} = 252$

P (2 are red and 1 other) = $\frac{m}{n} = \frac{252}{680} = 0.37058$

(iv) Favourable number of cases for all coloured balls (m) = ${}^{8}C_1 \times {}^{4}C_1 \times {}^{5}C_1 = 8 \times 4 \times 5 = 160$

Hence, P (all colour balls) = $\frac{m}{n} = \frac{160}{680} = \frac{4}{17} = 0.235$

Example 13

Five men in a group of 20 are graduates. If 3 are chosen out of 20 at random, what is the probability that

a) all are graduates b) none of them is graduate

c) al least one of them being graduate

Solution: Here,

Total number of men = 20

Number of graduates = 5

Number of non-graduates = 20 – 5 = 15

If three are choosen at random

Total number of possible outcomes (n) = ${}^{20}C_3$

$$= \frac{20!}{(20-3)!\ 3!}$$

$$= \frac{20!}{17!\ 3!}$$

$$= \frac{20 \times 19 \times 18 \times 17!}{17!\ 3!}$$

$$= 1140$$

(a) P (all are graduates) =?

Favourable number of cases (m) = ${}^5C_3 = \frac{5!}{(5-3)!\,3!} = 10$

$\therefore$ P (all are graduate) = $\frac{{}^5C_3}{{}^{20}C_3} = \frac{10}{1140} = \frac{1}{114}$

(b) P (none of them are graduate) =?

Favourable number of cases for none of them are graduate (m) = ${}^{15}C_3$

$= \frac{15!}{(15-3)!\,3!} = 455$

$\therefore$ P (none of them are graduate) = $\frac{{}^{15}C_3}{{}^{20}C_3} = \frac{455}{1140} = \frac{91}{228}$

(ii) P (at least one of them being graduate)

= 1 – P (none them are graduate)

= $1 - \frac{91}{228}$

= $\frac{228-91}{228} = \frac{137}{228}$

Example 14

A class consists of 40 boys and 60 girls. If two students are chosen at random, what will be the probability that

(a) both are boys (b) both are girls

(c) one boy and one girl

Solution:

Number of boys in a class = 40

Number of girls in a class = 60

Total number of student = 40 + 60 = 100

If two students are choosen at random, then total number of possible outcomes (n)

= ${}^{100}C_2 = 4950$

(i) P (both are boys) =?

Favourable case for both are boys (m) = ${}^{40}C_2 = 780$

$\therefore$ P (both are boys) = $\frac{{}^{40}C_2}{{}^{100}C_2} = \frac{780}{4950} = \frac{78}{495}$

(ii) P (both are girls) =?

Favourable case for both are girls (m) = ${}^{60}C_2 = 1770$

$\therefore$ P (both are girls) $= \frac{^{60}C_2}{^{10}C_2} = \frac{1770}{4950} = \frac{177}{495} = \frac{59}{165}$

(iii) P (one boy and one girl) =?

Favourable number of cases for one boy and one girl (m)

$= {}^{40}C_1 \times {}^{60}C_1$

$= 40 \times 60 = 2400$

$\therefore$ P (one boy and one girl) $= \frac{^{40}C_1 \times {}^{60}C_1}{^{100}C_2} = \frac{2400}{4950} = \frac{48}{99} = \frac{16}{33}$

Laws (or rules) of Probability)

Following are the laws of probability:

1. Additive law of probability
2. Multiplicative law of probability

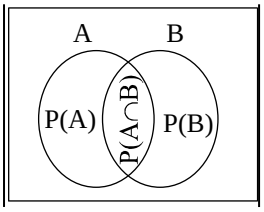
Additive Law of Probability (Addtion Theorem of Probability)

(At least one, one of them, or) "Union $\cup$" , +

Case I: when events are not mutually exclusive

If A and B are not mutually exclusive events, then the probability of occurrence of at least one of them is given by

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



The probability of occurrence of at least one of them can also be written as

$$P(A \cup B) = 1 - P(\overline{A \cup B})$$

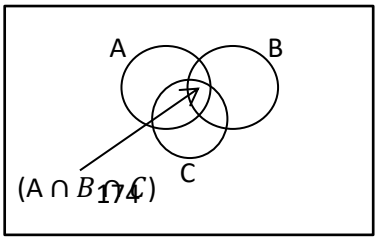
$$= 1 - P(\overline{A} \cap \overline{B}) \qquad \because P(A \cup B) + P(\overline{A \cup B}) = 1$$

Demorgan's law

$$P(\overline{A \cup B}) = P(\overline{A} \cap \overline{B}) \quad \& \quad P(\overline{A \cap B}) = P(\overline{A} \cup \overline{B})$$

If A, B and C are three not mutually exclusive events then the probability of occurrence of at least any one of them is given by

P (A or B or C) = P (A∪B∪C) = P (A) + P (B) + P(C) – P (A∩B) – P (B∩C) – P (C∩A) + P (A∩B∩C)



The probability of occurrence of at least one of them can also be written as

$P(A \cup B \cup C) = 1 - P(\overline{A \cup B \cup C}) \quad \because P(A \cup B \cup C) + P(\overline{A \cup B \cup C}) = 1$

$= 1 - P(\overline{A} \cap \overline{B} \cap \overline{C})$

In general,

$P(A_1 \cup A_2 \cup A_3 \cup A_4 \cup \ldots\ldots\ldots\ldots \cup A_n) = 1 - P(\overline{A_1 \cup A_2 \cup A_3 \cup \ldots\ldots\ldots \cup A_n})$

$= 1 - P(\overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3 \ldots\ldots\ldots \cap \overline{A}_n)$

Demorgan's law

$P(\overline{A \cup B \cup C}) = P(\overline{A} \cap \overline{B} \cap \overline{C}) \quad \& \quad P(\overline{A \cap B \cap C}) = P(\overline{A} \cup \overline{B} \cup \overline{C})$

Example 15

The probability that a new airport will get an award for its design is 0.16, the probability that it will get an award for the efficient use of materials is 0.24, and the probability that it will get both awards is 0.11.

a. What is the probability that it will get at least one of the two awards?

b. What is the probability that it will get only one of two awards?

Solution:

Probability that a new airport will get award for its design, P (A) = 0.16

Probability that a new airport will get award for efficient, P (B) = 0.24

Probability that a new airport will get award for both, P (A∩B) = 0.11

a. Probability that it will get at least one of the two awards,

P (A∪B) = P (A) + P (B) – P(A∩B)

= 0.16 + 0.24 – 0.11 = 0.29

b. Probability that it will get only one of two awards.

= [P (A) – P (A∩B)] or [P (B) – P (A∩B)]

= [P (A) – P (A∩B)] + [P (B) – P (A∩B)]

= (0.16 – 0.11) + (0.24 – 0.11)

= 0.05 + 0.13 = 0.18

Example 16

A construction company is bidding for two contracts A and B. The probability that the company will get contract A is $\frac{3}{5}$, the probability that the company will get contract B is $\frac{1}{3}$ and the probability that the company will get both the contracts is $\frac{1}{8}$. What is the probability that the company will get contract A or B?

Solution:

We have, $P(A) = \frac{3}{5}, P(B) = \frac{1}{3}, P(A \cap B) = \frac{1}{8}, P(A \text{ or } B) = ?$

The probability that the company will get contract A or B is given by

$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{3}{5} + \frac{1}{3} - \frac{1}{8} = \frac{97}{120}$$

Example 17

The probability that a boy will get a scholarship is 0.9 and that a girl will get is 0.8. What is the probability that at least one of them will get the scholarship?

Solution:

Let B = event of a boy getting scholarship

G = event of a girl getting scholarship

Then,

P (B) = 0.9, P (G) = 0.8

The probability that at least one of them will get scholarship is given by

$P(B \cup G) = 1 - P(\overline{B \cup G})$

$= 1 - P(\overline{B} \cap \overline{G})$ (Uing De-morgan's law)

$= 1 - P(\overline{B}) \times P(\overline{G})$ (Since, a boy and a girl getting scholarship are independent)

$= 1 - 0.1 \times 0.2 = 1 - 0.02 = 0.98$

Alternative method (i)

The probability that at least one of them will get scholarship is given by

P (B∪G) = P (Boy get the scholarship and girl not or girl get the scholarship and boy not or both get the scholarships)

$= P(B \cap \overline{G} \text{ or } G \cap \overline{B} \text{ or } B \cap G)$

$= P(B \cap \overline{G}) + P(G \cap \overline{B}) + P(B \cap G)$

$= P(B) \times P(\overline{G}) + P(G) \times P(\overline{B}) + P(B) \times P(G)$

$= 0.9 \times 0.2 + 0.8 \times 0.1 + 0.9 \times 0.8$

$= 0.18 + 0.08 + 0.72 = 0.98$

Alternative method (ii)

The probability that at least one of them will get scholarship is given by

$P(B \cup G) = P(B) + P(G) - P(B \cap G)$

$= P(B) + P(G) - P(B) \times P(G)$

$= 0.9 + 0.8 - 0.9 \times 0.8 = 1.7 - 0.72 = 0.98$

Example 18

A bag contains 24 balls numbered from 1 to 24. One ball is drawn at random. Find the probability that the ball drawn has a number which is multiple of 3 or 4.

Solution: Let A and B be the events of drawing a number which is multiple of 3 and 4 respectively.

Then, A = {3, 6, 9,12,15,18,21,24}, B = {4,8,12,16,20,24}

$P(A) = \frac{m}{n} = \frac{8}{24}$, $P(B) = \frac{m}{n} = \frac{6}{24}$

But 12 and 24 repeated. So, $P(A \cap B) = \frac{m}{n} = \frac{2}{24}$

$\therefore P(3 \text{ or } 4) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$

$$= \frac{8}{24} + \frac{6}{24} - \frac{2}{24} = \frac{12}{24} = \frac{1}{2}$$

OR

Total number of cases (n) =24

Favourable number of cases (m) = number of cases of multiple of 3 or 4

$$= 8 + 6 - 2 = 12$$

$\therefore P(3 \text{ or } 4) = P(A \cup B) = \frac{m}{n} = \frac{12}{24} = \frac{1}{2}$

Example 19

A product is made up of three components C_1, C_2 and C_3. The probability that thses components are defective are in the ratio 1:2:3. Find the probability that one product selected at random is non defective.

Solution

Since, the probability that these components C_1, C_2 and C_3 are defective are in the ratio 1:2:3.

Total ratio = 1+2+3 = 6

Then, $P(C_1) = \frac{1}{6}$, $P(\bar{C}_1) = 1 - \frac{1}{6} = \frac{5}{6}$

$P(C_2) = \frac{2}{6}$, $P(\bar{C}_2) = 1 - \frac{2}{6} = \frac{4}{6}$

$P(C_1) = \frac{1}{6}$, $P(\bar{C}_3) = 1 - \frac{3}{6} = \frac{3}{6}$

Hence, the probability that one product selected at random is non defective is given by

$P(\bar{C}_1, \bar{C}_2 \text{ and } \bar{C}_3) = P(\bar{C}_1 \cap \bar{C}_2 \cap \bar{C}_3)$

$= P(\bar{C}_1) \times P(\bar{C}_2) \times P(\bar{C}_3)$ $\quad \because$ the events $\bar{C}_1, \bar{C}_2$ and $\bar{C}_3$ are independent

$= \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} = \frac{5}{18} = 0.277$

Example 20

If the probability of machines M_1, M_2 and M_3 working without failure are 0.2, 0.3 and 0.5 respectively. Find the probability that

(i) at least one machine will work without failure.
(ii) at least two machine will work without failure.
(iii) all machines will work without failure.

Solution:

Let A, B and C denote the events that machines M_1, M_2 and M_3 working without failure respectively. Then

$P(A) = 0.2, \quad P(\bar{A}) = 1 - 0.2 = 0.8$

$P(B) = 0.3, \quad P(\bar{B}) = 1 - 0.3 = 0.7$

$P(C) = 0.5, \quad P(\bar{C}) = 1 - 0.5 = 0.5$

(i) P (at least one machine will work without failure) $= P(A \cup B \cup C)$

= 1– P(none of them will pass the exam)

$= 1 - P(\bar{A})\, P(\bar{B})\, P(\bar{C})$

$= 1 - 0.8 \times 0.7 \times 0.5 = 1 - 0.28 = 0.72$

(ii) P (at least two machine will work without failure) = $P(A \cap B \cap C \text{ or } A \cap B \cap \bar{C} \text{ or } A \cap \bar{B} \cap C \text{ or } \bar{A} \cap B \cap C)$

$= P(A)P(B)P(C) + P(A)P(B)P(\bar{C}) + P(A)\, P(\bar{B})\, P(C) + P(\bar{A})\, P(B)P(C)$

$= 0.2 \times 0.3 \times 0.5 + 0.2 \times 0.3 \times 0.5 + 0.2 \times 0.7 \times 0.5 + 0.8 \times 0.3 \times 0.5 = 0.25$

($\because$ Events are mutually exclusive and independent)

(iii) p (all machines will work without failure) = $P(A \cap B \cap C)$

$= P(A)P(B)P(C)$

$= 0.2 \times 0.3 \times 0.5 = 0.03$

Example 21

A, B and C will pass a certain examination in the proportion 2:4:6. What is the probability that

(i) at least two of them will pass the examination?
(ii) at least one of the them will pass the examination?
(iii) all will pass the examination?
(iv) none will pass the examination?
(v) B will pass but not A and C?

Solution:

Since, the pass proportion of A, B and C in a certain examination are in the rato 2:4:6.

Total ratio = 2 + 4 + 6 = 12. Then

$P(A) = \frac{2}{12} = \frac{1}{6}, \quad P(\bar{A}) = 1 - \frac{1}{6} = \frac{5}{6}$

$P(B) = \frac{4}{12} = \frac{1}{3}, \quad P(\bar{B}) = 1 - \frac{1}{3} = \frac{2}{3}$

$P(C) = \frac{6}{12} = \frac{1}{2}, \quad P(\bar{C}) = 1 - \frac{1}{2} = \frac{1}{2}$

(iv) P (at least two of them will pass the examination) = P $(A \cap B \cap C$ or $A \cap B \cap \bar{C}$ or $A \cap \bar{B} \cap C$ or $\bar{A} \cap B \cap C)$

$= P(A)P(B)P(C) + P(A)P(B)P(\bar{C}) + P(A)\,P(\bar{B})\,P(C) + P(\bar{A})\,P(B)P(C)$

$= \frac{1}{6} \times \frac{1}{3} \times \frac{1}{2} + \frac{1}{6} \times \frac{1}{3} \times \frac{1}{2} + \frac{1}{6} \times \frac{2}{3} \times \frac{1}{2} + \frac{5}{6} \times \frac{1}{3} \times \frac{1}{2} = \frac{9}{36} = 0.25$

($\because$ Events are mutually exclusive and independent)

(v) P (at least one of the them will pass the examination) = $P(A \cup B \cup C)$

$= 1 - P(\text{none of them will pass the exam})$

$= 1 - P(\bar{A})\,P(\bar{B})\,P(\bar{C})$

$= 1 - \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} = 1 - 0.2777 = 0.7222$

(vi) P(all will pass the examination) = $P(A \cap B \cap C)$

$= P(A)P(B)P(C)$

$= \frac{1}{6} \times \frac{1}{3} \times \frac{1}{2} = 0.02777$

(vii) P(none will pass the examination) = $P(\bar{A} \cap \bar{B} \cap \bar{C})$

$= P(\bar{A})\,P(\bar{B})\,P(\bar{C})$

$= \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} = 0.2777$

(viii) P(B will pass but not A and C) = $P(\bar{A} \cap B \cap \bar{C})$

$= P(\bar{A})\,P(B)\,P(\bar{C})$

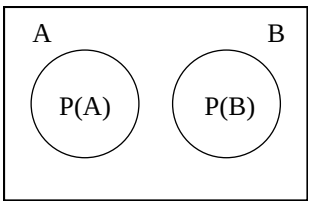
$= \frac{5}{6} \times \frac{1}{3} \times \frac{1}{2} = 0.1388$

Additive Law of Probability (Addtion Theorem of Probability)

Case II: when events are mutually exclusive

Let A and B be two mutually exclusive events. Then the probability of the occurrence of either event A or event B is the sum of their individual probabilities. Hence, the probability of occurance of either event A or event B is given by

P (A or B) = P (A∪B) = P (A) + P (B)



If A, B and C are three mutually exclusive events, then the probability of occurrence of either events A or B or C is given by

P (A or B or C) = P (A∪B∪C) = P (A) + P (B) + P(C)

In general, if events A_1, A_2,.... .,A_n are n mutually exclusive then

P ($A_1 \cup A_2 \cup$ ..... $\cup A_n$) = $P(A_n) + P(A_2) + P(A_3) + \ldots\ldots\ldots\ldots + P(A_n)$

Example 22

The probability that a company execute will travel by plane is $\frac{2}{3}$ and that he will travel by train is $\frac{1}{5}$. Find the probability of his travelling by plane or train.

Solution: Let A and B be the events that a company execute will travel by plane and train respectively.

$P(A) = \frac{2}{3}, \quad P(B) = \frac{1}{5}, P(plne \ \ or \ train) = \ P(A \ or \ B) = ?$

P ($A \cup B$) = P (A) + P (B)

$= \frac{2}{3} + \frac{1}{5} = \frac{10+3}{15} = \frac{13}{15}$

Example 23

If an experiment has the three possible and mutually exclusive outcomes A, B and C, check in each case whether the assignment of probabilities is permissible.

(a) P (A) = 1/3, P (B) = 1/3, P(C) = 1/3

(b) P (A) = 0.35, P (B) = 0.52, P(C) = 0.26

Solution:

(a) P (A) + P (B) + P(C) = $\frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$

Hence, this probability is permissible since total probability = 1.

(b) P (A) + P (B) + P(C) = 0.35 + 0.52 + 0.26 = 1.13 > 1

Which is impossible. Hence, this probability is not permissible.

Example 24

Three events A, B and C are mutually exclusive events and their respective probabilities are as follows.

P (A) = 2/3; P (B) = 1/4; P(C) = 1/6.

Comment on the result.

Solution:

If A, B and C are mutually exclusive events then

P (A or B or C) = P (A) + P (B) + P(C)

$$= \frac{2}{3} + \frac{1}{4} + \frac{1}{6} = \frac{8 + 3 + 2}{12} = \frac{13}{12}.$$

P (A or B or C) = $\frac{13}{12}$ = 1.08 > 1, which is not possible. Hence, the given information is not correct.

Example 25

Suppose that a manager of a large apartment complex provides the following subjective probability estimates about the number of vacancies that will exist next month.

Vacancies	0	1	2	3	4	5
Probability	0.05	0.15	0.35	0.25	0.10	0.10

List the sample points in each of the following events and provide the probability of the event.

(a) no vacancies (b) at least four vacancies (c) two or fewer vacancies

Solution:

Suppose the number of vacancies from 0 to 5 is denoted by A_0 to A_5 respectively.

(a) Probability of selecting no vacancies, $P(A_0) = 0.05$

(b) Probability of selecting at least four vacancies,

$$P(A_4 \cup A_5) = P(A_4) + P(A_5)$$

$$= 0.10 + 0.10 = 0.2$$

(c) Probability of selecting two or fewer vacancies,

$$P(A_2 \cup A_1 \cup A_0) = P(A_2) + P(A_1) + P(A_0)$$

$$= 0.35 + 0.15 + 0.05 = 0.55$$

Example 26

A card is drawn at random from a pack of 52 cards. Find the probability of drawing (i) a jack , a queen or a king (ii) a card which is neither a jack , a queen nor a king.

Solution:

(i) Let A, B & C be the events of drawing a jack, a queen and a king respectively. Then $P(A) = \frac{4}{52}, P(B) = \frac{4}{52}, P(c) = \frac{4}{52}$

∴ Probability of drawing a jack, a queen or a king

= P (a jack, a queen or a king) = P (A or B or C)

$$= P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$= \frac{4}{52} + \frac{4}{52} + \frac{4}{52} = \frac{12}{52} = \frac{3}{13}$$

(ii) P (Neither a jack, a queen nor a king) $= 1 -$ P (a jack, a queen or a king)

$$= 1 - \frac{3}{13} = \frac{10}{13}$$

Multiplicative Law of Probability (or Multiplication Theorem of Probability)

(And, both, as well as, all, three, two) "Intersection , ∩"

Case (i): When events are independent

Let A and B are two independent events, then the probability of occurrence of both the events is the product of their individual probabilities.Symbolically,

$$P(A \text{ and } B) = P(A \cap B) = P(A)\,P(B)$$

If A, B and C are three independent events, then

$$P(A \cap B \cap C) = P(A).P(B).P(C)$$

In general, if $A_1, A_2, \ldots .. , A_n$ are n independent events.Then

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\,P(A_2) \ldots . . P(A_n)$$

Where, $P(A \cap B)$ = Joint probability of events A and B

$P(A \cap B \cap C)$ = Joint probability of events A, B and C

$P(A_1 \cap A_2 \cap \ldots \cap A_n)$ = Joint probability of events $A_1, A_2, \ldots . . , A_n$

and P(A), P(B), P(C), $P(A_1)$, $P(A_2)$ etc. are the marginal probabilities of occurrence of events A, B, C, A_1, A_2 etc. respectively.

Example 27

Two brothers: Mr. X and Mr. Y appear in an interview for getting the scholarship. The scholarship can be provided for two persons.The probability of getting scholarship by Mr. X is $\frac{1}{7}$ and getting by Mr. Y is $\frac{1}{5}$. What is the probability that,

(a) both of them will get scholarship.

(b) Only one of them will get scholarship.

(c) None of them will get scholarship.

Solution:

Let P (A) = Probability that Mr. X will get scholarship

$P(\bar{A})$ = Probability that Mr. X will not get scholarship

P (B) = Probability that Mr. Y will get scholarship

$P(\bar{B})$= Probability that Mr. X will not get scholarship

Given that,

$$P(A) = \frac{1}{7} \text{ and } P(\bar{A}) = 1 - \frac{1}{7} = \frac{6}{7}$$

$P(B) = \frac{1}{5}$ and $P(\bar{B}) = 1 - \frac{1}{5} = \frac{4}{5}$

(a) Probability that both of them will get scholarship,

$P(A \cap B) = P(A)\,P(B)$ $\quad \because A$ & B Events A & B are independent

$$= \frac{1}{7} \times \frac{1}{5} = \frac{1}{35}$$

(b) Probability that only one of them will get scholarship

$= P(A \cap \bar{B}\ or\ \bar{A} \cap B)$

$= P(A \cap \bar{B}) + P(\bar{A} \cap B)$

$= P(A).\,P(\bar{B}) + P(\bar{A}).\,P(B)$

$$= \frac{1}{7} \times \frac{4}{5} + \frac{6}{7} \times \frac{1}{5} = \frac{4}{35} + \frac{6}{35} = \frac{10}{35}$$

(c) Probability that non of them will get scholarship

$P(\bar{A} \cap \bar{B}) = P(\bar{A})\,P(\bar{B})$

$$= \frac{6}{7} \times \frac{4}{5} = \frac{24}{35}$$

Example 28

Probability that a man will be alive 25 years hence is 0.3 and the probability that his wife will be alive 25 years hence is 0.4. Find the probability that 25 years hence (i) both will be alive (ii) only the man will be alive (iii) only the woman will be alive (iv) none will be alive (v) at least one of them will be alive. (vi) only one of them will be alive.

Solution:

Let A and B be the events that a man and a woman will be alive 25 years hence respectively. Then,

Probability that a man will be alive 25 years, hence i.e.

$P(A) = 0.3$ and then $P(\bar{A}) = 1 - 0.3 = 0.7$

Similarly, probability that his wife will be alive 25 years, hence i.e.

$P(B) = 0.4$, $P(\bar{B}) = 1 - 0.4 = 0.6$

(i) Required probability that both will be alive 25 years

$P(A \cap B) = P(A) \times P(B) = 0.3 \times 0.4 = 0.12$

(ii) Probability that only the man will be alive

$= P(A \cap \bar{B}) = P(A) \times P(\bar{B}) = 0.3 \times 0.6 = 0.18.$

(iii) Probability that the only woman will be alive

$= P(B \cap \bar{A}) = P(B) \times P(\bar{A}) = 0.4 \times 0.7 = 0.28$

(iv) Probability that none will be alive

$P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = 0.7 \times 0.6 = 0.42$

(v) Required probability that at least one of them will be alive

= P (A∩ $\bar{B}$ or $\bar{A}$ ∩B or A∩B)

= [P (A) × P ($\bar{B}$)] + [P ($\bar{A}$) × P (B)] + [P (A) × P (B)]

= 0.3 × 0.6 + 0.7 × 0.4 + 0.3 × 0.4 = 0.18 + 0.28 + 0.12 = 0.58

OR

Alternatively

P (A∪B) = 1 – P ($\bar{A}$) P ($\bar{B}$)

= 1 – 0.7 × 0.6 = 1 – 0.42 = 0.58

(vi) Probability that only one of them will be alive

= P (A∩ $\bar{B}$ or $\bar{A}$ ∩B) = P (A∩ $\bar{B}$) + $P(\bar{A} \cap B)$

= P (A) × P ($\bar{B}$) + P ($\bar{A}$) × P (B)

= 0.3 × 0.6 + 0.7 × 0.4

= 0.46

Example 29

A problem in statistics is given to three students A, B, and C whose chances of solving it are $\frac{1}{3}$, $\frac{1}{4}$ and $\frac{1}{5}$ respectively. Find the probability that

(a) the problem will be solved (or at least one of them will solve the problem)

(b) only one of them can solve the problem.

(c) none of them will solve the problem

(d) A solves it but B and C cannot.

(e) all three students A, B and C can solve the problem.

Solution:

Given,

Probability that A solves a problem i.e. P (A) = $\frac{1}{3}$ and P ($\bar{A}$) =1−$\frac{1}{3}$ = $\frac{2}{3}$

Probability that B solves a problem i.e. P (B) = $\frac{1}{4}$ and P ($\bar{B}$) = 1−$\frac{1}{4}$ = $\frac{3}{4}$

Probability that C solves a problem i.e. P(C) = $\frac{1}{5}$ and P ($\bar{C}$) =1−$\frac{1}{5}$ = $\frac{4}{5}$

(a) Required probability that the problem will be solved is given by

$P(A \cup B \cup C)$ = P (A) + P (B) + P(C) – P (A∩B) – P (B∩C) – P (C∩A) + P (A∩B∩C)

= P (A) + P (B) + P(C) – P (A) P (B) – P (B) P(C) – P(C) P (A)+ P (A) P (B) P(C)

$$= \frac{1}{3}+\frac{1}{4}+\frac{1}{5}-\frac{1}{3}\times\frac{1}{4}-\frac{1}{4}\times\frac{1}{5}-\frac{1}{5}\times\frac{1}{3}+\frac{1}{3}\times\frac{1}{4}\times\frac{1}{5}=\frac{3}{5} \quad \because A, B \text{ \& } C \text{ Events } A \text{ \& } B \text{ are independent}$$

Alternatively,

Required probability that the problem will be solved is given by

$P(A \cup B \cup C)$ = 1 – P $(\bar{A})$ P $(\bar{B})$ P $(\bar{C})$

$$= 1 - \frac{2}{3}\times\frac{3}{4}\times\frac{4}{5}$$

$$= 1 - \frac{2}{5} = \frac{3}{5}$$

(b) Probability that only one of them can solve the problem.

= P (A$\cap \bar{B} \cap \bar{C}$) + P ($\bar{A} \cap$B$\cap \bar{C}$) + P($\bar{A} \cap \bar{B} \cap$C)

= P (A) P $(\bar{B})$ P $(\bar{C})$ + P $(\bar{A})$ P (B) P $(\bar{C})$ + P $(\bar{A})$ P $(\bar{B})$ P(C)

$$= \frac{1}{3}\times\frac{3}{4}\times\frac{4}{5}+\frac{2}{3}\times\frac{1}{4}\times\frac{4}{5}+\frac{2}{3}\times\frac{3}{4}\times\frac{1}{5}=\frac{12}{60}+\frac{8}{60}+\frac{6}{60}=\frac{26}{60}=0.433$$

(c) Probability that non of them will solve the problem,

P ($\bar{A} \cap \bar{B} \cap \bar{C}$) = P $(\bar{A})$ P $(\bar{B})$ P $(\bar{C})$

$$= \frac{2}{3}\times\frac{3}{4}\times\frac{4}{5}=\frac{2}{5}=0.4$$

(d) Probability that A solve it but B and C cannot, P (A$\cap \bar{B} \cap \bar{C}$) = P (A) × P $(\bar{B})$ × P $(\bar{C})$

$$= \frac{1}{3}\times\frac{3}{4}\times\frac{4}{5}=\frac{1}{5}=0.2$$

(e) Probability that all three students can solve the problem,

P (A$\cap$B$\cap$C) = P (A) × P (B) × P(C) = $\frac{1}{3}\times\frac{1}{4}\times\frac{1}{5}=\frac{1}{60}$

= 0.017

Example 30

A salesperson has 65 percent chance of making a sale to a cutomer. The behaviour of each successive customer is independent. If three customers A, B and C entre together, (i) what is the probability that the salesperson will make a sale to at least one of the customers? (ii) What is the probability that the salesperson will make a sale to all three customers?

Solution:

The probabilities that a salesperson making a sale to customers A, B and C are

P (A) = 0.65, P $(\bar{A})$ = 1- P (A) = 0.35

P (B) = 0.65, P ($\bar{B}$) = 1- P (B) = 0.35

P (C) = 0.65, P ($\bar{C}$) = 1- P (C) = 0.35

(i) Probability that the salesperson will make a sale to at least one of the customers is given by

$P(A \cup B \cup C) =$ P (A) + P (B) + P(C) – P (A∩B) – P (B∩C) – P (C∩A) + P (A∩B∩C)

= P (A) + P (B) + P(C) – P (A) P (B) – P (B) P(C) – P(C) P (A)+ P (A) P (B) P(C)

= 0.65+ 0.65 + 0.65 –0.65×0.65 –0.65×0.65 –0.65× 0.65 + 0.65 × 0.65 × 0.65 = 0.957

∵ The behaviour of each successive customer is independent i.e. Events A & B are independent

Alternatively,

The required Probability that the salesperson will make a sale to at least one of the customers is given by

P(A ∪ B ∪ C) = 1 – P ($\bar{A}$) P ($\bar{B}$) P ($\bar{C}$)

= 1 –0.35× 0.35 × 0.35

= 0.957

(ii) Probability that the salesperson will make a sale to all three customers given by

P (A∩B∩C) = P (A) × P (B) × P(C)

= 0.65 × 0. 65× 0.65

= 0.274

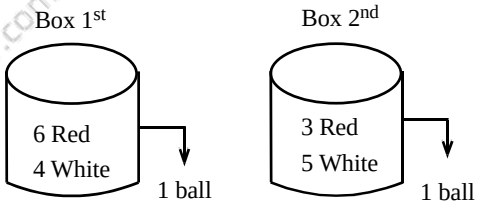
Example 31

A box contains 6 red and 4 white balls. Another box contains 3 red and 5 white balls. One ball is drawn at random from each bag, find the probability that (i) both balls are red; (ii) one is red and one is white; (iii) same colors

Solution:

Total number of balls in 1st box = 6 + 4 = 10

Total number of balls in 2nd box = 3 + 5 = 8



Let,

P (R_1) = Probability of getting a red ball from 1st box = $\frac{6}{10}$

$P(W_1)$ = Probability of getting a white ball from 1st box = $\frac{4}{10}$

$P(R_2)$ = Probability of getting a red ball from 2nd box = $\frac{3}{8}$

$P(W_2)$ = Probability of getting a white ball from 2nd box = $\frac{5}{8}$

Now,

(i) The probability of getting both red balls is given by,

$P(\text{both red balls}) = P(R_1 \cap R_2) = P(R_1) \times P(R_2) = \frac{6}{10} \times \frac{3}{8} = \frac{9}{40}$

(ii) The probability of one is red and one is white is given by,

P (one red and one white) = p (red ball from 1st box and white from 2nd box or white ball from 1st box and red from 2nd box)

$= P(R_1 \cap W_2 \quad \text{or, } W_1 \cap R_2)$

$= P(R_1 \cap W_2) + P(W_1 \cap R_2)$

$= P(R_1) \times P(W_2) + P(W_1) \times P(R_2)$

$= \frac{6}{10} \times \frac{5}{8} \times \frac{4}{10} \times \frac{3}{8} = \frac{21}{40}$

Note: The probability of different colors means probability of getting ball of each color. In above example probability of different colors is given by,

P (different colors) = P (One red and one white)

(iii) The probability of getting balls of same color is given by,

P (same colors) = P (both red or both white)

= P (both red) + P (both white)

$= P(R_1 \cap R_2) + P(W_1 \cap W_2)$

$= P(R_1) \times P(R_2) + P(W_1) \times P(W_2)$

$= \frac{6}{10} \times \frac{3}{8} + \frac{4}{10} \times \frac{5}{8} = \frac{19}{40}$

Example 32

The odds in favour that A speaks the truth are 3: 2 and the odds in favour that B speaks the truth are 5:3. In what percentage of cases are they likely to contradict and do not contradict each other on an identical point?

Solution: Given,

Probability that A speaks the truth = $P(A) = \frac{3}{3+2} = \frac{3}{5}$

Probability that A does not speak truth = P ($\bar{A}$) = $\frac{2}{3+2} = \frac{2}{5}$

Probability that B speaks the truth = P (B) = $\frac{5}{5+3} = \frac{5}{8}$

Probability that B does not speak truth = P ($\bar{B}$) = $\frac{3}{5+3} = \frac{3}{8}$

Required probability that they are likely to contradict each other on an identical point

= P (A$\cap \bar{B}$ or $\bar{A} \cap$B)

= P (A) × P ($\bar{B}$) + P ($\bar{A}$) × P (B)

= $\frac{3}{5} \times \frac{3}{8} + \frac{2}{5} \times \frac{5}{8} = \frac{19}{40}$

∴ Required percentage of contradiction = $\frac{19}{40} \times 100 = 47.5\%$

Required probability that they do not contradict each other on an identical point

= P (A$\cap B$ or $\bar{A} \cap \bar{B}$)

= P (A) P (B) + P ($\bar{A}$) P ($\bar{B}$) ∵ A and B are independent.

= $\frac{3}{5} \times \frac{5}{8} + \frac{2}{5} \times \frac{3}{8} = \frac{21}{40}$ P (A$\cap B$) =P (A) P (B)

= $\frac{21}{40} \times 100 = 52.5\%$

OR

Alternatively,

Required probability that they do not contradict each other on an identical point = $1 - \frac{19}{40} = \frac{21}{40}$

$= \frac{21}{40} \times 100 = 52.5\%$

Example 33

A product is made up of three components C_1, C_2 and C_3. The probability that thses components are defective are in the ratio 1:2:3. Find the probability that one product selected at random is non defective.

Solution

Since, the probability that these components C_1, C_2 and C_3 are defective are in the ratio 1:2:3.

Total ratio = 1+2+3 = 6

Then, $P(C_1) = \frac{1}{6}$, $P(\bar{C}_1) = 1 - \frac{1}{6} = \frac{5}{6}$

$P(C_2) = \frac{2}{6}, \quad P(\bar{C}_2) = 1 - \frac{2}{6} = \frac{4}{6}$

$P(C_1) = \frac{1}{6}, \quad P(\bar{C}_3) = 1 - \frac{3}{6} = \frac{3}{6}$

Hence, the probability that one product selected at randomis is non defective is given by

$P(\bar{C}_1,\ \bar{C}_2\ and\ \bar{C}_3) = P(\bar{C}_1 \cap \bar{C}_2 \cap \bar{C}_3)$

$= P(\bar{C}_1) \times P(\bar{C}_2) \times P(\bar{C}_3)$ $\quad \because$ *the events* $\bar{C}_1, \bar{C}_2\ and\ \bar{C}_3$ are independent

$= \frac{5}{6} \times \frac{4}{6} \times \frac{3}{6} = \frac{5}{18} = 0.277$

Example 34

If the probability of machines M_1, M_2 and M_3 working without failure are 0.2, 0.3 and 0.5 respectively. Find the probability that

(i) at least one machine will work without failure.
(ii) at least two machine will work without failure.
(iii) all machines will work without failure.

Solution:

Let A, B and C denote the events that machines M_1, M_2 and M_3 working without failure respectively. Then

$P(A) = 0.2, \quad P(\bar{A}) = 1 - 0.2 = 0.8$

$P(B) = 0.3, \quad P(\bar{B}) = 1 - 0.3 = 0.7$

$P(C) = 0.5, \quad P(\bar{C}) = 1 - 0.5 = 0.5$

(i) P (at least one machine will work without failure) $= P(A \cup B \cup C)$

$= 1 - P(none\ of\ them\ will\ pass\ the\ exam)$

$= 1 - P(\bar{A})\ P(\bar{B})\ P(\bar{C})$

$= 1 - 0.8 \times 0.7 \times 0.5 = 1 - 0.28 = 0.72$

(ii) P (at least two machine will work without failure) = $P(A \cap B \cap C\ or\ A \cap B \cap \bar{C}\ or\ A \cap \bar{B} \cap C\ or\ \bar{A} \cap B \cap C)$

$= P(A)P(B)P(C) + P(A)P(B)P(\bar{C}) + P(A)\ P(\bar{B})\ P(C) + P(\bar{A})\ P(B)P(C)$

$= 0.2 \times 0.3 \times 0.5 + 0.2 \times 0.3 \times 0.5 + 0.2 \times 0.7 \times 0.5 + 0.8 \times 0.3 \times 0.5 = 0.25$

($\because$ Events are mutually exclusive and independent)

(iii) p (all machines will work without failure) = $P(A \cap B \cap C)$

$= P(A)P(B)P(C)$

$= 0.2 \times 0.3 \times 0.5 = 0.03$

Example 35

A, B and C will pass a certain examination in the proportion 2:4:6. What is the probability that

(i) at least two of them will pass the examination?
(ii) at least one of the them will pass the examination?
(iii) all will pass the examination?

(iv) none will pass the examination?

(v) B will pass but not A and C?

Solution:

Since, the pass proportion of A, B and C in a certain examination are in the rato 2:4:6.

Total ratio = 2 + 4 + 6 = 12. Then

$P(A) = \frac{2}{12} = \frac{1}{6}, \quad P(\bar{A}) = 1 - \frac{1}{6} = \frac{5}{6}$

$P(B) = \frac{4}{12} = \frac{1}{3}, \quad P(\bar{B}) = 1 - \frac{1}{3} = \frac{2}{3}$

$P(C) = \frac{6}{12} = \frac{1}{2}, \quad P(\bar{C}) = 1 - \frac{1}{2} = \frac{1}{2}$

(i) P (at least two of them will pass the examination) = $P(A \cap B \cap C$ or $A \cap B \cap \bar{C}$ or $A \cap \bar{B} \cap C$ or $\bar{A} \cap B \cap C)$

$= P(A)P(B)P(C) + P(A)P(B)P(\bar{C}) + P(A)\,P(\bar{B})\,P(C) + P(\bar{A})\,P(B)P(C)$

$= \frac{1}{6} \times \frac{1}{3} \times \frac{1}{2} + \frac{1}{6} \times \frac{1}{3} \times \frac{1}{2} + \frac{1}{6} \times \frac{2}{3} \times \frac{1}{2} + \frac{5}{6} \times \frac{1}{3} \times \frac{1}{2} = \frac{9}{36} = 0.25$

($\because$ Events are mutually exclusive and independent)

(ii) P (at least one of the them will pass the examination) = $P(A \cup B \cup C)$

= 1– P(none of them will pass the exam)

= 1 – P $(\bar{A})$ P $(\bar{B})$ P $(\bar{C})$

$= 1 - \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} = 1 - 0.2777 = 0.7222$

(iii) P(all will pass the examination) = $P(A \cap B \cap C)$

$= P(A)P(B)P(C)$

$= \frac{1}{6} \times \frac{1}{3} \times \frac{1}{2} = 0.02777$

(iv) P(none will pass the examination) = $P(\bar{A} \cap \bar{B} \cap \bar{C})$

$= P(\bar{A})\,P(\bar{B})\,P(\bar{C})$

$= \frac{5}{6} \times \frac{2}{3} \times \frac{1}{2} = 0.2777$

(v) $P(B \text{ will pass but not } A \text{ and } C) = P(\bar{A} \cap B \cap \bar{C})$

$= P(\bar{A})\,P(B)\,P(\bar{C})$

$= \frac{5}{6} \times \frac{1}{3} \times \frac{1}{2} = 0.1388$

Multiplication Theorem of Probability

(And, both, as well, two, three, all, $\cap$), $\times$

Case (ii): When the events are dependent:

If A and B are two dependent events, then the probability of simultaneous happening of two events A and B is given by

$$P(A \cap B) = P(A)P(B/A)$$

Similarly,

$$P(A \cap B) = P(B)P(A/B)$$

Where, $P(B/A)$ is the conditional Probability of the occurrence (happening) of event B given that (if) event A has already occurred (happened).

& $P(A/B)$ is the conditional Probability of occurrence (happening) of event A given that event B has already occurred (happened).

If A, B and C are three dependent events, then

$$P(A \cap B \cap C) = P(A)P(B/A)P(C/A \cap B)$$

Where, $P(C/A \cap B)$ is the conditional probability of occurrence (happening) of event C given that (if) both events A and B have already occurred (happened).

In general, for n events $A_1, A_2, A_3, A_3, \ldots\ldots\ldots\ldots, A_n$, we have

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)P(A_2/A_1)P(A_3/A_1 \cap A_2)\ldots\ldots\ldots\ldots P(A_n/A_1 \cap A_2 \cap A_3 \cap \ldots\ldots \ldots \cap A_{n-1})$$

Note (i) If events A and B are independent, then P (A/B) = P (A) and P (B/ A) = P (B)

(ii) P (A∩B) can also be denoted by P (AB)

Conditional Probability

Definition: Conditional probability is the probability that an event will occur given that another event has already occurred. If A and B are two dependent events, then the conditional probability of event A given that (if) event B has already occurred is given by,

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) \neq 0$$

$$\Rightarrow P(A/B) = \frac{\frac{\text{No.of favourable cases to } A \cap B}{\text{Total number of exhaustive cases}}}{\frac{\text{No.of favourable cases to B}}{\text{Total number of exhaustive cases}}}$$

Similarly, the conditional probability of event B given that (if) event A has already occurred is given by $P(B/A) = \frac{P(A \cap B)}{P(A)}$, provided $P(A) \neq 0$

$$\Rightarrow P(B/A) = \frac{\frac{\text{No.of favourable cases to } A \cap B}{\text{Total number of exhaustive cases}}}{\frac{\text{No.of favourable cases to A}}{\text{Total number of exhaustive cases}}}$$

Where, $P(A)$ = Marginal (or simple) probability of an event A

& $P(A \cap B)$ = Joint probability of events A and B.

Example 36

The probability that a manufacturer will produce 'brand X' product is 0.13, the probability that he will produce 'brand Y' product is 0.28 and the probability that he will produce both brand is 0.06. What is the probability that the manufacturer who has produced 'brand Y' will also have produced 'brand X'?

Solution: Let X and Y be the events that a manufacturer will produce brand X and brand Y respectively.

P(X) = 0.13, P(Y) = 0.28, P $(X \cap Y) = 0.06$, $P(X/Y) =$?

$P(X/Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{0.06}{0.28} = 0.214$

∴ The probability that the manufacturer who has produced 'brand Y' will also have produced 'brand X' is 0.214

Example 37

In a certain school, 20% students failed in English, 15% students failed in Mathematics and 10% students failed in both English and Mathematics. A student is selected at random. If he failed in English, what is the probability that he also failed in Mathematics.

Solution:

Let E and M be the events that denote the students failed in English and failed in Mathematics respectively. Then

P (E) = 20% = 0.2

P (M) = 15% = 0.15

P (E∩M) = 10% = 0.10

If one student is selected at random, the Probability that if he failed in English then he also failed in Mathematics is given by

$P(M/E) = \frac{P(M \cap E)}{P(E)} = \frac{0.10}{0.2} = \frac{1}{2}$

Example 38

In a group of equal number of men and women, 75 % men and 60% women are employed. A person is selected at random and found to be unemployed. A person is selected at random and found to be unemployed. What is the probability that selected person is (i) men (ii) women?

Solution:

Since, the number of men and women are equal and their employment percentage is given. Therefore, the given information can be presented in the following tabular form:

Nature / Sex	Employed	Unemployed	Total
Men	75	25	100
Women	60	40	100
Total	135	65	200

Let U = an event of selecting unemployed

M = an event of selecting man

W = an event of selecting woman

(i) probability that the selected person is a man, given that (if) he is unemployed is

$$P(M/U) = \frac{P(M \cap U)}{P(U)} = \frac{P(M \text{ and } U)}{P(U)} = \frac{25/200}{65/200} = \frac{5}{13}$$

(ii) probability that the selected person is a woman, given that (if) he is unemployed is

$$P(W/U) = \frac{P(W \cap U)}{P(U)} = \frac{P(W \text{ and } U)}{P(U)} = \frac{40/200}{65/200} = \frac{8}{13}$$

Example 39

The personnel department of a company has records which show the following of its 120 engineers

Age (Years)	Bachelor degree only (B.E.)	Master degree only (M.E.)	Total
Under 25	60	8	68
25 to 30	7	26	33
over 30	9	10	19
Total	76	44	120

If one engineer is selected at random, find the probability that

(i) he has only a bachelor's degree
(ii) he is under 25 years.
(iii) he has a master's degree, given that he is over 30.
(iv) he is 25 to 30 years given that he has bachelor's degree.
(v) he is over 30 and has a master's degree.

Solution:

Total number of engineers = 120

Let , the event A = an engineer chsen is bachelor's degree

B = an engineer chsen is Mastrer's degree

C = an engineer chsen is under age 25 years

D = an engineer chsen is 25 to 30 years

E = an engineer chsen is over 30 years

If one engineer is selected at random, then

(i) P(A) = Probability that he has only a bachelor's degree

$$= \frac{76}{120} = \frac{19}{30}$$

(ii) P (B) = Probability that he he is under 25 years.

$$= \frac{68}{120} = \frac{17}{30}$$

(iii) P (B/E) = Probability that he has a master's degree, given that he is over 30.

$$= \frac{P(B \cap E)}{P(E)} = \frac{P(B \text{ and } E)}{P(E)} = \frac{10/120}{19/120} = \frac{10}{19}$$

(iv) P (D/A) = Probability that he is 25 to 30years, given that he has a bachelor's degree.

$$= \frac{P(D \cap A)}{P(A)} = \frac{P(D \text{ and } A)}{P(A)} = \frac{7/120}{76/120} = \frac{7}{76}$$

(v) P (E∩B) = P (E and B) = Probability that he is over 30 and has a master's degree

$$= \frac{10}{120} = \frac{1}{12}$$

Example 40

In a survey of 1000 customers about the quality and timely deliverance of "The Momo world" Marked by a company following information prevailed

Quality satisfaction		Timely deliverance		
		Yes	No	Total
	Yes	500	225	725
	No	200	75	275
	Total	700	300	1000

a. What is the probability that a customer satisfied with qulity but timely not delivered?
b. What is the probability that a customer timely delivered but not satisfied with quality?
c. What is the probability that a customer timely delivered and satisfied with quality?

Solution:

Let A = event that the customers satisfied with quality.

$\bar{A}$ = event that the customers not satisfied with quality

B = event that the customers with timely deliverance

$\bar{B}$ = event that the customers with timely not delivered

a. The probability that a customer satisfied with qulity but timely not delivered is given by

$$P(A/\bar{B}) = \frac{P(A \cap \bar{B})}{P(\bar{B})} = \frac{P(A \text{ and } \bar{B})}{P(\bar{B})} = \frac{225/1000}{300/1000} = \frac{225}{300} = 0.75$$

b. The probability that a customer timely delivered but not satisfied with quality is given by

$$P(B/\bar{A}) = \frac{P(\bar{A} \cap B)}{P(A)} = \frac{P(\bar{A} \text{ and } B)}{P(A)} = \frac{200/1000}{275/1000} = \frac{200}{275} = 0.7273$$

c. The probability that a customer timely delivered and satisfied with quality is given by

$$P(A \cap B) = P(A \text{ and } B) = \frac{5000}{1000} = 0.5$$

Example 41

What is the probability that a couple's second child will be

(a) a boy, given that their first child was a girl?

(b) a girl, given that their first child was a girl?

Solution:

Let B_1 and B_2 be the events that denote the first child is boy and second child is also boy respectively.

Similarly, G_1 and G_2 be the events that denote that first child is girl and second child is also girl respectively. There are only two possibilities; either boy child or girl child. Then

Probability of first boy child, $P(B_1) = 1/2$

Probability of first girl child, $P(G_1) = 1/2$

(a) Probability of being second child a boy given that their first child was a girl

$$P(B_2/G_1) = \frac{P(B_2 \cap G_1)}{P(G_1)}$$

$$= \frac{P(B_2)\,P(G_1)}{P(G_1)} \qquad \because B_2 \text{ and } G_1 \text{ are independent}$$

$$= P(B_2) = \frac{1}{2}$$

(b) Similarly, the probability of being second child a girl given that their first child was a girl,

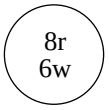
$$P(G_2/G_1) = \frac{P(B_2 \cap G_1)}{P(G_1)} = \frac{P(G_2)\,P(G_1)}{P(G_1)} \quad \because \text{First birth and second birth are independent}$$

$$= P(G_2) = \frac{1}{2}$$

Example 42

A bag contains 8 red and 6 white balls. Two balls are drawn randomly from the bag one after other without replacement. Find the probability that both balls are white.

Solution:



Probability of drawing a white ball in the first draw = $P(W_1) = \frac{6}{14}$

Probability of drawing a white ball in the second draw given that first drawn ball is white

$= P(W_2/W_1) = \frac{5}{13}$

Since, drawn is made without replacement. (i.e. dependent case)

Required probability of drawing both white balls, $P(W_1 \cap W_2) = P(W_1)\,P(W_2/W_1)$

$$= \frac{6}{14} \times \frac{5}{13} = 0.1648$$

Example 43

At a soup kitchen, a social worker gathers the following data. Of these visiting the kitchen, 59% are men, 32% are alcoholics, and 21% are male alcoholics. What is the probability that a random male visitor to the kitchen is an alcoholic?

Solution:

Let P (A) = Probability of person is a male = 0.59

P (B) = Probability of person is an alcoholic = 0.32

P (A∩B) = Probability of a person is a male alcoholic = 0.21

The probability that a random male visitor to the kitchen is an alcoholic is given by

$$P(B/A) = \frac{P(A \cap B)}{P(A)} = \frac{0.21}{0.59} = 0.3559$$

Example 44

Find the probability of drawing an ace, a king and a queen in that order from of a pack of cards in three consecutive draws, cards drawing not being replaced.

Solution:

Let A, K and Q denote the event of drawing an ace, a king and a queen respectively.

The probability of drawing an ace, a king and a queen in that order in three consecutive draws, when the drawn cards is not replaced (without replacement) in the pack of cards is given by

$$P(A \cap K \cap Q) = P(A) \times P(K/A) \times P(Q/A \cap K)$$

Where,

P (A) = Probability of drawing an ace card = $\frac{4}{52}$

P (K/A) = Probability of drawing a king card given that an ace has been already drawn = $\frac{4}{51}$

P (Q/A∩K) = Probability of drawing a queen card given that an ace and a king have been already drawn = $\frac{4}{50}$

$\therefore P(A \cap K \cap Q) = P(A) \times P(K/A) \times P(Q/A \cap K)$

$$= \frac{4}{52} \times \frac{5}{51} \times \frac{4}{50} = 4.826 \times 10^{-4}$$

Example 45

If A and B are two events such that P (A) = 0.4, P (B) = 0.6 and $P(B/A) = 0.5$, find $P(A/B)$ and

P (A ∪ B).

Solution: P (A) = 0.4, P (B) = 0.6

P (B/A) = 0.5

We have, $P(B/A) = \frac{P(A \cap B)}{P(A)}$

or, $0.5 = \frac{P(A \cap B)}{0.4}$

$\therefore P(A \cap B) = 0.5 \times 0.4 = 0.2$

Now,

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{0.2}{0.6} = 0.334$$

$$\therefore P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$= 0.4 + 0.6 - 0.2 = 0.8$$

Example 46

A bag contains 5 white and 3 black balls. Two balls are drawn at random one after the another without replacement. Find the probability that both balls drawn are (i) white, (ii) black, (iii) different colors (or one white and one black) (iv) same colors.

Solution:

(i) Let

W_1 = Event of drawing white ball in the first draw

W_2 = Event of drawing white ball in the second draw

Then,

The probability of drawing a white ball in the first draw is

$$P(W_1) = \frac{5}{8}$$

The probability of drawing a white ball in the second draw given that the first ball drawn is white is

$$P(W_2/W_1) = \frac{4}{7}$$

Thus, the probability that both balls drawn are white is given by

$$P(W_1 \cap W_2) = P(W_1) \times P(W_2/W_1) = \frac{5}{8} \times \frac{4}{7} = \frac{5}{14}$$

(Since, drown is made without replacement i.e. W_1 & W_2 are dependent events)

(ii) Let

B_1 = Event of drawing black ball in the first draw

B_2 = Event of drawing black ball in the second draw

Then,

The probability of drawing a black ball in the first draw is

$$P(B_1) = \frac{3}{8}$$

The probability of drawing a black ball in the second draw given that the first ball drawn is black is

$$P(B_2/B_1) = \frac{2}{7}$$

The probability that both balls drawn are black is given by

$$P(B_1 \cap B_2) = P(B_1) \times P(B_2/B_1) = \frac{3}{8} \times \frac{2}{7} = \frac{3}{28}$$

(Since, drawn is made without replacement i.e. W_1 & W_2 are dependent events)

(iii) The probability of different colors is given by

P (different colors) = P (one white and one black)

= P (frist drawn ball is white and second drawn ball is black or first drawn ball is black and second drawn ball is white)

$= P(W_1 \cap B_2 \text{ or } B_1 \cap W_2)$

$= P(W_1 \cap B_2) + P(B_1 \cap W_2)$

$= P(W_1) \times P(B_2/W_1) + P(B_1) \times P(W_2/B_1)$

$$= \frac{5}{8} \times \frac{3}{7} + \frac{3}{8} \times \frac{5}{7} = \frac{30}{56}$$

(vii) The probability of same colors is given by

P (same colors)

= P (both are white or both are black)

= P (both are white) + P (both are black)

$= P(W_1 \cap W_2) + P(B_1 \cap B_2)$

$= P(W_1) \times P(W_2/W_1) + P(B_1) \times P(B_2/B_1)$

$$= \frac{5}{8} \times \frac{4}{7} + \frac{3}{8} \times \frac{2}{7} = \frac{26}{56}$$

Example 47

A bag contains 5 white and 8 red balls. Two drawing of 3 balls are made such that (i) the balls are replaced before the second trial and (ii) the balls are not replaced before the second trial. Find the probability that the first drawing will give 3 white and the second 3 red balls in each case.

Solution:

Total number of balls = 5 + 8 = 13 balls

Let A and B denote the events of drawing 3 white balls in the first draw and 3 red balls in the second draw respectively.

(i) **With replacement:** If the balls drawn in the first draw are replaced back in the bag before the second draw, then events A and B are independent. Then the probability of drawing 3 white balls in the first draw and 3 red balls in the second draw is given by

$$P(A \cap B) = P(A) \times P(B) = \frac{{}^5C_a}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{13}C_3} = \frac{10}{286} \times \frac{56}{286} = 0.0068$$

(ii) **Without replacement :** If the balls drawn in the first draw are not replaced back in the bag before the second draw, then events A and B are dependent. The probability of drawing 3 white balls in the first draw and 3 red balls in the second draw is given by

$P(A \cap B) = P(A) \times P(B/A)$ .....(I)

Where,

P (A) = Probability of drawing 3 white balls in the first draw = $\frac{{}^5C_3}{{}^{13}C_3}$

Now, if 3 white balls drawn in the first draw are not replaced back in the bag, there are 13 – 3 = 10 balls left in which 5 – 3 = 2 are white and 8 are red balls.

P (B/A) = Probability of drawing 3 red balls from the bag containing 2 white and 8 red balls

$= \frac{{}^8C_3}{{}^{10}C_3}$

$\therefore P(A \cap B) = P(A)\,P(B/A) = \frac{{}^5C_3}{{}^{13}C_3} \times \frac{{}^8C_3}{{}^{10}C_3} = \frac{10}{286} \times \frac{56}{120} = 0.0163$

Theoretical Probability Distribution

Probability distributions are classified as either discrete or continuous. A discrete probability distribution is allowed to take on only a limited number of values (i.e. whole number values only) whereas a continuous probability distribution is allowed to take on any value within a given range. In this chapter, we shall study the following "univariate" probability distributions:

(1) Binomial Distribution
(2) Poisson Distribution
(3) Normal Distribution

The first two distributions are discrete probability distributions and the last one is continuous probability distribution.

Binomial Distribution

Binomial distribution is the most fundamental and important discrete probability distribution in probability theory and Statistics. The binomial distribution is the basis for the popular binomial test of statistical significance.

Probability function of Binomial Distribution

Definition: A random variable X is said to follow binomial distribution if it assumes only non- negative values and its probability mass function (p.m.f.) is given by

- $P(X = r) = p(r) = \begin{cases} {}^nC_r\, p^r q^{n-r}, & r = 0, 1, 2, 3, \ldots\ldots, n., \quad q = 1 - p \\ 0 & \text{otherwise} \end{cases}$

,

- The mean, variance and standard deviation of binomial distribution are

  Mean= $E(X) = np$

  Variance $= V(X) = npq$

  and standard deviation$(\sigma) = \sqrt{npq}$

$\therefore$ ***Mean*** $>$ ***Variance,*** **always holds in binomial distribution.**

[redacted]

Determine the binomial distribution for which the mean 4 and variance is 3.

Solution

Mean (np) = 4 . . . . . .(i)

Variance (npq) = 3 . . . .(ii)

X~ $B(n,p) = ?$

From (ii)

npq = 3

or, $4 \times q = 3$

or, $q = \frac{3}{4}$ and p = $1 - q = 1 - \frac{3}{4} = \frac{1}{4}$

Now, from (i)

np = 4 or, $n \times \frac{1}{4} = 4$

$\therefore$ n = 16

Hence, the required binomial distribution i.e. X~ $B\left(16, \frac{1}{4}\right)$

[redacted]

Point out the fallacy, if any, in the following statement:

For binomial distribution, mean = 7 and variance = 11.

Solution:

For a binomial distribution with parameters n and p

Mean = np = 7 . . . . . (i)

Variance = npq =11 . . . . . (ii)

Dividing (ii) by (i), we get

$\frac{npq}{np} = \frac{11}{7}$

q = $\frac{11}{7} = 1.6 > 1$ which is impossible, since q being the probability must lie between 0 and 1 i.e. $0 < q < 1$. Hence, the given statement is wrong (incorrect)

[redacted]

Ten unbiased coins are tossed simultaneously. Find the probability of obtaining

Exactly 6 heads (ii) At least 8 heads (iii) No head (iv) at least one head (v) Not more than three heads (vi) At least 4 heads (vii) At most 3 heads (viii) More than 2 heads (ix) Less than 8 heads (x) 3 heads and 7 tails.

Solution: Let P be the probability of getting a head, then P = $\frac{1}{2}$ and q = $1 - \frac{1}{2} = \frac{1}{2}$

Number of trials (n) = 10

Let r denotes the number of heads or tails. Then, the probability of getting r heads in binomial distribution is given by

$P(X = r) = p(r) = n_{C_r}\, p^r q^{n-r}$, $r = 0, 1, 2, 3, \ldots\ldots, n.$

$\therefore P(X = r) = 10_{C_r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{10-r} = 10_{C_r} \times \frac{1}{1024}$

(i) P(exactly 6 heads) = $P(X = 6)$

$= 10_{C_6} \times \frac{1}{1024} = \frac{210}{1024} = 0.205$

(ii) P(at least 8 heads) = $P(X \geq 8)$

$= P(X = 8) + P(X = 9) + P(X = 10)$

$= 10_{C_8} \times \frac{1}{1024} + 10_{C_9} \times \frac{1}{1024} + 10_{C_{10}} \times \frac{1}{1024}$

$= \frac{1}{1024}\left[10_{C_8} + 10_{C_9} + 10_{C_{10}}\right]$

$= \frac{1}{1024}(45 + 10 + 1) = \frac{56}{1024} = \frac{56}{128} = 0.4375$

(iii) P(no head) = $P(X = 0)$

$= 10_{C_0} \times \frac{1}{1024} = 1 \times \frac{1}{1024} = 0.00097$

(iv) P(at least one head) = $P(X \geq 1) = 1 -$ p(no head)

$= 1 - p(X < 1) = 1 - p(X = 0)$

$= 1 - 0.00097 = 0.99903$

(v) P(not more than three heads) = $P(X \leq 3)$

$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$= 10_{C_0} \times \frac{1}{1024} + 10_{C_1} \times \frac{1}{1024} + 10_{C_2} \times \frac{1}{1024} + 10_{C_2} \times \frac{1}{1024}$

$= \frac{1}{1024}\left[10_{C_0} + 10_{C_1} + 10_{C_2} + 10_{C_3}\right]$

$= \frac{1}{1024}[1 + 10 + 45 + 120] = \frac{176}{1024} = \frac{11}{64} = 0.17187$

(vi) P(at least 4 heads) = $P(X \geq 4)$

$= 1 - P(X < 4)$

$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\}$

$= 1 - \{10_{C_0} \times \frac{1}{1024} + 10_{C_1} \times \frac{1}{1024} + 10_{C_2} \times \frac{1}{1024} + 10_{C_3} \times \frac{1}{1024}\}$

$$= 1 - \frac{1}{1024}\left[10_{C_0} + 10_{C_1} + 10_{C_2} + 10_{C_3}\right]$$
$$= 1 - \frac{1}{1024}(1 + 10 + 45 + 120) = 1 - \frac{176}{1024} = 0.8281$$

(vii) P(at most 3 heads) = $P(X \leq 3)$

$$= \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\}$$

$$= 10_{C_0} \times \frac{1}{1024} + 10_{C_1} \times \frac{1}{1024} + 10_{C_2} \times \frac{1}{1024} + 10_{C_3} \times \frac{1}{1024}$$
$$= \frac{1}{1024}\left[10_{C_0} + 10_{C_1} + 10_{C_2} + 10_{C_3}\right]$$
$$= \frac{1}{1024}(1 + 10 + 45 + 120)$$
$$= \frac{176}{1024} = 0.1718$$

(viii) P(more than 2 heads) = $P(X > 2)$

$$= 1 - P(X \leq 2)$$

$$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

$$= 1 - \{10_{C_0} \times \frac{1}{1024} + 10_{C_1} \times \frac{1}{1024} + 10_{C_2} \times \frac{1}{1024}\}$$
$$= 1 - \frac{1}{1024}\left[10_{C_0} + 10_{C_1} + 10_{C_2}\right]$$
$$= 1 - \frac{1}{1024}(1 + 10 + 45) = 1 - \frac{56}{1024} = \frac{968}{1024} = 0.9453$$

(ix) P(less than 8 heads) = $P(X < 8)$

$$= 1 - P(X \geq 8)$$

$$= 1 - \{P(X = 8) + P(X = 9) + P(X = 10)\}$$

$$= 1 - \{10_{C_8} \times \frac{1}{1024} + 10_{C_9} \times \frac{1}{1024} + 10_{C_{10}} \times \frac{1}{1024}\}$$
$$= 1 - \frac{1}{1024}\left[10_{C_8} + 10_{C_9} + 10_{C_{10}}\right]$$
$$= 1 - \frac{1}{1024}(45 + 9 + 1) = 1 - \frac{55}{1024} = \frac{969}{1024} = 0.9462$$

(x) P(3 heads and 7 tails) = $P(X = 3)$

$$= P(X = 3)$$

$$= 10_{C_3} \times \frac{1}{1024}$$
$$= \frac{120}{1024} = 0.1171$$

[redacted]

The incidence of occupational disease in an industry is such that the workmen have a 20% chance of suffering from it. What is the probability that out of six workmen (i) 4 or more will contract the disease (ii) at most 3 will contract the disease (iii) more than 5 will contract the disease (iv) none is suffering from the disease?

Solution

Here,

Number of workmen (n) = 6

Probability of suffering from an occupational disease (p) = 0.20

Probability of not suffering from an occupational disease (q) = $1 - p = 1 - 0.20 = 0.80$

Let the random variable X denotes the number of workmen contract from the disease, then the probability of r workmen contract from the disease in binomial distribution is given by

$P(X = r) = p(r) = n_{C_r}\, p^r q^{n-r}, \quad r = 0, 1, 2, 3, 4, 5, 6.$

(i) P(4 or more will contract the disease)

$= P(X \geq 4)$

$= P(X = 4) + P(X = 5) + P(X = 6)$

$= 6_{C_4}\,(0.20)^4(0.80)^{6-4} + 6_{C_5}\,(0.20)^5(0.80)^{6-5} + 6_{C_6}\,(0.20)^6(0.80)^{6-6}$

= 0.0154+0.0015+0.000064 = 0.01696

(ii) P(at most 3 will contract the disease)

$= P(X \leq 3)$

$= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$

$= 6_{C_0}\,(0.20)^0(0.80)^{6-0} + 6_{C_1}\,(0.20)^1(0.80)^{6-1} + 6_{C_2}\,(0.20)^2(0.80)^{6-2}$

$+ 6_{C_3}\,(0.20)^3(0.80)^{6-3}$

= 0.2621+0.0393+0.2457+0.0819 = 0.629

(iii) P(more than 5 will contract the disease)

$= P(X > 5)$

$= P(X = 6) = 6_{C_6}\,(0.20)^6(0.80)^{6-6} = 0.000064$

(iv) P(none is suffering from the disease) $= P(X = 0)$

$= 6_{C_0}\,(0.20)^0(0.80)^{6-0} = 0.2621$

A merchant's file of 20 accounts contains 6 delinquent and 14 non –delinquent accounts. An auditor randomly selects 5 of these accounts for examination.

(i) What is the probability that the auditor finds exactly 2 delinquent accounts?
(ii) What is the probability that the auditor finds less than or equal to 3 delinquent accounts?
(iii) Find the expected number of delinquent accounts in the sample selected.

Solution

n = 5

Probability of a delinquent account (p) = $\frac{6}{20} = \frac{3}{10}$

Probability of a non- delinquent account (q) =$1-p = 1 - \frac{3}{10} = \frac{7}{10}$

Let random variable X denotes the number of delinquent accounts that the auditor finds. Then, the probability of getting r delinquent accounts out of n accounts in binomial distribution is given by

$P(X = r) = p(r) = n_{C_r}\, p^r q^{n-r}, \quad r = 0, 1, 2, \ldots\ldots, n \quad \& \quad q = 1-p$

(i) P(exactly 2 delinquent accounts) = $P(X = 2)$

$$= 5_{C_2} \left(\frac{3}{10}\right)^2 \left(\frac{7}{10}\right)^{5-2}$$

$$= 10 \times 0.09 \times 0.343 = 0.3087$$

(ii) P (less than or equal to 3 delinquent accounts) = $P(X \leq 3)$

$$= 1 - P(X > 3)$$

$$= 1 - \{P(X = 4) + P(X = 5)\}$$

$$= 1 - \{5_{C_4} \left(\frac{3}{10}\right)^4 \left(\frac{7}{10}\right)^{5-4} + 5_{C_5} \left(\frac{3}{10}\right)^5 \left(\frac{7}{10}\right)^{5-5}\}$$

$$= 1 - \{0.02835 + 0.00243\}$$

$$= 1 - 0.03078 = 0.96922$$

(iii) The expected number of delinquent accounts in the selected sample of n = 5 is given by

Mean = n p = $5 \times \frac{3}{10} = 1.5$

[redacted]

Out of 800 families of Kathmandu with 4 children each, what percentages would be expected to have (i) 2 boys and 2 girls (ii) at least one boy (iii) no girls (iv) at most 2 girls and (v) one boy and 3 girls. Assume equal probabilities for boys and girls.

Solution

Here,

Number of children (n) = 4

Let the probability of boys (p) = $\frac{1}{2}$

Probability of girls (q) = $\frac{1}{2}$

Let random variable X denotes the number of boys or girls.

Then, the probability of getting r boys out of n children in binomial distribution is given by

$P(X = r) = p(r) = n_{C_r} p^r q^{n-r}, \quad r = 0, 1, 2, \ldots, n.$

$\therefore P(X = r) = 4_{C_r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{4-r} = 4_{C_r} \times \frac{1}{16}$

(i) P (2 boys and 2 girls) = P(X=2) = $4_{C_2} \times \frac{1}{16} = \frac{6}{16} = \frac{3}{8} = 0.375 = 37.5\%$

(ii) P (at least one boy) $= P(X \geq 1)$

$= 1 - P(X < 1)$

$= 1 - P(X = 0)$

$= 1 - 4_{C_0} \times \frac{1}{16}$

$= 1 - \frac{1}{16} = \frac{15}{16} = 0.9375 = 93.75\%$

(iii) P(no girls) = P(all boys)

$= P(X = 0)$

$= {}^{4}C_0 \times \frac{1}{16} = \frac{1}{16} = 0.0625 = 6.25\%$

(iv) P (at most 2 girls) = $P(X \leq 2)$

$$= P(X = 0) + P(X = 1) + P(X = 2)$$

$$= {}^{4}C_0 \times \frac{1}{16} + {}^{4}C_1 \times \frac{1}{16} + {}^{4}C_2 \times \frac{1}{16}$$
$$= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} = \frac{11}{16} = 0.6875 = 68.75\%$$

(v) P(one boy and 3 girls) = $P(X = 1)$

$$= {}^{4}C_1 \times \frac{1}{16} = \frac{4}{16} = \frac{1}{4} = 0.25 = 25\%$$

[redacted]

If hens of a certain breed normally lay eggs on 5 days a week on an average, find how many days during a season of 100 days poultry keeper with 5 hens of this breed, will expect to receive (i) no eggs (ii) exactly 3 eggs (iii) at least 4 eggs (iv) at most 2 eggs.

Solution:

Number of hens (n) = 5

Total number of days (N) = 100

Probability of laying eggs (p) = $\frac{5}{7}$

Probability of not laying eggs (q) = 1 – p = $1 - \frac{5}{7} = \frac{2}{7}$

Let the random variable X denotes the number of laying eggs. Then, the probability of getting r eggs out of n hens in binomial distribution is given by

$P(X = r) = p(r) = {}^{n}C_r \, p^r q^{n-r}, \quad r = 0, 1, 2, \ldots\ldots, n$

(i) P(no eggs) = $P(X = 0)$

$$= {}^{5}C_0 \left(\frac{5}{7}\right)^0 \left(\frac{2}{7}\right)^{5-0} = \left(\frac{2}{7}\right)^5 = 0.0019$$

The expected number of days for receiving no eggs out of 100 days is

$f(0) = N \times P(X = 0) = 100 \times 0.0019 = 0.19 \approx 0$

(ii) P(exactly 3 eggs) = $P(X = 3)$

$$= {}^{5}C_3 \left(\frac{5}{7}\right)^3 \left(\frac{2}{7}\right)^{5-3} = 0.297$$

The expected number of days for exactly 3 eggs out of 100 days is

$f(3) = N \times P(X = 3) = 100 \times 0.297 = 29.7 \approx 30$

(iii) P (at least 4 eggs) = $P(X \geq 4)$

$$= P(X = 4) + P(X = 5)$$

$$= {}^{5}C_4 \left(\frac{5}{7}\right)^4 \left(\frac{2}{7}\right)^{5-4} + {}^{5}C_5 \left(\frac{5}{7}\right)^5 \left(\frac{2}{7}\right)^{5-5}$$

= 0.3719+0.1859 = 0.5578

The expected number of days for at least 4 eggs out of 100 days is

$f(X \geq 4) = N \times P(X \geq 4)$ = $100 \times 0.5578 = 55.78 \approx 56$

(iv) P(at most 2 eggs) = $P(X \leq 2)$

$= P(X = 0) + P(X = 1) + P(X = 2)$

$= 5_{C_0} \left(\frac{5}{7}\right)^0 \left(\frac{2}{7}\right)^{5-0} + 5_{C_1} \left(\frac{5}{7}\right)^1 \left(\frac{2}{7}\right)^{5-1} + 5_{C_2} \left(\frac{5}{7}\right)^2 \left(\frac{2}{7}\right)^{5-2}$

= 0.0019 +0.0238+0.0.1190 = 0.1447

The expected number of days for at most 2 eggs out of 100 days is

$f(X \leq 2) = N \times P(X \leq 2)$ = $100 \times 0.1447 = 14.47 \approx 15$

[redacted]

If mean and variance of binomial distribution are 3 and 1.5 respectively, find the probability (i) at least one successes (ii) exactly 2 successes (iii) at least 4 successes (iv) at most 3 successes (v) more than 3 successes (vi) less than 4 successes (vii) between 2 and 6 successes.

Solution

If n and p are the parameters of the binomial distribution, then we are given:

Mean = np = 3. . . . . (i)

Variance = npq =1.5 . . . . . (ii)

Dividing (ii) by (i), we get

$\frac{npq}{np} = \frac{1.5}{3} = \frac{1}{2}$

$\Rightarrow$ q = $\frac{1}{2}$

p =1$-q = 1 - \frac{1}{2} = \frac{1}{2}$

Substituting in (i), we get

np = 3

$n \times \frac{1}{2} = 3 \Rightarrow$ n = 6

Let random variable X denotes the number of successes.

Then, the probability of getting r successes out of n trials in binomial distribution is given by

$P(X = r) = p(r) = n_{C_r}\, p^r q^{n-r}, \quad r = 0, 1, 2, \ldots\ldots, n$

$\therefore P(X = r) = 6_{C_r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{6-r} = 6_{C_r} \times \frac{1}{64}$

(i) P(at least one success) = $P(X \geq 1)$

$= 1 - P(X < 1)$

$= 1 - P(X = 0)$

$= 1 - {}^{6}C_{0} \times \frac{1}{64}$

$= 1 - \frac{1}{64} = \frac{63}{64} = 0.9843$

(ii) P(exactly 2 successes) = P(X=2)

$= {}^{6}C_{2} \times \frac{1}{64} = \frac{15}{64} = 0.2343$

(iii) P(at least 4 success) = $P(X \geq 4)$

$= P(X = 4) + P(X = 5) + P(X = 6)$

$= {}^{6}C_{4} \times \frac{1}{64} + {}^{6}C_{5} \times \frac{1}{64} + {}^{6}C_{6} \times \frac{1}{64}$

$= \frac{15}{64} + \frac{6}{64} + \frac{1}{64}$

= 0.2343+0.09375+0.0156 = 0.3436

(iv) P(at most 3 successes) = $P(X \leq 3)$

$= 1 - P(X > 3)$

$= 1 - \{P(X = 4) + P(X = 5) + P(X = 6)\}$

$= 1 - \{ {}^{6}C_{4} \times \frac{1}{64} + {}^{6}C_{5} \times \frac{1}{64} + {}^{6}C_{6} \times \frac{1}{64} \}$

$= 1 - (\frac{15}{64} + \frac{6}{64} + \frac{1}{64})$

= 1−(0.2343+0.09375+0.0156) =1− 0.3436 = 0.6564

(v) P(more than 3 success) = $P(X > 3)$

$= P(X = 4) + P(X = 5) + P(X = 6)$

$= {}^{6}C_{4} \times \frac{1}{64} + {}^{6}C_{5} \times \frac{1}{64} + {}^{6}C_{6} \times \frac{1}{64}$

$= \frac{15}{64} + \frac{6}{64} + \frac{1}{64}$

= 0.2343+0.09375+0.0156 = 0.3436

(vi) P(less than 4 successes) = $P(X < 4)$

$= 1 - P(X \geq 4)$

$= 1 - \{P(X = 4) + P(X = 5) + P(X = 6)\}$

$= 1 - \{{}^{6}C_{4} \times \frac{1}{64} + {}^{6}C_{5} \times \frac{1}{64} + {}^{6}C_{6} \times \frac{1}{64}\}$

$= 1 - \{\frac{15}{64} + \frac{6}{64} + \frac{1}{64}\}$

= 1−(0.2343+0.09375+0.0156) =1− 0.3436 = 0.6564

(vii) P (between 2 and 6 success) = $P(2 < X < 6)$

$= (X = 3) + P(X = 4) + P(X = 5)$

$= {}^{6}C_{3} \times \frac{1}{64} + {}^{6}C_{4} \times \frac{1}{64} + {}^{6}C_{5} \times \frac{1}{64}$

$= \frac{20}{64} + \frac{15}{64} + \frac{6}{64}$

= 0.3125 + 0.2343+0.09375

= 0.6405

An experiment succeeds twice as often as it fails. Find the chance that in the next six trials there will at least be 4 successes.

Solution

Let p denotes the probability of success and q denotes its failure, then

By given condition

p = 2 q

Since, total probability is one

p + q = 1

⇒ 2 q + q = 1

or, 3 q = 1

or, q = $\frac{1}{3}$ & p =1−q = 1−$\frac{1}{3}$ =$\frac{2}{3}$

Number of trials (n) = 6

Let random variable X denotes the number of successes.

Then, the probability of getting r successes out of n trials in binomial distribution is given by

$P(X = r) = p(r) = n_{C_r}\, p^r q^{n-r}, \quad r = 0, 1, 2, \ldots, n.$

∴ The probability of at least 4 successes in 6 trials is given by

$P(X \geq 4) = P(X = 4) + P(X = 5) + P(X = 6)$

$= 6_{C_4} \left(\frac{2}{3}\right)^4 \left(\frac{1}{3}\right)^{6-4} + 6_{C_5} \left(\frac{2}{3}\right)^5 \left(\frac{1}{3}\right)^{6-5} + 6_{C_6} \left(\frac{2}{3}\right)^6 \left(\frac{1}{3}\right)^{6-6}$

= 0.3292 + 0.2633+0.0877

$= 0.68$

The mean and standard deviation of binomial distribution are 20 and 4 respectively. Calculate n, p and q. Also find P($3 < X < 6$).

Solution

If n and p are the parameters of the binomial distribution, then we are given:

Mean = np = 20. . . . . (i)

Standard deviation = $\sqrt{npq}$ = 4

Variance = npq =16 . . . . . (ii)

Dividing (ii) by (i), we get

$\frac{npq}{np} = \frac{16}{20} = \frac{4}{5}$

$\Rightarrow q = \frac{4}{5}$

$p = 1 - q = 1 - \frac{4}{5} = \frac{1}{5}$

Substituting in (i), we get

$np = 20$

$n \times \frac{1}{5} = 20 \Rightarrow n = 100$

$\therefore n = 100, p = \frac{1}{5}, q = \frac{4}{5}$

The probability of getting r successes out of n trials in binomial distribution is given by

$$P(X = r) = p(r) = {}_{n}C_{r}\, p^{r} q^{n-r}, \quad r = 0, 1, 2, \ldots\ldots, n$$

$$\therefore P(X = r) = {}_{6}C_{r} \left(\frac{1}{5}\right)^{r} \left(\frac{4}{5}\right)^{6-r}$$

Now,

$$P(3 < X < 6) = P(X = 4) + P(X = 5)$$

$$= {}_{100}C_{4} \left(\frac{1}{5}\right)^{4} \left(\frac{4}{5}\right)^{100-4} + {}_{100}C_{5} \left(\frac{1}{5}\right)^{5} \left(\frac{4}{5}\right)^{100-5}$$

$$= 0.00000321 + 0.00001496 = 0.00001497$$

[redacted]

A binomial random variable X satisfies the relation $6P(X = 3) = P(X = 2)$ when n = 5. Find the value of parameter 'p'.

Solution

If n and p are the parameters of the binomial distribution, then the probability of getting r successes out of n trials in binomial distribution is given by

$$P(X = r) = p(r) = {}_{n}C_{r}\, p^{r} q^{n-r}, \quad r = 0, 1, 2, \ldots\ldots, n \qquad q = 1 - p$$

$$P(X = 3) = {}_{5}C_{3}\, p^{3} q^{5-3} = 10p^{3}q^{2} \ldots\ldots \text{(i)}$$

$$\& \; P(X = 2) = {}_{5}C_{2}\, p^{2} q^{5-2} = 10p^{2}q^{3} \ldots\ldots\ldots \text{(ii)}$$

According to the question

$$6P(X = 3) = P(X = 2)$$

or, $6 \times 10p^{3}q^{2} = 10p^{2}q^{3}$

or, $6p = q$

or, $6p = 1 - p$

or, $7p = 1$

$\therefore p = \frac{1}{7}$

[redacted]

In a binomial distribution with 6 independent trials, the probability of 3 and 4 successes is found to be 0.2457 and 0.0819 respectively. Find the parameters p and q of the binomial distribution.

Solution:

If n and p are the parameters of the binomial distribution, let X denotes the number of successes. Then, the probability of getting r successes out of n trials in binomial distribution is given by

$P(X = r) = p(r) = n_{C_r}\, p^r q^{n-r}, \quad r = 0, 1, 2, \ldots\ldots, n \qquad q = 1-p$

By given condition

$P(X = 3) = p(3) = 6_{C_3}\, p^3 q^{6-3} = 20p^3q^2 = 0.2457 \ldots\ldots\ldots$(i)

& $P(X = 4) = p(4) = 6_{C_4}\, p^4 q^{6-4} = 15p^4q^2 = 0.0819 \ldots\ldots\ldots$ (ii)

Dividing (ii) by (i), we get

$\frac{p(4)}{p(3)} = \frac{15p^4q^2}{20p^3q^2} = \frac{0.0819}{0.2457}$

$\Rightarrow \frac{3}{4}\frac{p}{q} = \frac{1}{3}$

$\Rightarrow 9p = 4q \Rightarrow 9p = 4(1-p)$

$\Rightarrow 9p+4p = 4 \Rightarrow 13p = 4$

$\Rightarrow p = \frac{4}{13}$ & $q = 1-p = 1 - \frac{4}{13} = \frac{9}{13}$

$\therefore p = \frac{4}{13}$ & $q = \frac{9}{13}$

Fitting of Binomial distribution

expected or theoretical frequencies of getting r successes is given by

$f(r) = NP(X = r) \qquad r = 0, 1, 2, 3, \ldots\ldots, n$

$= Np(r) \qquad N = \sum f = \text{Total frequency}$

$\therefore f(r) = N \times n_{C_r}\, p^r q^{n-r} \qquad p + q = 1 \Rightarrow q = 1 - p$

Case I: If probability of success 'p' is given:

[redacted]

Fit the binomial distribution for the following data. Assuming that the coin is unbiased.

No. of heads	0	1	2	3	4
Observed frequency	56	114	92	40	8

Solution

We have n = 4, N = $\sum f = 310$

Since, the coin is unbiased

∴ Probability of getting head or success (p) = $\frac{1}{2}$

Probability of getting tail or failure (q) = $\frac{1}{2}$ i.e. p = q = $\frac{1}{2}$

Let, the random variable X denotes the number of heads (success) turned up, then the expected or theoretical frequency of getting r heads (successes) in binomial distribution is given by

$f(r) = N\,P(X = r)$

$= N \times n_{C_r}\, p^r q^{n-r}$; $r = 0, 1, 2, 3, 4$ & p + q = 1, q = 1- p

$= 310 \times 4_{C_r} \left(\frac{1}{2}\right)^r \left(\frac{1}{2}\right)^{4-r}$

$= 310 \times 4_{C_r} \times \frac{1}{16}$

∴ $f(r) = 310 \times 4_{C_r} \times \frac{1}{16}$

Calculation of expected frequencies is as follows:

Number of heads (r)	$f(r) = 310 \times 4_{C_r} \times \frac{1}{16}$
0	$f(0) = 310 \times 4_{C_0} \times \frac{1}{16} = 19.375 \approx 19$
1	$f(1) = 310 \times 4_{C_1} \times \frac{1}{16} = 77.5 \approx 78$
2	$f(2) = 310 \times 4_{C_2} \times \frac{1}{16} = 116.25 \approx 116$
3	$f(3) = 310 \times 4_{C_3} \times \frac{1}{16} = 77.5 \approx 78$
4	$f(4) = 310 \times 4_{C_4} \times \frac{1}{16} = 19.375 \approx 19$

Hence, the fitted binomial distribution is

No. of heads	0	1	2	3	4	Total
Expected frequency	19	78	116	78	19	310

Case II: If probability of success 'p' is not given:

In this case, at first we find the mean of the given frequency distribution by using the formula

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

Equating it to np, which gives the mean of the binomial distribution.

$np = \bar{x}$

$\Rightarrow p = \frac{\bar{x}}{n}$

then $q = 1 - p$ and with these estimated values of p and q, the expected frequencies can be obtained by the above relation.

[redacted]

Fit the binomial distribution to the following data.

X	0	1	2	3	4
f	4	10	46	62	28

Solution

We have n = 4, $N = \sum f = 150$

Since, the values of parameters p and q of the binomial distribution are unknown (i.e. the nature of the coin is unknown) therefore we use

$$np = \overline{X}$$

$$\Rightarrow \quad np = \frac{\sum fx}{N} = \frac{4\times0+10\times1+46\times2+62\times3+28\times4}{150} = \frac{400}{150}$$

$$\Rightarrow 4\,p = \frac{400}{150}$$

$$\text{or, } p = \frac{400}{150\times4} = \frac{2}{3}$$

$$\therefore p = \frac{2}{3}, \ \& \ q = 1-p = 1-\frac{2}{3} = \frac{1}{3}$$

Let, the random variable X denotes the number of successes, then the expected or theoretical frequency of getting successes in binomial distribution is given by

$$f(r) = N\,P(X = r)$$

$$= N \times n_{C_r}\, p^r q^{n-r}; \qquad r = 0, 1, 2, 3, 4 \quad \& \quad p + q = 1$$

$$\therefore f(r) = 150 \times 4_{C_r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r}$$

Calculation of expected frequencies as follows:

Number of heads (r)	$f(r) = 150 \times 4_{C_r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r}$
0	$f(0) = 150 \times 4_{C_r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r} = 19.375 \approx 19$
1	$f(1) = 150 \times 4_{C_r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r} = 77.5 \approx 78$
2	$f(2) = 150 \times 4_{C_r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r} = 116.25 \approx 116$
3	$f(3) = 150 \times 4_{C_r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r} = 77.5 \approx 78$
4	$f(4) = 150 \times 4_{C_r} \left(\frac{2}{3}\right)^r \left(\frac{1}{3}\right)^{4-r} = 19.375 \approx 19$

Hence, the fitted binomial distribution is

No. of heads	0	1	2	3	4	Total
Expected frequency	19	78	116	78	19	310

Poisson distribution

Poisson distribution is a discrete probability distribution developed by French mathematician, Simeon Denis Poisson (1781-1840) in the year 1837. It is also one of the most widely used discrete probability distribution among other probability distributions. Poison distribution explains the behaviour of those discrete random variables for which the probability of occurrence of an event is very small and the total number of possible cases is very large (i.e. number of trials is very large). Poisson distribution deals with the evaluation of probabilities of rare events.

Conditions of Poisson distribution

The Poisson distribution is the limiting case of Binomial distribution under the following conditions:

(i) The number of trials (n) is indefinitely large i.e. $n \to \infty$.

In practice, $n > 20$

(ii) The probability of success(p) for each trial is indefinitely small i.e. $p \to 0$

In practice , $p < 0.05$

(iii) $np = \lambda$ (Say) is finite so that $p = \frac{\lambda}{n}$ & $q = 1 - \frac{\lambda}{n}$. Where λ is read as lamda.

Definition: A random variable X is said to follow Poisson distribution if it assumes only non-negative integer values 0, 1, 2, 3, 4, ... , and its probability mass function (pmf) is given by

$$P(X = r) = p(r) = \begin{cases} \frac{e^{-\lambda}\lambda^r}{r!}, & r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0 \\ 0 & \text{otherwise} \end{cases},$$

Where, λ = Mean of the Poisson distribution = Average number of occurrences per specified interval and is called parameter of the Poisson distribution.

$r =$ Number of successes of a given event.

$e = 2.7183$ is the base of natural logarithm

Symbolically, if X follows Poisson distribution with parameter 'λ ' then it is written as $X \sim P(\lambda)$.

Note: λ can also be denoted by other notations such as 'm' or 'θ'

Uses (applications) of Poisson distribution

Poisson distribution is used in a wide variety of fields such as queuing theory, (waiting time problems), insurance, physics, biology, business, Economics, industry etc. Some practical important applications of the Poisson distribution are as follows:

1. Poisson distribution is mostly used in statistical quality control to find the number of defects per unit of manufactured product and hence to construct the control chart and in sampling inspection plan when the chance of any defective items is very small and lot is very large.

2. Poisson distribution is used in business, economics etc. where the probability of occurrence of an event is very small and the number of trials is very large.
3. It is used in queuing theory or waiting time problems to count the number of incoming customers or arrivals in a service station.
4. It is used in insurance problems to count the number casualties during a certain time interval.
5. It is used in Physics to determine the number of particles emitted from a radioactive substance.
6. It is also used in biology to count the number of bacteria per unit.
7. The number of accidents taking place per day on a busy road.
8. The number of printing mistakes per page in a book etc.

Characteristics (Properties) of Poisson distribution

1. Poisson distribution is a discrete probability distribution since the number of occurrences can take only integral values 0, 1, 2, 3, ... ,∞.
2. It is also known as uni-parametric discrete distribution as it is characterized by only one parameter λ.
3. The mean, variance and standard deviation of poison distribution are:
   Mean$(\overline{X})$ = E(X) = λ, Variance(σ^2) = V(X) = λ, Standard deviation $(\sigma) = \sqrt{\lambda}$. Hence, for Poisson distribution mean and variance are equal, each being equal to the parameter λ. i.e. Mean = Variance = λ
4. The Poisson distribution is positively skewed (as λ is always positive) and the curve approaches to symmetrical form when the value of λ increases.
5. Poisson distribution could be unimodal or bimodal depending upon the value of the parameter λ.
6. Poisson distribution follows additive property but does not follow subtraction property.
7. It is applied in the situations when the number of trials n is very large and probability of success of an event 'p' is very small.

[redacted]

Comment on the following:

For Poisson distribution, Mean = 8 and Variance = 7.

Solution:

Since, for a Poisson distribution mean and variance are equal, therefore the given statement is wrong.

[redacted]

Between the hours 2 P.M. and 4 P.M. the average number of phone calls per minute coming into the switch board of a company is 2.5. Find the probability that during one particular minute there will be (i) no phone call at all (ii) at least 3 calls (iii) at most 4 calls (iv) more than 4 calls (v) at least one call (v) less than 3 calls (vii) between 3 calls and 6 calls (viii) 2 or less calls .(ix) upto 3 calls.

Solution

Let the random variable X denotes the number of phone calls per minute coming into the switch board of a company such that X∼P(λ). Then, the probability of getting r calls per minute by Poisson distribution is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0$$

Here, average number of phone calls per minute $(\lambda) = 2.5$

$$\therefore P(X = r) = p(r) = \frac{e^{-2.5}(2.5)^r}{r!}$$

(i) P (no phone call at all) = $P(X = 0)$

$$= \frac{e^{-2.5}\times(2.5)^0}{0!}$$

$$= \frac{0.0821\times1}{1} = 0.0821$$

(ii) P (at least 3 calls) = $P(X \geq 3)$

$$= 1 - P(X < 3)$$

$$= 1 - P(X \leq 2)$$

$$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

$$= 1 - \{\frac{e^{-2.5}\times(2.5)^0}{0!} + \frac{e^{-2.5}\times(2.5)^1}{1!} + \frac{e^{-2.5}\times(2.5)^2}{2!}\}$$

$$= 1 - e^{-2.5}\{\frac{(2.5)^0}{0!} + \frac{(2.5)^1}{1!} + \frac{(2.5)^2}{2!}\}$$

$$= 1 - 0.0821\ (1 + 2.5 + 3.125)$$

$$= 1 - 0.05438 = 0.4562$$

(iii) P (at most 4 calls) = $P(X \leq 4)$

$$= \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)\}$$

$$= \{\frac{e^{-2.5}\times(2.5)^0}{0!} + \frac{e^{-2.5}\times(2.5)^1}{1!} + \frac{e^{-2.5}\times(2.5)^2}{2!} + \frac{e^{-2.5}\times(2.5)^3}{3!} + \frac{e^{-2.5}\times(2.5)^4}{4!}\}$$

$$= e^{-2.5}\{\frac{(2.5)^0}{0!} + \frac{(2.5)^1}{1!} + \frac{(2.5)^2}{2!} + \frac{(2.5)^3}{3!} + \frac{(2.5)^4}{4!}\}$$

$$= 0.0821\ (1 + 2.5 + 3.125 + 2.6042 + 1.6276)$$

$$= 0.0821 \times 10.8568 = 0.8912$$

(iv) p(more than 4 calls) = $P(X > 4)$

$$= 1 - P(X \leq 4)$$

$$= 1 - 0.8912 = 0.1088$$

(v) P (at least one call) = $P(X \geq 1)$

$$= 1 - P(X < 1)$$

$$= 1 - P(X = 0)$$

$$= 1 - 0.0821 = 0.9179$$

(vi) P(less than 3 calls) = $P(X < 3)$

$$= P(X \leq 2)$$

$$= \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

$$= \{\frac{e^{-2.5}\times(2.5)^0}{0!} + \frac{e^{-2.5}\times(2.5)^1}{1!} + \frac{e^{-2.5}\times(2.5)^2}{2!}\}$$

$$= e^{-2.5}\{\frac{(2.5)^0}{0!} + \frac{(2.5)^1}{1!} + \frac{(2.5)^2}{2!}\}$$

$$= 0.0821\ (1 + 2.5 + 3.125)$$

$$= 0.0821 \times 6.525 = 0.05438$$

(vii) p(between 3 calls and 6 calls) = $P(3 < X < 6)$

$$= P(X = 4) + P(X = 5)$$

$$= \frac{e^{-2.5}\times(2.5)^4}{4!} + \frac{e^{-2.5}\times(2.5)^5}{5!}$$

$$= e^{-2.5}\{\frac{(2.5)^4}{4!} + \frac{(2.5)^5}{5!}\}$$

$$= 0.0821(1.6276 + 0.81380)$$

$$= 0.0821 \times 2.44140 = 0.200438$$

(viii) P(2 or less calls) $= P(X \leq 2)$

$$= \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

$$= \{\frac{e^{-2.5}\times(2.5)^0}{0!} + \frac{e^{-2.5}\times(2.5)^1}{1!} + \frac{e^{-2.5}\times(2.5)^2}{2!}\}$$

$$= e^{-2.5}\{\frac{(2.5)^0}{0!} + \frac{(2.5)^1}{1!} + \frac{(2.5)^2}{2!}\}$$

$$= 0.0821\ (1 + 2.5 + 3.125)$$

$$= 0.0821 \times 6.525 = 0.05438$$

(ix) P (up to 3 calls) = $P(X \leq 3)$

$$= \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= \frac{e^{-2.5}\times(2.5)^0}{0!} + \frac{e^{-2.5}\times(2.5)^1}{1!} + \frac{e^{-2.5}\times(2.5)^2}{2!} + \frac{e^{-2.5}\times(2.5)^3}{3!}$$

$$= e^{-2.5}\{\frac{(2.5)^0}{0!} + \frac{(2.5)^1}{1!} + \frac{(2.5)^2}{2!} + \frac{(2.5)^3}{3!}\}$$

$$= 0.0821\ (1 + 2.5 + 3.125 + 2.604)$$

$$= 0.0821 \times 9.229 = 0.7577$$

[redacted]

It is known from past experience that in a certain plant there are on the average 4 industrial accidents per month. Find the probability that in a given year there will be less than 4 accidents. Assume Poisson distribution. ($e^{-4} = 0.0183$)

Solution:

Let the random variable X denotes the number of accidents in the plant per month such that $X \sim P(\lambda)$. Then, the probability of getting r accidents per month by Poisson distribution (i.e. Poisson probability law) is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0$$

Here, average number of accidents per month $(\lambda) = 4$

$$\therefore P(X = r) = p(r) = \frac{e^{-4}(4)^r}{r!}$$

(i) The probability that there will be less than 4 accidents = $P(X < 4)$

$$= \{P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)\}$$

$$= \{\frac{e^{-4}\times(4)^0}{0!} + \frac{e^{-4}\times(4)^1}{1!} + \frac{e^{-4}\times(4)^2}{2!} + \frac{e^{-4}\times(4)^3}{3!}\}$$

$$= e^{-4}\{\frac{(4)^0}{0!} + \frac{(4)^1}{1!} + \frac{(4)^2}{2!} + \frac{(4)^3}{3!}\}$$

$$= 0.0183\ (1 + 4 + 8 + 10.67)$$

$$= 0.0183 \times 23.67 = 0.4332$$

[redacted]

A car hire firm has two cars which it hires out day by day. The number of demands for a car on each day is distributed as a Poisson variate with mean 1.5. Calculate the proportion of days on which (i) neither car is used (ii) some demand is refused.

Solution

Let the random variable X denotes the number of demands for a car on any day such that $X \sim P(\lambda)$. Then, the probability of getting r demands for a car on any day by Poisson distribution (i.e. Poisson probability law) is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0$$

Here, average (mean) number of demands of car per day $(\lambda) = 1.5$

$$\therefore P(X = r) = p(r) = \frac{e^{-1.5}(1.5)^r}{r!}$$

(i) The Proportion of days on which neither car is used = P(neither car is used)

$$= P \text{ (no demand)}$$
$$= P(X = 0)$$
$$= \frac{e^{-1.5} \times (1.5)^0}{0!}$$
$$= \frac{0.02231 \times 1}{1}$$
$$= 0.02231$$

(ii) Since, the firm has only two cars, some demands will be refused if the number of demands per day is greater 2.Hence,

P (some demand is refused) = $P(X > 2)$

$$= 1 - P(X \leq 2)$$
$$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$$
$$= 1 - \{\frac{e^{-1.5} \times (1.5)^0}{0!} + \frac{e^{-1.5} \times (1.5)^1}{1!} + \frac{e^{-1.5} \times (1.5)^2}{2!}\}$$
$$= 1 - e^{-1.5}\{\frac{(1.5)^0}{0!} + \frac{(1.5)^1}{1!} + \frac{(1.5)^2}{2!}\}$$
$$= 1 - 0.2231\ (1 + 1.5 + 1.125)$$
$$= 1 - 0.2231 \times 3.625$$
$$= 1 - 0.80874 = 0.19126$$

Poisson approximation to Binomial distribution

The Poisson distribution is a good approximation of the binomial distribution when

(i) the number of trials ' n' is greater than or equal to 20 (i.e.$n \leq 20$)

(ii) and the probability of success 'p' for each trial is less than or equal to 0.05 ($p \leq 0.05$)

Then binomial distribution is converted into Poisson distribution by converting mean of binomial distribution = mean of Poison distribution.

Symbolically, $np = \lambda$

[redacted]

Suppose on an average 1 house in 1000 in a certain district has a fire during a year. If there are 2000 houses in that district, what is the probability that exactly 5 houses will have a fire during the year?

Solution: Here, probability of a house fire during a year (p) = $\frac{1}{1000}$

Number of houses (n) = 2000

Since, p is small and n is large, therefore Poisson distribution is a good approximation to binomial distribution.

Mean number of houses that will have a fire, $\lambda = np = 2000\times\frac{1}{1000} = 2$

Let the random variable X denotes the number of houses that will have a fire during the year such that $X\sim P(\lambda)$. Then, the probability of getting r houses that will have a fire by Poisson distribution (i.e. Poisson probability law) is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0,1,2,3,\ldots\ldots,\infty, \ \lambda > 0$$

P (exactly 5 houses will have a fire during the year) = P $(X = 5)$ =?

$$P(X = 5) = \frac{e^{-2}\,2^5}{5!} = 0.0360$$

[redacted]

If 5% of the electric bulbs manufactured by a company are defective. Use Poisson distribution to find the probability that in a sample of 100 bulbs

(i) none is defective
(ii) 5 bulbs will be defective
(iii) at most 2 will be defective
(iv) more than 4 will be defective
(v) at least one will defective
(vi) less than 3 will be defective
(vii) between 2 and 6 will be defective
(viii) 2 or less will be defective.

Solution:

Here, n = 100

Probability of a defective bulb (p) = 5% = 0.05

Since, p is small and n is large, therefore Poisson distribution is a good approximation to binomial distribution.

$\lambda = np$

$= 100\times 005 = 5$

The average number of defective bulbs $(\lambda) = 5$

Let the random variable X denotes the number of defective bulbs such that $X\sim P(\lambda)$. Then, the probability of getting r defective bulbs by Poisson distribution (i.e. Poisson probability law) is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0,1,2,3,\ldots\ldots,\infty, \ \lambda > 0$$

$$\therefore P(X = r) = p(r) = \frac{e^{-5}(5)^r}{r!}$$

(i) P (none is defective) = $P(X = 0)$

$$= \frac{e^{-5}\times(5)^0}{0!}$$

$$= \frac{0.0067\times 1}{1} = 0.0067$$

(ii) P (5 bulbs will be defective) = $P(X = 5)$

$$= \frac{e^{-5}\times(5)^5}{5!}$$

$$= \frac{0.0067\times(5)^5}{5!}$$

$$= 0.0067 \times 26.041$$

$$= 0.1744$$

(iii) P (at most 2 will be defective) = $P(X \leq 2)$

$$= \{P(X=0) + P(X=1) + P(X=2)\}$$

$$= \{\frac{e^{-5}\times(5)^0}{0!} + \frac{e^{-5}\times(5)^1}{1!} + \frac{e^{-5}\times(5)^2}{2!}\}$$

$$= e^{-5}\{\frac{(5)^0}{0!} + \frac{(5)^1}{1!} + \frac{(5)^2}{2!} + \frac{(5)^3}{3!} + \frac{(5)^4}{4!}\}$$

$$= 0.0067\ (1 + 5 + 12.5)$$

$$= 0.0067 \times 18.5 = 0.1239$$

(iv) p(more than 4 will be defective) = $P(X > 4)$

$$= 1 - P(X \leq 4)$$

$$= 1 - \{P(X=0) + P(X=1) + P(X=2) + P(X=3) + P(X=4)\}$$

$$= \{\frac{e^{-5}\times(5)^0}{0!} + \frac{e^{-5}\times(5)^1}{1!} + \frac{e^{-5}\times(5)^2}{2!} + \frac{e^{-5}\times(5)^3}{3!} + \frac{e^{-5}\times(5)^4}{4!}\}$$

$$= e^{-5}\{\frac{(5)^0}{0!} + \frac{(5)^1}{1!} + \frac{(5)^2}{2!} + \frac{(5)^3}{3!} + \frac{(5)^4}{4!}\}$$

$$= 0.0067\ (1 + 5 + 12.5 + 20.833 + 26.041)$$

$$= 0.0067 \times 52.874 = 0.3542$$

(v) P (at least one will be defective) = $P(X \geq 1)$

$$= 1 - P(X < 1)$$

$$= 1 - P(X = 0)$$

$$= 1 - 0.0067 = 0.9933$$

(vi) P(less than 3 will be defective) = $P(X < 3)$

$$= P(X \leq 2)$$

$$= \{P(X=0) + P(X=1) + P(X=2)\}$$

$$= \{\frac{e^{-5}\times(5)^0}{0!} + \frac{e^{-5}\times(5)^1}{1!} + \frac{e^{-5}\times(5)^2}{2!}\}$$

$$= e^{-5}\{\frac{(5)^0}{0!} + \frac{e^{-5}\times(5)^1}{1!} + \frac{e^{-5}\times(5)^2}{2!}\}$$

$$= 0.0067\ (1 + 2.5 + 3.125)$$

$$= 0.0067 \times 6.625 = 0.04438$$

(vii) p (between 2 and 6 will be defective) = $P(2 < X < 6)$

$$= P(X=3) + P(X=4) + P(X=5)$$

$$= \frac{e^{-5}\times(5)^3}{3!} + \frac{e^{-5}\times(5)^4}{4!} + \frac{e^{-5}\times(5)^5}{5!}$$

$$= e^{-2.5}\{\frac{(5)^3}{3!} + \frac{(5)^4}{4!} + \frac{(5)^5}{5!}\}$$

$$= 0.0067(20.8333 + 26.0416 + 26.0416)$$

$$= 0.0067 \times 72.9165 = 0.4885$$

(viii) P(2 or less will be defective) $= P(X \leq 2)$

$$= \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

$$= \{\frac{e^{-5} \times (5)^0}{0!} + \frac{e^{-5} \times (5)^1}{1!} + \frac{e^{-5} \times (5)^2}{2!}\}$$

$$= e^{-5}\{\frac{(5)^0}{0!} + \frac{(5)^1}{1!} + \frac{(5)^2}{2!}\}$$

$$= 0.0067\ (1 + 5 + 12.5)$$

$$= 0.0067 \times 18.5 = 0.1239$$

[redacted]

A manufacturer of pins knows that on an average 2% in production is defective. He sells pins in boxes of 100 and guarantee that not more than two pins will be defective. What is the probability that a box randomly selected (i) will meet the guaranteed quality (ii) will not meet the guaranteed quality.

$(e^{-1} = 0.3679, e^{-2} = 0.1353)$ (T.U. 2053 MBA)

Solution

Here, sample size (per box) (n) = 100

Probability of a defective pin (p) = 2% = 0.02

Since, p is small and n is large, therefore Poisson distribution is a good approximation to binomial distribution.

The average number of defective pins (λ) $= np$

$$= 100 \times 0.02 = 2$$

Let the random variable X denotes the number of defective pins in a box of 100 such that X~P(λ). Then, the probability of getting r defective pins in the selected box by Poisson distribution (i.e. Poisson probability law) is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0$$

$$\therefore P(X = r) = p(r) = \frac{e^{-2}(2)^r}{r!}$$

(i) P (the box will meet the guaranteed quality) = P (number of defective pins per box is not more than 2)

$$= P(X \leq 2)$$

$$= \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

$$= \{\frac{e^{-2} \times (2)^0}{0!} + \frac{e^{-2} \times (2)^1}{1!} + \frac{e^{-2} \times (2)^2}{2!}\}$$

$$= e^{-2}\{\frac{(2)^0}{0!} + \frac{(2)^1}{1!} + \frac{(2)^2}{2!}\}$$

$$= 0.1353\ (1 + 2 + 2)$$

$$= 0.1353 \times 5 = 0.6765$$

(ii) P (will not meet the guaranteed quality) = P (number of defective pins per box is more than 2)

$= P(X > 2)$
$= 1 - P(X \leq 2)$
$= 1 - 0.6765 = 0.3335$

[redacted]

If mean and variance of a Poisson distribution is 3, find

(i) $P(X \geq 3)$ (ii) $P(X < 3)$ (iii) $P(2 \leq X \leq 4)$

Solution

Let X be a random variable following Poisson distribution with mean and variance is 3. i.e. $X \sim P(\lambda)$.

$\therefore$ Mean = Variance = $\lambda = 3$

Then, the probability of getting r successes by Poisson distribution (i.e. Poisson probability law) is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0$$

(i) $P(X \geq 3) = P(X \geq 3)$

$= 1 - P(X < 3)$

$= 1 - P(X \leq 2)$

$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$

$= 1 - \{\frac{e^{-3}\times(3)^0}{0!} + \frac{e^{-3}\times(3)^1}{1!} + \frac{e^{-3}\times(3)^2}{2!}\}$

$= 1 - e^{-3}\{\frac{(3)^0}{0!} + \frac{(3)^1}{1!} + \frac{(3)^2}{2!}\}$

$= 1 - 0.04978\ (1 + 3 + 4.5)$

$= 1 - 0.04978 \times 8.5$

$= 1 - 0.4232 = 0.5768$

(ii) $P(X < 3) = 1 - P(X \geq 3)$

$= 1 - 0.5768 = 0.4232$

OR

$P(X < 3) = P(X \leq 2)$

$= P(X = 0) + P(X = 1) + P(X = 2)$

$= \frac{e^{-3}\times(3)^0}{0!} + \frac{e^{-3}\times(3)^1}{1!} + \frac{e^{-3}\times(3)^2}{2!}$

$= e^{-3}\{\frac{(3)^0}{0!} + \frac{(3)^1}{1!} + \frac{(3)^2}{2!}\}$

$= 0.04978\ (1 + 3 + 4.5)$

$= 0.04978 \times 8.5$

$= 0.4232$

(iii) $P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4)$

$$= \frac{e^{-3}\times(3)^2}{2!} + \frac{e^{-3}\times(3)^3}{3!} + \frac{e^{-3}\times(3)^4}{4!}$$

$$= e^{-3}\{\frac{(3)^2}{2!} + \frac{(3)^3}{3!} + \frac{(3)^4}{4!}\}$$

$$= 0.04978\ (4.5 + 4.5 + 3.375)$$

$$= 0.04978 \times 12.375$$

$$= 0.61602$$

[redacted]

If a random variable X follows Poisson distribution such that $P(X = 1) = P(X = 2)$, find

(i) the mean and variance of the poison distribution (ii) $P(X = 0)$ (iii) $P(X > 2)$

Solution

Let X be a random variable following Poisson distribution with parameter λ . i.e. $X \sim P(\lambda)$.

Then, the probability function of getting is given by

$$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \qquad r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0$$

(i) we have,

For r =1

$$P(X = 1) = \frac{e^{-\lambda}\times(\lambda)^1}{1!}$$

For r = 2

$$P(X = 2) = \frac{e^{-\lambda}\times(\lambda)^2}{2!}$$

According to question

$$P(X = 1) = P(X = 2)$$

$$\Rightarrow \frac{e^{-\lambda}\times(\lambda)^1}{1!} = \frac{e^{-\lambda}\times(\lambda)^2}{2!}$$

$$\Rightarrow 1 = \frac{\lambda}{2}$$

$$\Rightarrow \lambda = 2$$

∴ Mean = Variance = λ = 2

(ii) $P(X = 0) = \frac{e^{-2}\times(2)^0}{0!} = \frac{0.1353\times1}{1} = 0.1353$

(iii) $P(X > 2) = 1 - P(X \leq 2)$

$$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$$

$$= 1 - \{\frac{e^{-2}\times(2)^0}{0!} + \frac{e^{-2}\times(2)^1}{1!} + \frac{e^{-2}\times(2)^2}{2!}\}$$

$$= 1 - e^{-2}\{\frac{(2)^0}{0!} + \frac{(2)^1}{1!} + \frac{(2)^2}{2!}\}$$
$$= 1 - 0.1353\,(1 + 2 + 2)$$
$$= 1 - 0.1353 \times 5$$
$$= 1 - 0.6765 = 0.3235$$

Fitting of Poisson distribution

The process of finding the expected or theoretical frequencies on the basis of available information supplied by observed or experimental frequencies for a poison distribution is called fitting of Poisson distribution.

Suppose, a random experiment consisting of n trials is repeated N times and satisfying the conditions of Poisson distribution then expected or theoretical frequencies of getting 'r' successes is given by

$f(r) = NP(X = r)$

$= Np(r)$

$\therefore f(r) = N \times \frac{e^{-\lambda}\lambda^r}{r!}$; …….(*)

Where $r = 0, 1, 2, 3, \ldots\ldots, \infty$

$r =$ Number of occurrence of given event.

λ = mean or average number of occurrences per specified interval.

e = 2.7183 is the base of natural logarithm.

$N = \sum f =$ Total observed frequency

The expected or theoretical frequencies of Poisson distribution for different values of

$r = 0, 1, 2, 3, \ldots\ldots$ can be calculated as follows:

No. of success $(X = r)$	$f(r) = NP(X = r) = N \times \frac{\lambda^r\, e^{-\lambda}}{r!}$
0	$f(0) = N \times \frac{\lambda^0\, e^{-\lambda}}{0!}$
1	$f(1) = N \times \frac{\lambda^1\, e^{-\lambda}}{1!}$
2	$f(2) = N \times \frac{\lambda^2\, e^{-\lambda}}{2!}$.
3	$f(3) = N \times \frac{\lambda^3\, e^{-\lambda}}{3!}$
.	.
.	.
.	.

Case I: If average (mean) number of occurrence 'λ' is given (known):

The expected or theoretical frequencies can be obtained by using above relation (*) as shown in the above table.

Fit a Poisson distribution to the following data with mean $\lambda = 3$

Number of arrivals per hour	0	1	2	3	4	5 or more
Number of hours	20	57	98	85	78	62

Solution

We have,

Mean (average) number of arrivals per hour $(\lambda) = 3$

$N = \sum f = 20 + 57 + 98 + 85 + 78 + 62 = 400$

Let the random variable X denotes the number of arrival per hour following Poisson distribution with mean $\lambda = 3$, then the expected or theoretical frequency of getting r arrivals (successes) is given by

$f(r) = NP(X = r)$

$= Np(r)$

$= N \times \frac{e^{-\lambda}\lambda^r}{r!}$; Where $r = 0, 1, 2, 3, \ldots\ldots, \infty$

$= 400 \times \frac{e^{-3}(3)^r}{r!}$

$= 400 \times \frac{0.049787 \times (3)^r}{r!}$

$\therefore f(r) = \frac{19.9148 \times (3)^r}{r!}$

Calculation of expected frequencies as follows:

Number of arrivals per hour (r)	$f(r) = \frac{19.9148 \times (3)^r}{r!}$
0	$f(0) = \frac{19.91248 \times (3)^0}{0!} = 19.9148 \approx 20$
1	$f(1) = \frac{19.91248 \times (3)^1}{1!} = 59.7444 \approx 60$
2	$f(2) = \frac{19.91248 \times (3)^2}{2!} = 89.6166 \approx 90$
3	$f(3) = \frac{19.91248 \times (3)^3}{3!} = 89.60616 \approx 90$
4	$f(4) = \frac{19.91248 \times (3)^4}{4!} = 67.2046 \approx 67$
5 or more	$f(5 \text{ or more}) = 73$

$f(5 \text{ or more}) = 400 - 20 - 60 - 90 - 90 - 67 = 73$

Hence, the fitted Poisson distribution is

Number of arrivals per hour (r)	0	1	2	3	4	5 or more	Total
Expected frequency	20	60	90	90	67	73	310

Case II: If average (mean) number of occurrence 'λ' is not given (known):

In this case, at first we find the mean of the given frequency distribution by using the formula

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

Equating it to, which gives the mean of the Poisson distribution.

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

and with the estimated value of λ the expected frequencies can be obtained by the above relation (*).

Note: The value of $e^{-\lambda}$ can calculated by using scientific calculator or obtained from tables.

[redacted]

Fit a Poisson distribution to the following data. Also compute the mean and variance of fitted (theoretical) distribution.

Number of mistake per page (X)	0	1	2	3	4	5
Number of pages (f)	40	30	20	15	10	5

Solution

Let the random variable X denotes the number of mistake per page such that $X \sim P(\lambda)$ i.e. follows Poisson distribution, then the expected or theoretical frequency of getting r mistake per page is given by

$$f(r) = NP(X = r)$$

$$= Np(r)$$

$$= N \times \frac{e^{-\lambda}\lambda^r}{r!} \ldots\ldots\ldots \text{(i)} \qquad \text{Where} \quad r = 0, 1, 2, 3, \ldots\ldots, \infty$$

Here, $N = \sum f = 40 + 30 + 20 + 15 + 10 + 5 = 120$

If the above distribution is approximated by a Poisson distribution, then

$$\lambda = \bar{x} = \frac{\sum fx}{N} = \frac{40\times0+30\times1+20\times2+15\times3+10\times4+5\times5}{120} = \frac{180}{120} = 1.5$$

$$\Rightarrow \lambda = 1.5$$

From (i)

$$f(r) = N \times \frac{e^{-\lambda}\lambda^r}{r!}$$

$$= 120\times \frac{e^{-1.5}(1.5)^r}{r!} = 120\times \frac{0.22313\times(1.5)^r}{r!} = \frac{26.77562\times(1.5)^r}{r!}$$

$$\therefore f(r) = \frac{26.77562\times(1.5)^r}{r!}$$

Calculation of expected frequencies as follows:

Number of mistake per page (r)	$f(r) = \frac{26.77562\times(1.5)^r}{r!}$
0	$f(0) = \frac{26.77562\times(1.5)^0}{0!} = 26.7756 \approx 27$

1	$f(1) = \frac{26.77562 \times (1.5)^1}{1!} = 40.1634 \approx 40$
2	$f(2) = \frac{26.77562 \times (1.5)^2}{2!} = 30.1225 \approx 30$
3	$f(3) = \frac{26.77562 \times (1.5)^3}{3!} = 15.0612 \approx 15$
4	$f(4) = \frac{26.77562 \times (1.5)^4}{4!} = 5.6479 \approx 6$
5	$f(5) = \frac{26.77562 \times (1.5)^4}{5!} = 1.6943 \approx 2$

Hence, the fitted Poisson distribution is

Number of mistake per hour (r)	0	1	2	3	4	5	Total
Expected frequency	27	40	30	15	6	2	120

Since, for Poisson distribution, mean and variance are equal, the mean and variance of theoretical (fitted) distribution are given by:

Mean = Variance = $\lambda = 1.5$

[redacted]

In a certain factory turning out optical lenses, there is a small chance $\frac{1}{500}$ for any one lens to be defective. The lenses are supplied in packets of 10. Use Poisson distribution to calculate the approximate number of packets containing (i) no defective (ii) one defective (iii) two defective (iv) three defective (v) between 1 and 3 defective lenses in a consignment of 20,000 packets.(Given that $e^{-0.02} = 0.9802$)

Solution

Here, we are given

N = 20,000, n = 100

Probability of a defective optical lens (p) = $\frac{1}{500}$

Since, p is small and n is large, therefore Poisson distribution is a good approximation to binomial distribution.

$\lambda = np$

$= 10 \times \frac{1}{500} = 0.02$

∴The average number of defective optical lens (λ) $= 0.02$

Let the random variable X denotes the number of defective optical lenses in a packet of 10 such that $X \sim P(\lambda)$. Then, the probability of getting r defective optical lenses in a pocket by Poisson distribution (i.e. Poisson probability law) is given by

$P(X = r) = p(r) = \frac{e^{-\lambda}\lambda^r}{r!}, \quad r = 0, 1, 2, 3, \ldots\ldots, \infty, \ \lambda > 0$

$\therefore P(X = r) = p(r) = \frac{e^{-0.02}(0.02)^r}{r!} = \frac{0.9802 \times (0.02)^r}{r!}$

The number of packets containing r defective optical lenses in a consignment of 20,000 packets is $f(r) = NP(X = r) = 20{,}000 \times \frac{0.9802 \times (0.02)^r}{r!}$

(i) p (no defective) = $P(X = 0)$

$$= \frac{e^{-0.02} \times (0.02)^0}{0!}$$

$$= \frac{0.9802 \times 1}{1} = 0.9802$$

The expected number of packets containing no defective optical lenses = 20000× 0.9802 =19604

(ii) P (one defective) = $P(X = 1)$

$$= \frac{e^{-0.02} \times (0.02)^1}{1!}$$

$$= \frac{0.9802 \times (0.02)^1}{1!}$$

$$= 0.019604$$

The expected number of packets containing one defective optical lenses = 20000× 0.019604 =392.08≈ 392

(iii) P (two defective) = $P(X = 2)$

$$= P(X = 2)\}$$

$$= \frac{e^{-0.02} \times (0.02)^2}{2!}$$

$$= \frac{0.9802 \times (0.02)^2}{2!}$$

$$= 0.00019604$$

The expected number of packets containing two defective optical lenses = 20000× 0.00019604 =3.9208≈ 4

(iv) P (three defective) = $P(X = 3)\}$

$$= \frac{e^{-0.02} \times (0.02)^3}{3!}$$

$$= \frac{0.9802 \times (0.02)^3}{3!}$$

$$= 0.00000130693$$

The expected number of packets containing three defective optical lenses = 20000× 0.00019604 =0.026138≈ 0

(v) P(between 1 and 3 defective) = $P(1 < X < 3)$

$$= P(X = 2)$$

$$= \frac{e^{-0.02} \times (0.02)^2}{2!}$$

$$= \frac{0.9802 \times (0.02)^2}{2!}$$

$$= 0.00019604$$

The expected number of packets containing between 1 and 3 defective optical lenses = 20000× 0.00019604 =3.9208≈ 4

Normal Distribution (Gaussian distribution)

Normal probability distribution is one of the most important and widely used theoretical continuous probability distributions in statistics. Because of its characteristics, most of the data relating to economics, business or even in social and physical sciences conform to this distribution. Therefore, this distribution has made a notable revolution in the entire field of probability theory and Statistics. The normal distribution was first discovered by English Mathematician De-Moivre (1667–1754) in 1733 obtained the mathematical equation for this distribution while dealing with problems arising in the game of chance. It is also called Gaussian distribution (Gaussian law of errors) after Karl Friedrich Gauss (1777–1855.

Importance of Normal Distribution

Normal distribution plays vital role in Statistics theory. So, it is of great importance in Statistics because of the following reasons:

1. Most of the discrete probability distributions such as binomial, Poisson , hyper-geometric distributions etc. tend to normal distribution as n increases sufficiently large i.e. $n \to \infty$.
2. Almost all sampling distributions like, t, F, χ^2 etc. for their large degree of freedom conform (tend to) to normal distribution.
3. Normal distribution is the most frequently used probability distribution and it has wide applications in Statistics, physical, biological, social, medical sciences, engineering, economics, industry etc.
4. Even if many variables many variables which are not normally distributed can be normalized through suitable transformation.
5. The theory of exact sample test like t, F, χ^2 tests etc. is based on the fundamental assumption that the parent population (from which the samples are drawn) follows the normal distribution.
6. Normal distribution is largely (widely) applied in statistical quality control (SQC) in industry for finding control limits.
7. Normal distribution is extensively used to find confidence limits (interval) of the population parameters from sample statistics.
8. One of the greatest applications of normal distribution is that the central limit theorem follows the normal distribution when number of trials tends to infinity.
9. In theoretical statistics as well as applied field many problems can be solved only under the assumption of a normal population like age, weight , height ,output, income , wage etc.

Hence, to sum up, normal distribution is considered as ocean in which entire facts can be extracted.

Standard Normal Distribution

If X is a random variable following normal distribution with mean μ and standard deviationσ, then normal variable Z defined as follows:

$Z = \frac{X-E(X)}{\sigma} = \frac{X-\mu}{\sigma}$ is called standard normal variate (S.N.V.)

How to compute areas under Normal probability curve

Mathematically, the area bounded by the curve $f(x)$, X-axis and the ordinates at X = a and X = b is given by the definite integral:

$P(a < X < b) = \int_a^b f(x)\,dx$

Since, the area under normal probability curve has been tabulated in terms of the standard normal variate Z. Therefore, for practical problems in order to find the area $P(a < X < b)$under normal probability curve we do not deal with the variable X but first convert the normal variable X into standard normal variate Z by using

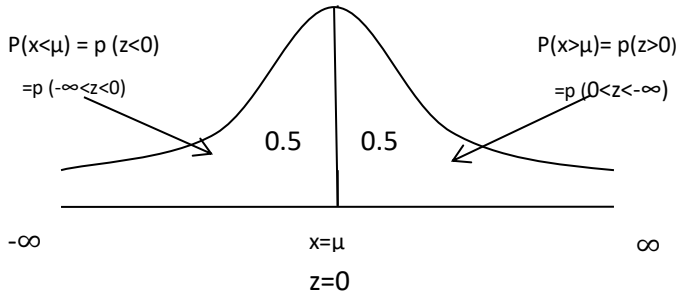
$$Z = \frac{X-\mu}{\sigma}$$

Next try to convert the required area in the form $P(o < Z < Z_1)$ and it is obtained from the area under standard normal curve using the following results:

$$P(X > \mu) = P(Z > 0) = 0.5$$

$$P(X < \mu) = P(Z < 0) = 0.5$$

and making use of the symmetry property of the distribution.



Note: In normal distribution, Mean = Median = Mode.

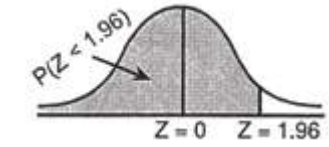
[redacted]

Given a standardized normal distribution, determine the following probabilities:
(i) P (Z < 1.96) (ii) P (Z > –0.34) (iii) P (Z > 1.64) (iv) P (Z < – 1.64) (v) P (0 < Z < 2.34)
(vi) P (–1.25 < Z < 0) (vii) P (0.17 < Z < 1.64) (viii) P $(-2 < Z < -1.76)$ (ix) P$(-2 < Z < 3.45)$
(x) Z is less than –0.84 or greater than +2.08 and (xi) what is the value of Z if only 31.87 % of all possible Z values are smaller?

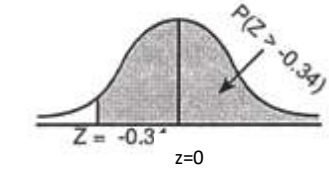
Solution:

Here, Z is a standard normal variate (S.N.V.)

(i) P (Z < 1.96) = P (–∞ < Z < 0) + P (0 < Z < 1.96)
= 0.50 + 0.475
= 0.975

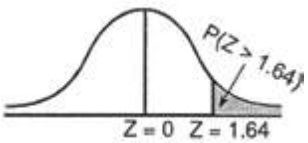


(ii) P (Z > –0.34) = P (–0.34 < Z < 0) + P(0 < Z <∞)
= 0.1331 + 0.50
= 0.6331

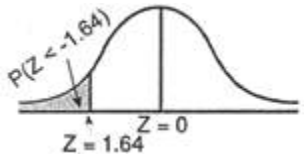


(iii) P (Z > 1.64) = P(1.64 < Z < ∞)

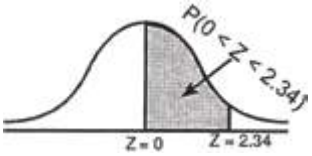
= P (0 < Z < ∞) – P (0 < Z < 1.64)
= 0.50 – 0.4495
= 0.0505



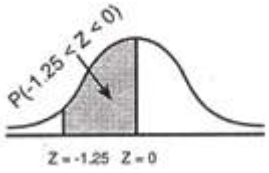
(iv) P (Z < – 1.64) = P (–∞ < Z < 0) – P (–1.64 < Z < 0)
= 0.5 – P (0 < Z < 1.64) [By symmetry]
= 0.50 – 0.4495
= 0.0505



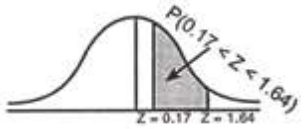
(v) P(0 < Z < 2.34) = 0.4904



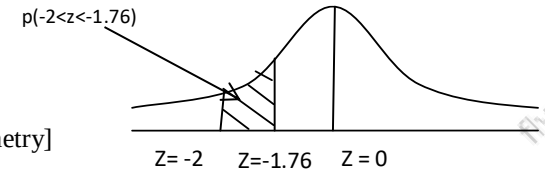
(vi) P(–1.25 < Z < 0) = P(0 < Z < 1.25) [By symmetry]
= 0.3944



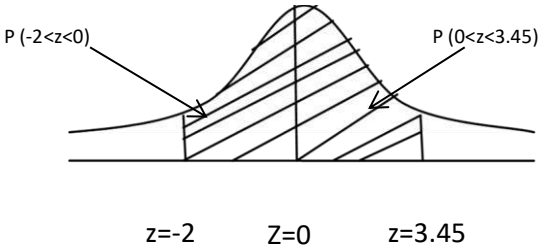
(vii) P (0.17 < Z < 1.64) = P(0 < Z < 1.64) – P(0 < Z < 0.17)
= 0.4495 – 0.0675
= .382



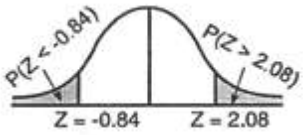
(viii) $P(-2 < Z < -1.76)$
P (-2 < Z < 0) – P (-1.76 < Z < 0)
P (0 < Z < 2) – P (0 < Z < 1.76) [By symmetry]
= 0.4772– 0.4608
= 0.0164



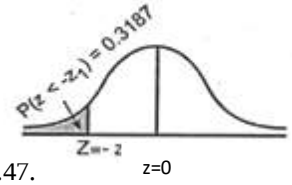
(ix) $P(-2 < Z < 3.45)$
= $P(-2 < Z < 0) +$ P (0 < Z < .3.45)
= P (0 < Z < 2) + P (0 < Z < 3.45) [By symmetry]
= 0.4772 + 0.4997
= 0.976



(x) The probability of Z is less than –0.84 or greater than +
= P (Z< – 0.84 or Z > 2.08)
= P (Z < –0.84) + P (Z > 2.08)
= [P (–∞ < Z < 0) – P (–0.84 < Z < 0)] + [P (0 < Z < ∞) – P (0 < Z < 2.08)]
= [0.50 – P (0 < Z < 0.84)] + [0.50 – 0.4812] [By symmetry]
= 0.50 – 0.2995 + 0.50 – 0.4812
= 0.2193



(xi) Left the value of z be $-z_1$ such that only 31.87% of all values of z are smaller than $-z_1$ i.e.,
P ($Z < -z_1$) = 0.3187
or, P(–∞ < Z < 0) – P($-z_1$< Z < 0) = 0.3187
or, 0.50 – P (0 < Z <z_1) = 0.3187
or, P (0 < Z < z_1) = 0.50 – 0.3187
or, P (0 < Z < $-z_1$) = 0.1813
The value of probability closer to 0.1813 in Z -table is 0.1808 at Z = 0.47.
z_1 = 0.47
Hence , the required value of Z is –0.47.



[redacted]

Given a normal distribution with mean μ = 200 and σ = 20, find the probability that
(i) P(X > 180) (ii) P(X < 220) (iii) P (160 < X < 240) (iv) P(X > 220) (v) P(X < 180 or X > 220) (vi) 10 % of the Values are less than what X values?

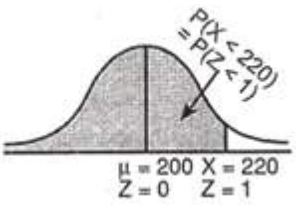
Solution:
Here, let X be the normal variable and Z be the corresponding standard normal variate (S.N.V.).
Then,

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

Here, mean (μ) = 200 and standard deviation (σ) = 20.

(i) For P(X > 180) =?

When X = 180
$Z = \frac{X-\mu}{\sigma} = \frac{180-200}{20} = -1$
Now,



$P(X > 180) = P(Z > -1)$
$= P(-1 < Z < 0) + P(0 < Z < \infty)$
$= P(0 < Z < 1) + 0.50$ by symmetry
$= 0.3413 + 0.50 = 0.8413$

(ii) $P(X < 220) = ?$

For $P(X < 220)$:

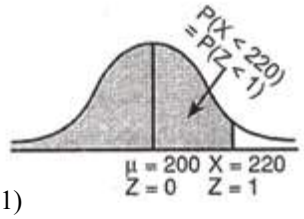
When $X = 220$

$Z = \frac{X-\mu}{\sigma} = \frac{220-200}{20} = 1$

$\therefore P(X < 220) = P(Z < 1)$

$= P(-\infty < Z < 0) + P(0 < Z < 1)$

$= 0.50 + 0.3413 = 0.8413$



(iii) $P(160 < X < 240) = ?$

When $X = 160$ $Z = \frac{X-\mu}{\sigma} = \frac{160-200}{20} = -2$

When $X = 240$ $Z = \frac{X-\mu}{\sigma} = \frac{240-200}{20} = 2$

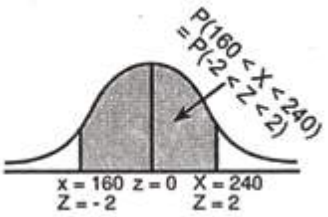
$\therefore P(160 < X < 240) = P(-2 < Z < 2)$

$= P(-2 < Z < 0) + P(0 < Z < 2)$

$= P(0 < Z < 2) + P(0 < Z < 2)$ by symmetry

$= 2 \times 0.4772$

$= 0.9544$



(iv) $P(X > 220) = ?$

When $X = 220$

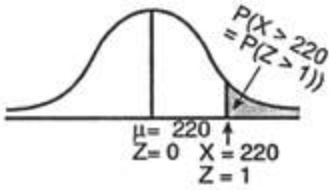
$Z = \frac{X-\mu}{\sigma} = \frac{220-200}{20} = 1$

$\therefore P(X > 220) = P(Z > 1)$

$= P(0 < Z < \infty) - P(0 < Z < 1)$

$= 0.50 - 0.3413$

$= 0.1587$



(v) $P(X < 180 \text{ or } X > 220) = ?$

When $X = 180$ $Z = \frac{180-200}{20} = -1$

When $X = 220$ $Z = \frac{220-200}{20} = 1$

$\therefore P(X < 180 \text{ or } X > 220)$

$= P(X < 180) + P(X > 220)$

$= P(Z < -1) + P(Z > 1)$

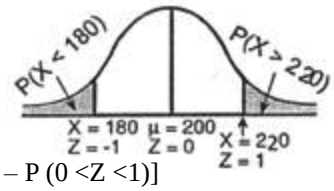
$= [P(-\infty < Z < 0) - P(-1 < Z < 0)] + [P(0 < Z < \infty) - P(0 < Z < 1)]$

$= [0.50 - P(0 < Z < 1)] - [0.50 - P(0 < Z < 1)]$ by symmetry

$= 1 - 2 \times P(0 < Z < 1)$

$= 1 - 2 \times 0.3413$

$= 0.3174$



(vi) Let x_1 be the value of X for which 10% of the values are less than x_1, i.e.

$P(X < x_1) = 0.10$

When $X = x_1$,

$Z = \frac{X-\mu}{\sigma} = \frac{x_1-200}{20} = z_1 \ldots\ldots\ldots$ (i)

$\therefore P(X < x_1) = 0.10$

$= P(Z < -z_1) = 0.10$

$= P(-\infty < Z < 0) - P(-z_1 < Z < 0) = 0.10$

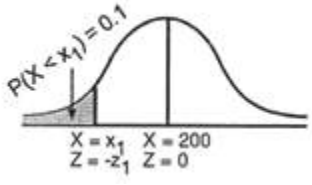
$= 0.50 - P(0 < Z < z_1) = 0.10$ by symmetry

$= P(0 < Z < z_1) = 0.40$

The value of probability closer to 0.40 in Z -table is 0.3997 at Z = 1.28

$\therefore z_1 = -1.28$ (from table) $\because$ it lies below mean

From equation (i)



$\frac{x_1-200}{20} = z_1$

or, $\frac{x_1-200}{20} = -1.28$

or, $x_1 = 200 - 1.28 \times 20$

or, $x_1 = 174.4$

Hence, the required value of X is 174.4

[redacted]

Kathmandu municipality puts 10,000 light bulbs on the streets of a city. If lives of bulbs follow a normal distribution with a mean of 60 days and a standard deviation of 20 days, how many bulbs will have to be replaced after (i) 40 days? and (ii) 80 days?

Solution:

Let the random variable X denotes the life of the electric bulbs such that it follows normal distribution with mean life of bulbs (μ)= 60 days

Standard deviation of life of bulbs (σ) = 20 days

Here, total number of light bulbs (N) = 10000

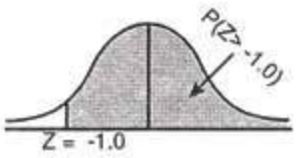
Then, the corresponding standard normal variate (S.N.V.) is

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

(i) The probability that the bulbs will have to be replaced after 40 days is given by P(X > 40)

When X = 40

$Z = \frac{X-\mu}{\sigma} = \frac{40-60}{20} = -1$



$\therefore P(X > 40) = P(Z > -1)$

$= P(-1 < Z < 0) + P(0 < Z < \infty)$

$= P(0 < Z < 1) + 0.5$ by symmetry

$= 0.3413 + 0.5$

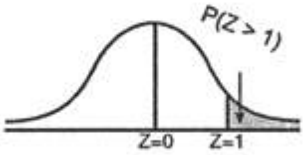
$= 0.8413$

Hence, the expected number of bulbs that will have to be replaced after 40 days = N × P (X > 40) = 10000 × 0.8413 = 8413

ii. The probability that the bulbs will have to be replaced after 80 days is given by P(X > 80).

For, X = 80

$Z = \frac{X-\mu}{\sigma} = \frac{80-60}{20} = 1$



$\therefore P(X > 80) = P(Z > 1)$

$= P(0 < Z < \infty) - P(0 < Z < 1)$

$= 0.5 - 0.3413$

$= 0.1587$

Hence, the expected number of bulbs that will have to be replaced after 80 days = N × P(X > 80) = 10000 × 0.1587 = 1587

[redacted]

The mean and a standard deviation of the wages of 6,000 workers engaged in a factory are Rs.1,200 and Rs.400 respectively. Assuming that the distribution to be normal. Estimate:

(i) the percentage of workers getting wages above Rs.1600 (ii) number of workers getting wages between Rs.400 and Rs.800 (iii) number of workers getting wages between Rs.1000 and Rs.1400.

Solution:

Let the random variable X denotes the wage of the workers engaged in a factory that follows normal distribution with mean wages(μ) = Rs.1200

Standard deviation of workers (σ) = Rs.400

Here, total number of workers (N) = 6000

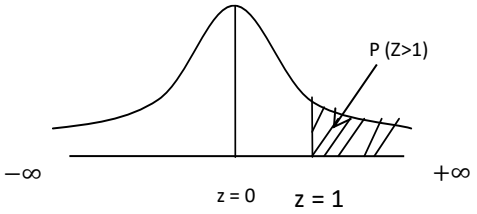
Then, the corresponding standard normal variate (S.N.V.) is

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

(i) The probability that workers getting wages above Rs.1600 is given by P(X > 1600).

For X = 1600

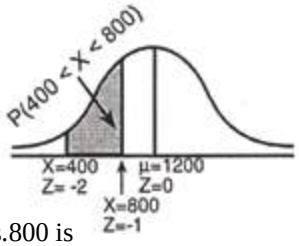
$Z = \frac{X-\mu}{\sigma} = \frac{1600-1200}{400} = 1$



∴ P(X > 1600) = P (Z > 1)
= P (0 < Z < ∞) – P (0 < Z < 1)
= 0.5 – 0.3413
= 0.1587
= 15.87%
Hence, the percentage of workers getting wages above Rs.1600 is 15.87%.

(ii) The probability of workers getting wages between Rs.400 to Rs.800 is given by P(400 < X < 800)
For, X = 400 For, X = 800

$Z = \frac{X-\mu}{\sigma} = \frac{400-1200}{400} = -2$ $Z = \frac{800-1200}{400} = -1$

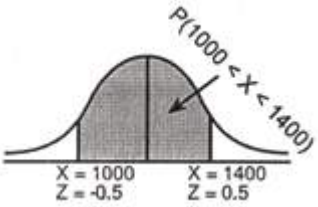


∴ P (400 < X < 800) = P (–2 < Z < –1)
= P (–2 < Z < 0) – P (–1 < Z < 0)
= P (0 < Z < 2) – P (0 < Z < 1) by symmetry
= 0.4772 – 0.3413
= 0.1359
Hence, the expected number of worker getting wages between Rs.400 and Rs.800 is
= N × P (400 < X < 800) = 6000 × 0.1359 = 815.4 = 815

(iii) The probability of workers getting wages between Rs.1000 to Rs.1400 is P(1000 < X < 1400)

For, X = 1000 for X = 1400

$Z = \frac{X-\mu}{\sigma} = \frac{1000-1200}{400} = -0.5$ $Z = \frac{1400-1200}{400} = -0.5$



∴ P (1000 < X < 1400) = P (–0.5 < Z < 0.5)
= P (–0.5 < Z < 0) + P (0 < Z < 0.5)
= P (0 < Z < 0.5) + P (0 < Z < 0.5) by symmetry
= 2 × P (0 < Z < 0.5)
= 2 × 0.1915
= 0.3830
Hence, the expected number of worker getting wages between Rs.1000 and Rs.1400 is
= N × P (1000 < X < 1400) = 6000 × 0.3830 = 2298

[redacted]

An industrial sewing machine uses ball bearings that are targeted to have a diameter of 0.75 inch. The lower and upper specification limits under which the 4 ball bearings can operate are 0.74 and 0.76 inch, respectively. Experience has indicated that the actual diameter of the ball bearings is approximately normally distributed with a mean of 0.753 inch and a standard deviation of 0.004 inch. What is the probability that a ball bearing will be (i) between the target and the actual mean, (ii) between the lower specification limit and the actual mean, (iii) above the upper specification limit, (iv) below the lower specification limit, (v) above which value in diameter will 93% of the ball bearing be?

Solution:
Let X be the normal variable which denotes the diameters of the ball bearings that follows normal distribution with
Mean diameter of ball the bearings (μ) = 0.753 inch.
Standard deviation of the ball bearings (σ) = 0.004inch.
Here, targeted value = 0.75
Lower specification limit = 0.74
Upper specification limit = 0.76
Then, the corresponding standard normal variate (S.N.V.) is

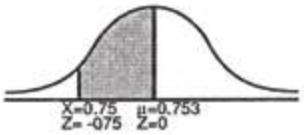
$Z = \frac{X-\mu}{\sigma}$

(i) The probability that a ball bearing will be between the targeted value and the actual mean is P(0.75 < X < 0.753).

For X = 0.75 For, X = 0.753

$Z = \frac{X-\mu}{\sigma} = \frac{0.75-0.753}{0.004} = 0.75$ $Z = \frac{0.753-0.753}{0.004} = 0$

∴ P (0.75 < X < 0.753) = P (–0.75 < Z < 0)
= P (0 < Z < 0.75) by symmetry
= 0.2734

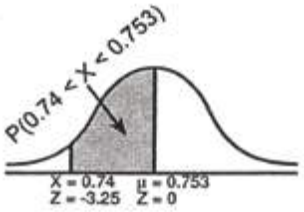


(ii) The probability that a ball bearing will be between the lower specification limit and the actual mean is P(0.74 < X < 0.753)

For, X = 0.74 For, X = 0.753

$z = \frac{0.74-0.753}{0.004} = -3.25$ $z = \frac{0.753-0.753}{0.004} = 0$

∴ P (0.74 < X < 0.753) = P (–3.25 < Z < 0)
= P (0 < Z < 3.25) by symmetry
= 0.4994

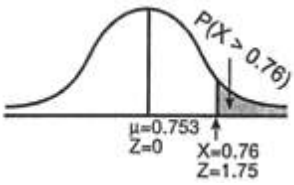


(iii) The probability that a ball bearing will be above the upper specification limit is P(X > 0.76).

For, X = 0.76

$z = \frac{0.76-0.753}{0.004} = 1.75$

∴ P(X < 0.76) = P (Z > 1.75)
= P (0 < Z < ∞) – P (0 < Z < 1.75)
= 0.5 – 0.4599
= 0.0401

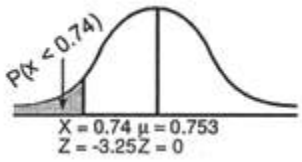


(iv) The probability that a ball bearing will be below the lower specification limit is P(X > 0.74).

For X = 0.74

$z = \frac{0.74-0.753}{0.004} = 3.25$

∴ P(X < 0.74) = P (Z < –3.25)
= P (–∞ < Z < 0) – P (–3.25 < Z < 0)
= 0.5 – P (0 < Z < 3.25) by symmetry
= 0.5 – 0.4994
= 0.0006

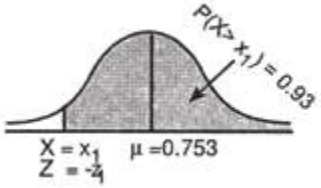


(v) Let x_1 be the value above which there are 93% of the ball bearings i.e. $P(X > x_1) = 0.93$

For $X = x_1$

$Z = \frac{X-\mu}{\sigma} = \frac{x_1-0.753}{0.004} = z_1$ . . . . . (i)

∴ $P(X > x_1) = P(Z > -z_1) = 0.93$
$= P(-z_1 < Z < 0) + P(0 < Z < \infty)$
$= P(0 < Z < z_1) + 0.5 = 0.93$ by symmetry
$= P(0 < Z < z_1) = 0.43$



The value of probability closer to 0.43 in the normal table is 0.4306 at Z = 1.48.

∴ $z_1 = -1.48$

From equation (i)

$\frac{X_1-0.753}{0.004} = z_1$

or, $\frac{X_1-0.753}{0.004} = -1.48$

or, $X_1 = 0.753 - 1.48 \times 0.004$

or, $X_1 = 0.7471$
Hence, the required value of X is 0.7471.

The CMAT score of MBS students in T.U. is normally distributed with mean score is 65 and standard deviation 10.

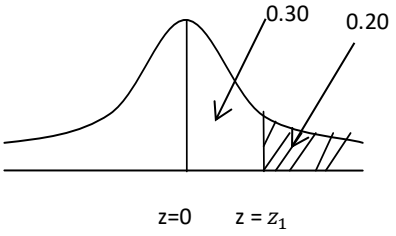
(i) Find the lowest score of top 20% students.
(ii) Find the highest score of lowest 30% students.
(iii) What are the limits within which the middle (or central) 50% of the scores lie?

Solution: Let X be the random variable which denotes the score of the students follows normal distribution with
Mean score $(\mu) = 65$
Standard deviation $(\sigma) = 10$
Then, the corresponding standard normal variate (S.N.V.) is

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

(i) Let the X_1 be the lowest score of top 20% students. Then

$P(X > x_1) = 0.20$
or, $P(z > z_1) = 0.20$
or, $P(0 < z < \infty) - P(0 < z < z_1) = 0.20$
or, $0.5 - P(0 < z < z_1) = 0.20$
or, $P(0 < z < z_1) = 0.30$



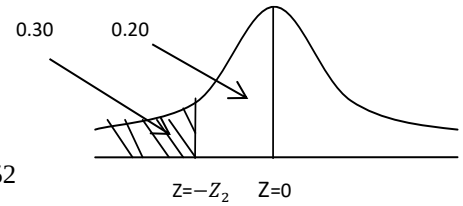
From normal table, the value of z closer to 0.3 is 0.2995 at z = 0.84

$\therefore z_1 = 0.84$
$z_1 = \frac{x_1-\mu}{\sigma}$
or, $0.84 = \frac{x_1-65}{10}$
or, $x_1 - 65 = 0.84 \times 10$
or, $x_1 = 65 + 0.84 \times 10 = 73.4$

Hence, the lowest score of top 20% students is 73.4

(ii) Let the X_2 be the highest score of lowest 30% students. Then

$P(X < x_2) = 0.30$
or, $P(z < -z_2) = 0.30$
or, $P(-\infty < z < 0) - P(-z_2 < z < 0) = 0.30$
or, $0.5 - P(-z_2 < z < 0) = 0.30$
or, $P(-z_2 < z < 0) = 0.20$ by symmetry
or, $P(0 < z < z_2) = 0.20$



From normal table, the value of z closer to 0.20 is 0.1285 at z = 0.52

$\therefore z_2 = -0.52$

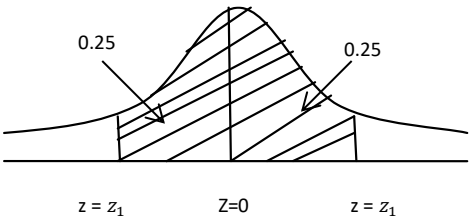
$z_2 = \frac{x_2-\mu}{\sigma}$
or, $-0.52 = \frac{x_1-65}{10}$
or, $x_2 - 65 = -0.52 \times 10$
or, $x_2 = 65 - 0.52 \times 10 = 59.8$

Hence, the highest score of lowest 30% students is 59.8

(iii) Let x_1 and x_2 be the lower and upper limits within which the middle (or central) 50% of the scores lie.

$P(-x_1 < X < x_2) = 0.5$
or, $P(-z_1 < z < z_2) = 0.5$
or, $P(-z_1 < z < 0) = P(0 < z < z_2) = 0.25$

Taking
$P(-z_1 < Z < 0) = 0.25$ by symmetry
$P(0 < Z < z_1) = 0.25$



From normal table, the value of z closer to 0.25 is 0.2486 at z = 0.67

$\therefore z_1 = -0.67$

$z_1 = \frac{x_1-\mu}{\sigma}$

or, $-0.67 = \frac{x_1-65}{10}$

or, $x_1 - 65 = -0.67 \times 10$

or, $x_1 = 65 - 0.67 \times 10 = 58.3$

Taking

$P(0 < X < z_2) = 0.25$

From normal table, the value of z closer to 0.25 is 0.2486 at z = 0.67

$\therefore z_1 = 0.67$

$$z_1 = \frac{x_1-\mu}{\sigma}$$

or, $0.67 = \frac{x_1-65}{10}$

or, $x_1 - 65 = 0.67 \times 10$

or, $x_1 = 65 + 0.67 \times 10 = 71.7$

Hence, Lower limit = 58.3

Upper limit = 71.7

Range = 71.7−58.3

[redacted]

The marks obtained by 1000 students in an examination are known to be normally distributed. If 15% of the students got less than 30 marks and 10% of the students got over 90, find the mean and standard deviation of the distribution.

Solution:

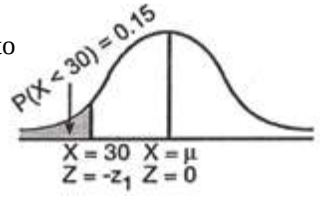
Let X be the normal variable which denotes the marks obtained by the students with mean μ and standard deviation σ.

According to question,

P(X < 30) = 0.05 . . . . . (i)

P(X > 90) = 0.10 . . . . . (ii)

Let z_1 and z_2 be the standard normal variates corresponding to Marks 30 and 90 respectively. Then,



From equation (i)

For, X = 30 = x_1

$\therefore$ P(X < 30) = 0.15

or, P (Z < $-z_1$) = 0.15

or, P ($-\infty$ < Z < 0) – (P (z_1 < Z < 0) = 0.15

or, 0.5 – P (0 < Z < z_1) = 0.15 [By symmetry]

or, P (0 < Z < z_1) = 0.35

The value of probability nearest to 0.35 in the normal table is 0.3508 at Z = 1.04.

$\therefore z_1 = -1.04$

Now,

For X = 30 = x_1

$$z_1 = \frac{x_1-\mu}{\sigma}$$

or, $-1.04 = \frac{30-\mu}{\sigma}$

or, $-1.04\,\sigma = 30 - \mu$

or, $\mu = 30 + 1.04\,\sigma$ . . . . . (iii)

Again, from equation (ii)

For, X = 90 = x_2

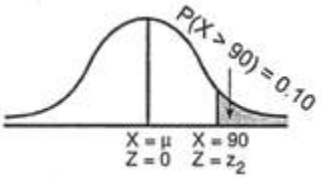
P(X > 90) = 0.10

or, P (Z > z_2) = 0.10

or, P (0 < Z < ∞) – P (0 < Z < z_2) = 0.10

or, 0.5 – P (0 < Z < z_2) = 0.10

or, P (0 < Z < z_2) = 0.40



The value of probability nearest to 0.40 in the normal table is 0.3997 at Z = 1.28.

$\therefore z_2 = 1.28$

$z_2 = \frac{x_2 - \mu}{\sigma}$

or, $1.28 = \frac{90 - \mu}{\sigma}$

or, $1.28\,\sigma = 90 - \mu$

or, $\mu = 90 - 1.28\,\sigma$ ..... (iv)

From (iii) & (iv)

$30 + 1.04\,\sigma = 90 - 1.28\,\sigma$

or, $2.32\,\sigma = 60$

or, $\sigma = \frac{60}{2.32} = 25.86$

From equation (iii), we get

$\mu = 30 + 1.04\,\sigma$

$= 30 + 1.04 \times 25.86 = 56.89$

Hence, the mean and standard deviation of the distribution are 56.89 and 25.86 respectively.

Unit 5 Sampling Survey

Introduction

Sampling is a day to day phenomenon for all of us and it is an essential part of any statistical and research investigation as well. Therefore, Sampling has been employing not only in any statistical investigation and research works but also in the case of daily life of human beings. Almost all research studies involve sampling. For examples, a businessman gives order for the commodities by examining a small unit of the same commodity. Customers often examine only a handful of rice, pulses or any commodity from a packet (sack) in a shop to check its quality before purchasing. A housewife makes judgement about the taste of curry by a spoonful of curry. Similarly, a pathologist takes a sample of blood to test any change in blood of the whole human body. Therefore, sampling is the process by which inference is made to the whole by examining only a part.

Population:

When we think of the term "population," we usually think of people in our town, region, state or country. But in statistics, the term "population" takes a slightly different meaning. In statistics, "population or universe" is the aggregative of objects under in any statistical investigation. That is, statistical population refers the totality or aggregative of all units or items or animate / inanimate subjects under investigation or study. It is also called universe e.g. all the patients in a hospital, all the fruit trees in garden, all the books in a library, all the fishes in a pond, all the people in a town, all the cattle in the cattle farm etc.

There are various types of population which are discussed as follows:

Finite population

A population containing a finite number of objects or items is known as finite population. For example, total no. of students of T.U. in BCA, the total number of books in a library, etc.

Infinite population

A population having an infinite number of objects is called infinite population. For examples, the number of stars in the sky, the total number of fishes in Pacific Ocean etc.

Population may be further be classified as real population and hypothetical population

Study population: Study population is a subset of the target population from which samples are drawn. The study population is the group of individuals in a study. It is also known as the accessible population. It is the population in research to which the researcher can apply their conclusions.

Sampling frame: A complete list of the sampling units, which represents the population to be covered, is called the sampling frame popularly known as frame. In other words, a list of all the elements that is in the population. The list of all the units in the reference population (target population) from which samples are to be picked for statistical investigation is known as sampling frame.

Sampling unit: A sampling unit refers to any single person, animal, plant, product or 'thing' being selected in the sampling process for research. In other words, a case or a single unit that is selected from a target population and measured in some way on the basis of analysis (e.g., a person, a thing) is called sample unit. For example, when studying a group of college students, a single student could be a sampling unit. Therefore, the sampling unit is the basic unit containing the elements of the population to be sampled. The sampling unit selected is often dependent upon the sampling frame. The selection of sampling units is also partially dependent upon the overall design of the project.

Sample:
A part or portion of the population which is considered for study and analysis is called a sample. In other words, a finite subset of the population with the objective of investigating its properties is called a sample and the number of units in the sample is known as the sample size. The process of selecting some units from the population in order to draw conclusion about the population is known as sampling. Thus, sampling is a process of selecting a representative portion of a population.

Sample is classified into two types.

(i) Random sample
(ii) Non random sample.

Random sample
A random sample is a set of items or units that have been drawn from a population in such a way that each time an item is selected, every item or unit of the population has equal chance to appear in the sample. In other words, a set of items or units in which the units are selected in such a way that each and every unit in the population has equal and independent chance of being selected from the population is called random sample.

Non random sample
A non-random sample is a set of items or units that have been drawn from a population in such a way that each time an item is selected, every item or unit of the population has no equal chance to appear in the sample.
In other words, a set of items or units in which the units are selected in such a way that each and every unit in the population has no equal and independent chance of being selected from the population is called non-random sample. Therefore, in non-random sample, the choice of selection of sampling units depends entirely on the discretion or judgment or convenience, beliefs, biases of sampler or investigator.
Sampling: The process (technique) of selecting some units (i.e. study units) from a defined population in order to draw conclusion about the population under study or investigation is known as sampling. Thus, sampling is a process of selecting a representative portion of a population. Usually, a representative subgroup of the population (sample) is included in the investigation.

Sampling Design:
A sample design is a definite plan for obtaining a sample from the sampling frame. It refers to the technique or procedure that the researcher would adopt in selecting some sampling units from which inferences about the population are drawn.
Sampling units: The sampling unit is the basic unit containing the elements of the population to be sampled. It may be the element itself or a unit in which the element is contained. The sampling unit selected is often dependent upon the sampling frame. The selection of sampling units is also partially dependent upon the overall design of the project.

Census Survey (or complete enumeration Survey)
Survey is the technique of investigation by direct observation of a phenomenon or individuals of collection of information through questionnaires or interview. The survey is used for describing current practice and events and analyzing the facts.
A census is a study of every unit, everyone or everything, in a population. That is, a census survey is the complete enumeration or count of all units of the population for certain character of the population. The term census is used mostly in connection with National population and Housing

Censuses and other common censuses include agriculture census, business census etc. Census requires more money, manpower and time.

Sample Survey

A sample is a subset of units in a population, selected to represent all units in a population. It is a partial enumeration because it is a count from part of the population. Therefore, the technique (or survey) in which only part of the population is selected and examined to estimate the certain character of the population is known as sample survey. Information from the sampled units is used to estimate the characteristics for the entire population. A sample survey will usually be less expensive than a census survey and the desired information will be obtained in less time.

When and Where Sampling/Census is Appropriate:

A Sampling Technique is more desired (more appropriate) than census

- When the universe is very large i.e. infinite and it would be impossible to conduct census survey.
- When quick results are required, it would be appropriate to conduct sample survey rather than census surveys.
- When to save time and money.
- When the universe possess homogeneous characteristics
- When utmost accuracy is not required
- Where census is impossible i.e. in destructive/explosive nature of testing.
- In the hypothetical population, sampling is the only one process to estimate population parameter.

A census (or complete enumeration) is more appropriate than sampling when

- The universe is small
- The population is heterogeneous
- Hundred percent accuracy is required.
- The population frame is incomplete.

Advantages of Sampling over Census

the parameters of the population.

Demerits of Sampling Technique:

Less accuracy. The following are the principal advantages or merits of sampling over complete enumeration (or census survey)

- Reduced cost.

  Sampling is much more economical than a complete enumeration. In sample survey, there is reduction in the cost of collection of the information, administration, transport, training and man hours. Although, the labour and the expenses of obtaining information per unit are generally large in sample enquiry than in the census survey, the overall expenses of a survey are relatively much less , since only a fraction of the population is to be enumerated.

- Greater speed.

  Since only a part of the population is to be inspected and examined, the sampling results in considerable saving in time and labour. Therefore, the sampling results can be obtained more rapidly and the data can be analyzed much faster as relatively fewer data have to be collected and processed. There is saving in time not only in conducting the sampling inquiry but also in the processing, editing and analyzing the data.

- Greater accuracy.

  Because of highly qualified, skilled and trained personnel, intensive training to the investigator and using more sophisticated equipment and statistical technique for the processing and analysis of the data, sampling thus provides relatively more accurate and reliable results than complete enumeration.

- Greater scope.

Sample survey has generally greater scope as compared with complete enumeration. The complete enumeration is impracticable if the survey requires a highly trained personnel and more sophisticated equipment for the collection and analysis of the data. Sampling procedure is more readily adaptable than census for statistical investigations.

- Destructive testing

If the testing of units is destructive i.e. if in the course of inspection the units are destroyed, the complete enumeration is impracticable and in such cases sampling technique is the only method to be used. For example, to test the quality of explosives, crackers, shells etc., to test the breaking strength of chalks manufactured in a factory.

- Infinite population

If the population is infinite or too large, then sampling procedure is only used for estimating the parameters of a population. For instance, the number of fish in the sea, the number of wild elephants in a dense forest can be estimated only by sampling method.

- Hypothetical population
    - In case of hypothetical population as for example in the problem of throwing a die or tossing a coin , where the process may continue large number of times or indefinitely, the sampling procedure is the only scientific technique of estimating
    - Inaccurate and Misleading conclusions.
    - Need for specialized knowledge and trained and qualified personnel.
    - When sampling is not possible.
    - Need sophisticated equipment for its planning, execution and analysis.
    - It cannot be used if the information of each and every unit of the population is required.

Demerits of Census:

- Expensiveness
- Excessive time and effort.
- Not applicable for destructive testing.

Distinction between census and Sampling:

Census	**Sampling**
• Census is the complete enumeration of population in which each and every element or unit of population is taken to study. • It is free from sampling error. • It is costly and time consuming i.e. it requires more money, manpower and time. • It has less scope. • It provides higher accuracy than sampling. • It is appropriate when the universe is small. • It is appropriate when the population is heterogeneous.	• Sampling is the process of measuring the characteristics of the population or universe by studying only a part or portion of the population chosen. • It is not free from sampling error. • It is less expensive and less time consuming i.e. it requires less money, manpower and time. • It has a greater scope than census. • It has less accuracy. • It is appropriate when the universe is large. • It is appropriate when the universe possess homogeneous characteristics.

Objective of Sampling

The objective of sampling procedure is to study about the population by selecting representative units from the concerned population through minimum resources viz. time, cost and energy or effort without loosing the accuracy. Sampling is used not only in case of research works but also in the case of daily life. Hence, sampling method is more desirable under limited cost and quick decision. The major objectives are:

- Obtain the maximum possible information about universe or population with minimum time, cost and effort.
- To obtain the limits of accuracy and degree of confidence of estimates of population parameter.
- To test the significance of population parameters on the basis of sample statistics.

Principal (Major) Steps in a Sample Survey (Organizational aspect of Sample survey)

The main (major) steps involved in the planning and execution of the sample survey are as follows:

1. Objectives of the survey:

   The first and foremost step in a sample survey is to define the objective of the survey in clear and concrete terms. It should be kept in mind that these objectives of the survey are commensurate with the available resources in terms of money, manpower and the time limit within which the results of the survey are desired.

2. Defining the population to be sampled.

   The population from which sample is to be chosen should be clearly and unambiguously defined. The geographical, demographic and other boundaries of the population must be specified so that no ambiguity arises regarding the coverage of the survey.

3. Getting a sampling frame.

   Before selecting the sample, the list of sampling units should be constructed. The list of sampling units from which the sample is to be drawn is called sampling frame. The sampling frame is the base for drawing a sample; therefore it should be made up to date, free from omission and duplications.

4. Choice of sampling unit

   Before selecting the sample, the population must be divided into parts that are called sampling units or units. These units must be the representative of the entire population. The sampling unit selected is often dependent upon the sampling frame. The selection of sampling units is also partially dependent upon the overall design of the project. In sample survey, sampling units should be unambiguous, specific, non-overlapping and appropriate to the enquiry.

5. Selection of proper sampling design.

   From among the several available sampling designs like Simple Random sampling, stratified Random Sampling, Systematic Sampling etc., an appropriate sampling design should be selected, keeping in view the objective of the survey, the nature of the population, cost involved and time required for the availability of the results and the degree of precision aimed at. If a number of sampling designs for selecting a sample is available, then the cost and precision should be considered before making a final selection of the sampling design. The relative costs and time involved for each design (plan) should also be compared before making a decision. The selection of proper sampling design provides quite reliable estimates.

6. Data to be collected.

   The data should be collected keeping in view with the nature, objectives and scope of the survey. An attempt should be made to eliminate the collection of irrelevant and unnecessary data which are never used subsequently and inessential information is omitted.

7. Questionnaire or schedule

Having decided about the type of the data to be collected, the next important part of the sample survey is the construction of the questionnaire or schedule of enquiry. The questions should be clear, unambiguous and to the point. The questionnaire or the schedule should be designed with utmost care and caution, keeping in view the knowledge, understanding and the general educational level of respondents.

8. Method of collecting of data (information)

There are two methods commonly used for obtaining numerical data for human population:

(i) Direct personal Investigation or Interview Method.

(ii) Mailed Questionnaire Method

A choice between the two methods depends on the objectives of the enquiry, the expenses involved and the accuracy of the results aimed at.

9. Organization of field work

To obtain reliable results in a sample survey, the sampling errors should be minimized. For this it is essential that the field work is properly organized and the personnel engaged in conducting and execution of the survey should be properly trained to handle the problems of the survey like locating the sample units, use of the equipment and recording of the observations, the methods of collecting the desired information, dealing with non-response etc.

10. Summary and Analysis of the data

After the planning and execution of the sample survey, the last step is the analysis of collected data. It basically involves the following steps:

(i) Scrutiny and Editing of the data.

(ii) Tabulation of data.

(iii) Statistical analysis.

(iv) Reports, summary and conclusions.

Questionnaire Design

Questionnaire is a formal list of question designed to gather responses from respondents on given topic. Thus, a questionnaire is an efficient data collecting mechanism when the researcher knows exactly what is required and how to measure the variable of interest. A questionnaire translates the research objectives into specific questions.

A questionnaire can be designed to secure different types of primary data from the respondents:

(a) intentions (b) attitudes and opinions (c) activities or behavior and (d) demographic characteristics.

Requisites of a questionnaire

Questionnaire depends upon the nature of the study, type of the respondents, quality of the field workers etc. There is no hard and fast rule for preparing a good questionnaire. However, the following general consideration might be taken while preparing questionnaire.

- The questions should be few, short, clearly worded, simple and easy to reply.
- The questions should be within the information scope of the respondents.
- The questions should be precise and long questions should be avoided as far as possible.
- The questions should be so sequenced that the respondents are motivated to answer all questions.
- They should have direct relation with objectives of the investigation or research problem.

- Units and technical terms are not used in non- technical question as far as possible.
- The questions should be inter-related with each other.
- The questions should proceed in logical sequence moving from easy to more difficult questions.
- Personal and intimate questions are not to be included as for as possible.
- Emotional questions should be avoided.
- The questions should be free from ambiguity.
- The questions should not be vague.
- The questions should be so framed. Questions may be dichotomous or multiple choice. Open ended questions should be avoided.
- The questions should be such that there is minimum of writing.
- The questions should be crossed checked so as to avoid manipulation by respondents.
- Why, what, when, how etc. can be used for measuring the degree of intensity of feeling, mood, gestures, convictions etc.

Sample Selection

Determination of Sample Size

Determination of sample size for estimating a population mean or population proportion is a crucial question because if sample size is too small it will not represent the population precisely and the objective will not be achieved. Similarly, if sample size is too large it represents population more precisely but it will waste more time, resources for gathering sample information. So, the determination of appropriate sample size (i.e. optimum sample size) so that the population parameters may be estimated with a desired degree of accuracy is one of the major problems in the sampling theory.

1) Determination of sample size (n) to estimate population mean

Case I: For SRSWR (i.e. for infinite population i.e. population size N is not given)

$\boldsymbol{n = \left(\frac{Z_{\alpha/2}\,\sigma}{d}\right)^2}$ This gives the required sample size for estimating population mean in case of infinite population (i.e. SRSWR)

Where, n = required sample size

SRSWR = Simple random sampling with replacement,

d = Permissible error or difference between sample mean and population mean.

$= |\bar{X} - \mu|$

$\boldsymbol{\sigma}$ = Population standard deviation

$\mathbf{Z_{\alpha/2} = Z_{tab}}$ = Critical value or tabulated value or significant value at $\boldsymbol{\alpha}$ level of significance

If population standard deviation (σ) in unknown

$$\boxed{n = \left(\frac{Z_{\alpha/2}\,s}{d}\right)^2} \qquad \because \hat{\sigma} = s \qquad \text{For large sample}$$

s = Sample standard deviation

Example 1

The research division of Nepal dairy wants to estimate the mean amount of milk fills in one litre packet within ± 0.01 litres with 99% confidence. The division assumes the standard deviation as 0.05 litre, what sample size is required for survey?

Solution:

Sample size (n) =?

Error (E) = $|\pm \mathbf{0.01}| = \mathbf{0.01}$

Confidence level, $\mathbf{1 - \alpha = 99}\%$ = 0.99

Standard deviation $(\boldsymbol{\sigma}) = 0.05$

$$\frac{1-\alpha}{2} = \mathbf{0.495}$$

$\mathbf{Z_{\alpha/2} = 2.575}$ (From normal table)

$$\boldsymbol{n} = \left(\frac{Z_{\alpha/2}\,\sigma}{d}\right)^2$$

$$= \left(\frac{2.575 \times 0.05}{0.01}\right)^2$$

$= 165.765 \approx \mathbf{166}$

Example 2

A researcher wants to estimate the universe mean by using sampling technique. What should be the sample size when permissible error between parameter value and sample statistic in 98 % of chance will not be more than 1.5 and population standard deviation is 15.

Solution:

Sample size (n) =?

Error (d) = 1.5

Population standard deviation $(\boldsymbol{\sigma})$ = 15

Confidence level, $\mathbf{1 - \alpha = 98}\%$ = 0.98

$\frac{1-\alpha}{2} = \mathbf{0.49}$ (Two tailed test)

$\mathbf{Z_{\alpha/2} = 2.33}$ (From normal table)

$$\mathbf{n} = \left(\frac{\mathbf{Z_{\alpha/2}\,\sigma}}{\mathbf{d}}\right)^2$$

$$= \left(\frac{2.33 \times 15}{1.5}\right)^2$$

$= 542.89 \approx \mathbf{543}$

Example 3

A researcher wishes to estimate the mean of a population by using sufficiently large sample. The probability is 0.95 that the sample mean will not differ from the true mean by more than 25% of the standard deviation. How large a sample should be taken?

Solution:

Sample size (n) =?

Error (E) = 25% of standard deviation = $\frac{25\,\sigma}{100} = \frac{\sigma}{4}$

Confidence level, $\mathbf{1 - \alpha = 95}\%$ = 0.95

$\frac{1-\alpha}{2} = 0.475$ (Two tailed test)

$Z_{\alpha/2} = 1.96$ (From normal table)

$$n = \left(\frac{Z_{\alpha/2}\,\sigma}{d}\right)^2$$

$$= \left(\frac{1.96\ \sigma}{\frac{\sigma}{4}}\right)^2$$

$$= (1.96 \times 4)^2$$

$$= 61.4656 \approx 62$$

Example 4

The average outstating balance of loan issued by a bank varies from month to month. From the past experience it is known that the amounts are normally distributed with standard deviation of Rs. 5000. The bank wishes to estimate the average by drawing a random sample such that the probability is 0.95 that the mean of the sample will not deviate by more than Rs. 600 from the population mean. What should be sample size?

Solution:

Sample size (n) =?

Error (d) = Rs. 600

Population standard deviation (σ) = Rs. 5000

Confidence level, $1 - \alpha = 95\%$ = 0.95

$\frac{1-\alpha}{2} = 0.475$ (Two tailed test)

$Z_{\alpha/2} = 1.96$ (From normal table)

$$n = \left(\frac{Z_{\alpha/2}\,\sigma}{d}\right)^2$$

$$= \left(\frac{1.96\times\ 5000}{600}\right)^2$$

$$= 266.77 \approx 267$$

Case II: For SRSWOR (i.e. for finite population i.e. Population size N is given)

Sample size $(n_0) = \frac{n}{1+\frac{n}{N}}$ is the required sample size for estimating population mean in case of finite population (i.e. SRSWOR)

SRSWOR = Simple random sampling without replacement.

Example 5

The mean systolic blood pressure of a certain group of people was found to be 125 mm of Hg with standard deviation of 15 mm Hg . Calculate sample size at 5 % level of significance if error does not exceed 2 and population of size 500.

Sample size (n) =?

Error (d) = 2

standard deviation (σ) = 15 mm Hg

Level of significance, $\alpha = 5\%$

Confidence level, $1 - \alpha = 95\%$ = 0.95

$\frac{1-\alpha}{2} = 0.475$ (Two tailed test)

$\mathbf{Z_{\alpha/2} = 1.96}$ (From normal table)

$$\mathbf{n} = \left(\frac{\mathbf{Z_{\alpha/2}\,\sigma}}{\mathbf{d}}\right)^{\mathbf{2}}$$

$$= \left(\frac{\mathbf{1.96\times\ 15}}{\mathbf{2}}\right)^{\mathbf{2}}$$

$$= 216.09 \approx \mathbf{216}$$

When Population size (N) = 500

Sample size $(\boldsymbol{n_0})$ = $\frac{\boldsymbol{n}}{\mathbf{1+\frac{n}{N}}}$ = $\frac{\mathbf{216}}{\mathbf{1+\frac{216}{500}}} = \mathbf{150.83 \approx 151}$

2) Determination of sample size (n) to estimate population proportion

Case I: For SRSWR (i.e. for infinite population i.e. population size N is not given)

$$n = PQ\left(\frac{Z_{\alpha/2}}{d}\right)^2$$

This gives the required sample size for estimating population proportion in case of infinite population.

Where, n = required sample size

$\mathbf{d = |p - P|}$ = The maximum allowable error or permissible error or difference between sample proportion and population proportion.

$\mathbf{Z_{\alpha_{/2}} = Z_{tab}}$ = Critical value or tabulated value or significant value at $\boldsymbol{\alpha}$ level of significance

If population proportion P is unknown, then we may use its unbiased estimate (p) $\boldsymbol{i.e.\widehat{P} = p}$

$$\mathbf{n = pq\left(\frac{Z_{\alpha/2}}{d}\right)^2} \qquad \because \mathbf{\widehat{P} = p} \qquad \text{For large sample}$$

Note: *If no pilot study (No previous surveys have been taken) so for the most conservative sample we take P=0.5, Q = 0.5*

Example 6

Suppose a certain hotel management is interested in determining the percentage of the hotel's guests who stay for more than 3 days. The reservation manager wants to be 95% confident that the percentage has been estimated to be within $\pm\,\mathbf{3}\%$ of the true value. What is the most conservative sample size needed for this problem?

Solution:

d = $|\pm\mathbf{3}\%|$ = 0.03

Sample size, n = ?

Since, confidence level, $\mathbf{1 - \alpha = 95}\%$ = 0.95

$$\frac{\mathbf{1-\alpha}}{\mathbf{2}} = \mathbf{0.475}$$

$$\boldsymbol{Z_{\alpha/2}} = \mathbf{1.96}$$

Since, no pilot study (No previous surveys have been taken) so for the most conservative sample we take P=0.5, Q = 0.5

Then, the sample size needed for the given problem is given by

$$\mathbf{n = P\,Q\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2}$$

$$= \mathbf{0.5 \times 0.5\left(\frac{1.96}{0.03}\right)^2}$$

$$= 1067.11 \approx \mathbf{1067}$$

Example 7

The principal of a college wants to estimate the proportion of smoker among his students. What size of a sample should he select so as to have the proportion of smokers not to exceed by 10% with almost certainty? It is believed from previous records that the proportion of smokers was 0.30?

Solution:

Error, d = 10% = 0.1

Sample size, n = ?

Since, for almost certainty, confidence level, $\boldsymbol{Z_{\alpha/2} = 3}$

Since, from previous records, proportion of smokers, p = 0.30, q = 1- p = 1-0.30 = 0.70

Then, the sample size needed for the given problem is given by

$$\mathbf{n = pq\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2}$$

$$= \mathbf{0.30 \times 0.70\left(\frac{3}{0.10}\right)^2}$$

$$= 189$$

Case II: For SRSWOR (i.e. for finite population i.e . Population size N is given)

Sample size $(\boldsymbol{n_0})$ = $\frac{n}{1+\frac{n}{N}}$ is the required sample size for estimating population proportion in case of finite population.

Example 8

What should be the size of sample if a simple random sample from a population of 4000 items is to be drawn to estimate the percent defective within 2% of the true value with 95%? What would be the size of sample if the population is assumed to be infinite?

Solution:

Sample size (n) =?

N = 4000

E = 2% = 0.02

Since, confidence level, $\mathbf{1 - \alpha = 95\%} = 0.95$

$\frac{1-\alpha}{2} = \mathbf{0.475}$ (Two tailed test)

$\mathbf{Z_{\alpha/2} = 1.96}$ (From normal table)

Since, no previous surveys have been taken, so we assume P = 0.5, Q = 0.5

$$\mathbf{n = P\,Q\left(\frac{Z_{\frac{\alpha}{2}}}{E}\right)^2}$$ (For infinite population i.e. SRSWR)

$= \mathbf{0.5 \times 0.5 \left(\frac{1.96}{0.03}\right)^2}$

= 2401

For finite population (i.e. SRSWOR)

When population size, N = 4000

Sample size $(\boldsymbol{n_0})$ = $\frac{\boldsymbol{n}}{\mathbf{1+\frac{n}{N}}}$

$= \frac{\mathbf{2401}}{\mathbf{1+\frac{2401}{4000}}}$

$= \mathbf{1500.391 \approx 1500}$

Example 9

Determine the minimum sample size required so that the sample estimate lies within 10% of the true value with 95% confidence when coefficient of variation is 60%.

Solution:

Sample size (n) =?

Confidence level, $\mathbf{1 - \alpha = 95}$% = 0.95

$\frac{\mathbf{1-\alpha}}{\mathbf{2}} = \mathbf{0.475}$

$\boldsymbol{Z_{\alpha/2} = 1.96}$ (from normal table)

C.V. = 60%

E = 10%

Confidence level, $\mathbf{1 - \alpha = 95}$% = 0.95

$\frac{\mathbf{1-\alpha}}{\mathbf{2}} = \mathbf{0.475}$

$\boldsymbol{Z_{\alpha/2} = 1.96}$ (From normal table)

C.V. = 60%

Then, the required minimum sample size is given by

$\boldsymbol{n = \left(\frac{Z_{\frac{\alpha}{2}} \text{ C.V.}}{E}\right)^2}$

$= \left(\frac{\mathbf{1.96 \times 60}}{\mathbf{10}}\right)^2$

$= 138.297 \approx \mathbf{138}$

Example 10

Determine the minimum sample size required so that the sample estimate lies within 5% of the true value with 99% confidence when coefficient of variation is 40%.

Solution:

Sample size (n) =?

E = 5%% of true value = 0.05 $\bar{\mathbf{Y}}$

Confidence level, $\mathbf{1-\alpha = 99\%}$ = 0.99

$\frac{\mathbf{1-\alpha}}{\mathbf{2}} = \mathbf{0.495}$

$\boldsymbol{Z_{\alpha/2} = 2.575}$ (from normal table)

C.V. = 40%

Sample size (n) =?

E = 5%

Confidence level, $\mathbf{1-\alpha = 99\%}$ = 0.99

$\frac{\mathbf{1-\alpha}}{\mathbf{2}} = \mathbf{0.495}$

$\boldsymbol{Z_{\alpha/2} = 2.575}$ (From normal table)

C.V. = 40%

Then, the required minimum sample size is given by

$$\boldsymbol{n} = \left(\frac{\frac{\mathbf{Z_\alpha}}{\mathbf{2}}\,\mathbf{C.V.}}{\boldsymbol{E}}\right)^{\mathbf{2}}$$

$$= \left(\frac{\mathbf{2.575 \times 40}}{\mathbf{5}}\right)^{\mathbf{2}}$$

$= 424.36 \approx \mathbf{424}$

Example 11

Suppose it is required to estimate the average value of output of a group of 5000 pharmaceutical industries in an industrial city so that the sample estimate lies within 10% of the true value with a confidence coefficient of 95%. Determine the minimum sample size required. It is also known that the population coefficient of variation is 40%.

Solution:

Sample size (n) =?

N = 5000

E = 10%

Confidence level, $\mathbf{1-\alpha = 95\%}$ = 0.95

$\frac{\mathbf{1-\alpha}}{\mathbf{2}} = \mathbf{0.475}$ (For two tailed test)

$\boldsymbol{Z_{\alpha/2} = 1.96}$ (From normal table)

C.V. = 40 %

Then, the required minimum sample size is given by

$$\boldsymbol{n} = \frac{\mathbf{1}}{\left(\frac{\boldsymbol{E}}{\boldsymbol{Z_{\alpha/2}} \times \boldsymbol{C.V}}\right)^{\mathbf{2}} + \frac{\mathbf{1}}{\boldsymbol{N}}}$$

$$= \frac{\mathbf{1}}{\left(\frac{\mathbf{10}}{\mathbf{1.96 \times 40}}\right)^{\mathbf{2}} + \frac{\mathbf{1}}{\mathbf{5000}}}$$

$$= \frac{\mathbf{1}}{\mathbf{0.016269 + 0.0002}}$$

$= \frac{\mathbf{1}}{\mathbf{0.016469}} = 60.72.57 \approx 61$

Example 12

Determine the minimum sample size required so that the sample estimate lies within 10% of the true value with 95% confidence when coefficient of variation is 60%.

Solution:

Sample size (n) =?

E = 10%

Confidence level, $\mathbf{1-\alpha = 95\%} = 0.95$

$\frac{\mathbf{1-\alpha}}{\mathbf{2}} = \mathbf{0.475}$

$\boldsymbol{Z_{\alpha/2} = 1.96}$ (From normal table)

C.V. = 60%

Then, the required minimum sample size is given by

$$\boldsymbol{n} = \left(\frac{\mathbf{Z_{\frac{\alpha}{2}}}\ \text{C.V.}}{\boldsymbol{E}}\right)^{\mathbf{2}} \qquad \because \text{C.V.} = \frac{\boldsymbol{\sigma}}{\bar{\mathbf{Y}}} \times \mathbf{100}$$

$$= \left(\frac{\mathbf{1.96\times 60}}{\mathbf{10}}\right)^{\mathbf{2}}$$

$= 138.297 \approx \mathbf{138}$

Sampling error and non-sampling error

An 'error' refers the difference to the difference between the true value of a population parameter and its estimate provided by corresponding value of a sample statistic.

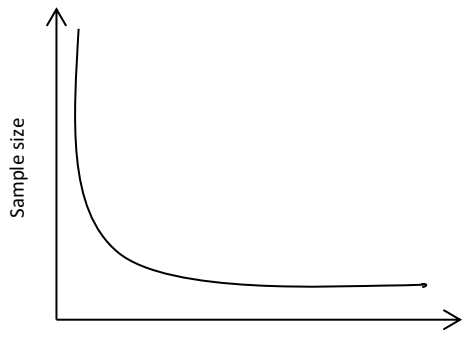
The inaccuracies or errors in any statistical investigation involved in collection, processing and analysis of the data may be broadly classified as follows:

(i) Sampling error (ii) Non-sampling error

(i) Sampling error

Sampling errors arise due to the fact that when a part of population (i.e. sample) has been used to estimate the population parameters and draw inferences about the population. So, the sampling errors are absent in complete enumeration survey or census Survey. The magnitude of sampling error depends upon the nature of the population and size of the sample. If the sample size increases, the sampling error will decrease.

Relationship between sample size and sampling error, graphically as shown in the following figure.



Sampling error

Sampling errors are occurred due to the following reasons:

- Faulty selection of the sample.
- Substitution of convenient unit of the population.
- Faulty demarcation of sampling units.
- Improper choice of the statistics for estimating population parameter.

(ii) Non sampling error

Non–sampling errors arise at the stages of planning, observation, ascertainment and processing of the data and are thus present in both the complete enumeration survey (or census) and sample survey. Thus, the data obtained from a complete census are free from sampling errors but may not free from non-sampling errors whereas data obtained from the sample survey would be both sampling and non-sampling errors. So, non- sampling errors are more serious in census survey with compared to a sample survey.

Non sampling errors arise from the following factors:

- Faulty planning or definitions
- Response errors
- Non –response errors
- compiling errors
- publication errors
- Coverage error.

Objectives of sampling

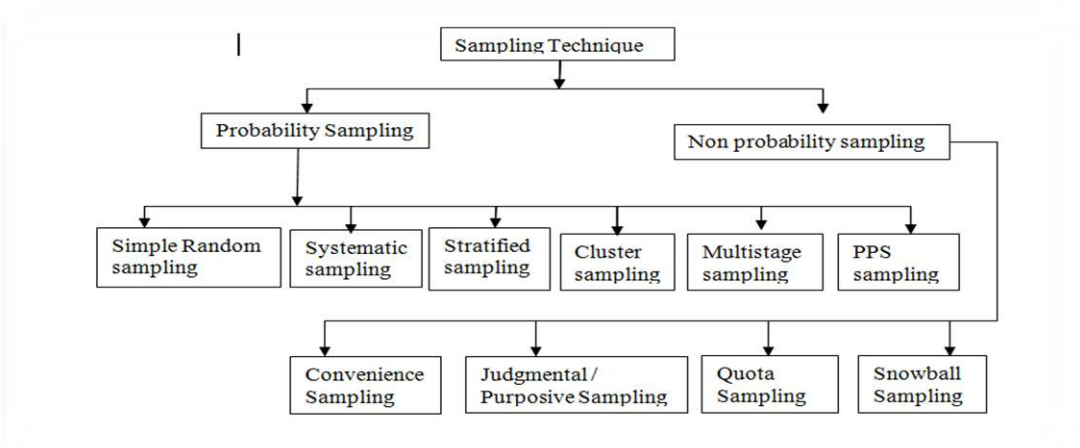
The objective of sampling is to study about the characteristics of population by selecting a portion of concerned population through minimum resources viz. time, cost and energy without loosing the accuracy. The main objectives of sampling are given below:

(i) To obtain as much information as possible about the population with the minimum time, cost and effort or limited resources.

(ii) To obtain the limits of accuracy and degree of confidence of the estimates of the population parameters.

(iii) To test the significance of population parameters on the basis of sample statistics.

Methods of Sampling (Sampling Techniques)



Types of Sampling

The method or process of selecting a sample from a population under study is called sampling. There are various methods of sampling techniques used for the selection of sample from the population. The method of selecting a sample from the population is of fundamental importance in sampling theory and depends very much on the objective and scope of the inquiry or investigation , nature of the universe or population, the size of the sample , available resources etc. Sampling methods which are commonly used can be classified into the following broad categories.

a. Random sampling (Probability Sampling)

b. Non random sampling (Non- Probability Sampling)

c. Mixed sampling

a. Random sampling (Probability Sampling)

Random sampling or probability sampling is the scientific method of selecting samples from the population according to some laws of chance in which each and every unit in the population has some definite pre-assigned probability of being selected in the sample. Probability samples are selected in such a way so as to be representative of the population. The probability samplings have the following characteristics:

(i) Each sample unit has an equal chance of being selected.

(ii) Sampling units have different probability of being selected.

(iii) Probability of selection of a unit is proportional to the sample size.

b. Non random sampling (Non- Probability Sampling)

Non-random sampling is the method of selecting samples, in which the choice of selection of sampling units depends entirely on the discretion or judgment or convenience, beliefs, biases of sampler or investigator. This method is mainly used for opinion surveys but cannot be recommended for general use as it is subject to the drawbacks of prejudice and bias of the investigator.

However, if the researcher is experienced and expert, it is possible that judgment sampling may yield useful results. However, this method suffers from a serious defect that is not possible to compute the degree of precision of estimate from the sample values.

c. Mixed sampling

If the samples are selected partly according to some laws of chance and partly according to a fixed sampling rule (i.e. no assignment of probabilities) they are termed as fixed samples and the technique of selecting such samples is known as mixed sampling.

Distinction between probability and non-probability samplings

Probability sampling	**Non probability sampling**
Probability sampling can be rigorously analyzed to the determine possible bias and likely error.	But, non-probability sampling does not provide this advantage.
It is free from personal biases.	It is not free from personal biases.
In comparison to non- probability sampling, it is less prefer to pilot study.	It is more appropriate in pilot study.
In this method, the degree of precision of estimate from sample value can be computed i.e. sampling error can be estimated.	But in this method, the degree of precision of estimate from the sample values cannot be computed.i.e. sampling error cannot be estimated.

Types of Random Sampling (Probability Sampling)

Some of the most important methods of selecting a random sampling are as follows:

1. Simple random sampling
2. Systematic sampling
3. Stratified sampling
4. Cluster sampling
5. Multi-stage sampling

1. Simple Random sampling (SRS)

The random sampling (i.e. probability sampling) in which the units are selected in such a way that each and every unit in the population has equal and independent chance of being selected from the population. It is the simplest and most common method of sampling in which the sample is drawn unit by unit, with equal probability of selection for each unit at each draw. This is also known as the equal probability sampling. It is the most elementary random sampling.

If a sample is to be taken at random, the following two methods can be used:

(i) Lottery method (ii) Use of table of random numbers.

(i) Lottery method: It is the simplest method of simple random sampling. Suppose one is interested to select 'n' units out of 'N'. Distinct numbers from 1 to N are assigned for each unit and these numbers are written on 'N' homogenous slips. These slips are put in a box or a bag and mixed thoroughly. Then 'n' slips are drawn one by one. But it will not be practicable if the number of items in the universe is very large and the slips are not identical.

(ii) Use of table of random numbers: Another easiest way of selecting samples is through the use of random number tables. These random numbers are generally generated by computer or by a table of random numbers. Tippet's random number tables, Fisher and Yates tables, Kendall and Smith's tables are some of commonly used random number tables. If the population is very large, in this case the lottery method for selecting random samples is more time consuming and tedious. In such case, the most

practical method of selecting a random sample is the random table method. So, this method can also be considered as better than the lottery method since lottery method is time consuming.

There are two types of simple random sampling (SRS)

- **SRSWOR:** If the unit selected in any draw is not replaced in the population before making the next draw, then it is known as simple random sampling without replacement (SRSWOR). If a sample of size n is drawn without replacement from a population of size N then there are N_{C_n} possible samples. Then, the probability of selecting (drawing) of a sample of size n from a population of size N in SRSWOR is $\frac{1}{N_{C_n}}$
- **SRSWR**: If a unit is selected and noted and then returned back or replaced back in the population before making the next draw, then the sampling procedure is called simple random sampling with replacement (SRSWR). If a sample of size n is drawn with replacement from a population of size N then there are N^n possible samples. The probability of selecting (drawing) of a sample of size n from a population of size N in SRSWR is $\frac{1}{N^n}$

Merits:

- Each item has equal chance of being selected. So, it depends upon the chance but not on the personal judgment.
- This method is quite economic and saves time and money as compared.
- It is more representative of the population as compared to the judgment or purposive sampling.

Demerits:

- Expensive and time consuming especially when the population is large.
- If the population is heterogeneous in nature, it may not give accurate results.
- It requires completely up-to-date list of population units from which samples to be drawn.

2. Systematic sampling

A random sampling in which only the first unit is selected at random and the remaining units are automatically selected according to pre-determined pattern (i.e. at fixed equal intervals from one another) is known as systematic sampling. Systematic sampling is a commonly used technique if the complete and up-to-date sampling frame is available i.e. complete and up-to-date list of sampling units is available.

Suppose N units of the population are numbered from 1 to N in some order. Let N = n k where n sample size and K is an integer known as sampling interval. Thus, $k = \frac{N}{n}$. Systematic sampling is a statistical method involving the selection of elements from an ordered sampling frame. This is random sampling with a system. From the sampling frame, a starting point is chosen at random, and choice thereafter is at regular intervals. For example, suppose you want to sample 8 houses from a street of 120 houses. 120/8=15, so every 15th house is chosen after a random starting point between 1 and 15. If the random starting point is 11, then the houses selected are 11, 26, 41, 56, 71, 86, 101, and 116.

This sampling is mostly used in forest surveys, fisheries surveys etc.

Sample selection procedure:

Assuming that population consists N units which can be expressed as N = n k + C

Here, n is desirable size of sample and k & c are some positive constants, depending upon the value of C the two different approaches can be used to select the systematic sample of desirable size.

(i) Linear systematic sampling (L.S.S.)

(ii) Circular systematic sampling (C.S.S)

- **Linear systematic sampling (L.S.S.):** If C = 0. This implies that N = n k be the units of population are numbered from 1 to N in some order. Where n is sample size & k = $\frac{N}{n}$ is an integer known as sampling interval and random number called random start is selected between 1 to K. Then every k^{th} unit will be selected automatically. This type of selection procedure of systematic sampling is known as linear systematic sampling (L.S.S).
- **Circular systematic sampling (C.S.S):** If N ≠ nk. Sometimes it is not possible to express N = nk i.e. N ≠ nk then k is taken as an integer nearest to $\frac{N}{n}$. Then a random number $1 \leq i \leq N$ is selected and then after every k^{th} unit is selected. This type of selection procedure of systematic sampling is known as circular systematic sampling (C.S.S).

Merits:

- This method is simple and convenient to use.
- In selecting the sample by this method, it takes less time and labour.
- Most of the results obtained from this method are satisfactory.
- If the complete list of the population is available and if the items are arranged systematically, this method is more efficient.

Demerits:

- The sample selected may not be the representative of the population in some cases.
- The items of the population must be arranged in some order otherwise the result obtained will be misleading.

3. Stratified sampling

Stratified random sampling is used when the characteristics of elements of population are heterogeneous. In a stratified random sampling approach, a population is first divided into subpopulations or strata, based upon one or more classification criteria in such a way that the characteristics of units are homogeneous within strata and heterogeneous between strata. Then sample is drawn from each stratum using simple random sampling method. These samples are combined (pooled) to form the required sample of the population. Thus, the procedure in which first stratification and then simple random sampling is known as stratified random sampling. The number of units to be sampled from each stratum depends on its size relative to the population.

Stratified sampling is appropriate to apply when the characteristics of elements of population are heterogeneous and there are different sub-groups in the population.

For example, if we want to study about the examination results of 2400 students of Mahendra Multiple Campus, Nepalgunj. At first we divide the no. of students into different faculties such as Law, management, humanities, Education and science. If the number of students in Law, management, humanities, Education and science are 900, 600, 200, 400 & 300 respectively. Then we select 20% of each for the samples i.e. selecting samples consisting of 180, 120, 40 & 60 students from each faculty respectively. The samples taken from each stratum are then pooled (combined) to form the overall sample (required sample). Similarly, in survey of average crop yield per hectare etc. A stratified sample is a mini-reproduction of the population. Before sampling, the population is divided into characteristics of importance for the research. For example, by gender, social class, education level, religion etc. Then the population is randomly sampled within each category or stratum.

A stratified sampling may be either

(i) Proportionate

(ii) or Disproportionate

In a proportionate stratified sampling, the number of items drawn from each stratum is proportional to the size of the strata. On the other hand, if an equal number of items are drawn from each stratum regardless of how the stratum represented in the population (universe), it is called disproportionate sampling.

Purpose of Stratification

- To make more representative.
- Greater accuracy.
- Administrative convenience

Merits:

- The units selected represents whole universe.
- The estimation of population parameters is more efficient.
- For large and heterogeneous population, stratified sampling is the best design.

Demerits

- This method requires more time and cost.
- If each stratum of the population be not homogeneous the result obtained may not be reliable.
- The samples from each stratum should be selected only by the experts or experience persons.

4. Cluster sampling

Cluster sampling is used in a situation where the population characteristics are heterogeneous. A Cluster sampling is a method or technique of random sampling in which the population is divided into different groups, called clusters, in such a way that the characteristics of units or elements within the cluster are heterogeneous and between the clusters are homogenous so that the number of sampling units in each cluster should be approximately same. Then, a cluster is selected as a sample by using simple random sampling. This method of random sampling is called cluster sampling. In cluster sampling individual clusters are representative of the population as a whole. For example, Let us suppose that we want to study about the economic condition of people of Nepalgunj sub- metropolitan city. First of all, sub-metropolitan city is divided into different wards in such a way that the economic condition of people within ward is heterogeneous and between wards are homogeneous. In this way, in each ward (i.e. cluster) each type of economic condition of people may lie i.e. poor, middle, richest and so on. Then, a ward (a cluster) is selected as sample by using simple random sampling method and it will represent the whole population. Then, we can study about the economic condition of people in Nepalgunj sub- metropolitan city.

Cluster sampling is appropriate to apply when

- The entire population unclear or unknown.
- The sample clusters are geographically convenient.
- The clusters are 'Natural' in population.

Merits

- It is less costly than simple random sampling and stratified sampling
- It is useful even when the sampling frame of elements may not be available.
- Elements (units) selected by well-designed cluster sampling procedures in easier, faster cheaper and more convenient than simple random sampling and stratified sampling.

Demerits

- The efficiency decreases with increase in cluster size.
- The efficiency cost per unit may be more in cluster sampling.
- Enumeration of the sampling units within the selected clusters is difficult when the population is large.

Difference between Stratified sampling and Cluster sampling

Stratified Sampling	Cluster Sampling
In this sampling, the population is divided into various groups or classes called strata in such a way that the characteristics of units are homogeneous within strata and heterogeneous between strata.	In this sampling, the population is divided into different groups called clusters in such a way that the characteristics of units within clusters are heterogeneous and between clusters is homogeneous.
The population is separated into strata, and then sampling is conducted within each stratum.	The population is separated into clusters, and then clusters are sampled.
In order to minimize sampling error, within group differences among strata should be minimized, and between/group differences among strata should be maximized.	In order to minimize sampling error, within group differences should be consistent with those in the population, and between-group differences among the clusters should be minimized.
Main purpose: increase precision and representation.	Main purpose: decrease costs and increase operational efficiency.
Categories are imposed by the researcher	Categories are naturally occurring pre-existing groups.

More precision compared to simple random sampling.	Lower precision compared to simple random sampling.
The variables used for stratification should be related to the research problem	The variables used for clustering should not be related to the research problem.
Common stratification variables: age, gender, income, race.	Common classification variables: Geographical area, school, grade level.
In this sampling, a random sample is drawn from all strata for study.	But in this sampling only the selected clusters are studied

Difference between Systematic sampling and cluster sampling

Systematic sampling	Cluster sampling
Systematic sampling uses fixed intervals from the larger population to create the sample.	While cluster sampling breaks the population down into clusters.
Systematic sampling selects a random stating point from the population and then a sample is taken from regular fixed intervals of the population depending on the size.	Cluster sampling divides the population into clusters and then a cluster is taken as a simple random sample.
Systematic sampling is used when complete and up to date sampling frame is available.	It is useful even the sampling frame of all elements or units is not available.
Systematic sampling is more precise than cluster sampling.	It is less precise.
More costly	Less costly.

5. Multi- -stage sampling

Multi-stage sampling is a further development of the principle of cluster sampling. Multi-stage sampling is also a random sampling in which sampling procedure is carried out i.e. Done in various stages. At first stage, the population is divided into large sample units i.e. large groups called first stage units (fsu) or primary stage units and cluster is selected as the sample at random from them. At the second stage the selected clusters at the first stage are further divided or sub-sampled into smaller sample units. The method selecting only some of the units of selected cluster is called two –stage sampling. And if it is generalized into third stage or more stages until get to ultimate units of sample size is known as multi-stage sampling. Multistage sampling is a complex form of cluster sampling. Although cluster sampling and stratified sampling bear some superficial similarities, they are substantially different. In stratified sampling, a random sample is drawn from all the strata, where in cluster sampling only the selected clusters are studied, either in single- or multi-stage.

Multi-stage sampling has been found to be very useful in practice specially in large scale survey. Multi-stage sampling can be considered as the combination of cluster sampling and random sampling. For example, in crop surveys for estimating yield to a crop in a district, rural municipality can be considered as primary sampling unit (psu), the villages as the second stage units, crop fields as third stage units and a plot of fixed size as the ultimate unit of sampling.

Similarly, for a survey of the living standard of families in the villages of Nepal, a sample of Anchals may be taken first. From each of these selected Anchals, a sample of districts is taken. From each of these selected districts, a sample of rural municipalities is selected. Then, a few villages are selected from each of the selected rural municipalities. Lastly, some families are selected from each of the selected villages to get the required sample for the survey.

Merits:

- It is more convenient when area of investigation is very large.
- It is commonly used in large scale survey.
- This method is also more flexible than other sampling methods.
- As sample size is reduced in each stage, this sampling technique saves time and cost.

Demerits:

- If the samples are not carefully taken from the different stages, it may give the faulty result.
- In this method, there is high chance of occurring sampling error when the selected sampling units are decreased.

Non probability sampling (Non –probability Sampling)

Non-random sampling is the method of selecting samples in which the choice of selection of sampling units depends entirely on the discretion or judgment, convenience, beliefs, biases of sampler or investigator. This method is mainly used for opinion surveys but cannot be recommended for general use as it is subject to the drawbacks of prejudice and bias of the investigator. However, if the researcher is experienced and expert, it is possible that judgment sampling may yield useful results. However, this method suffers from a serious defect that it is not possible to compute the degree of precision of estimate from the sample values.

Types of Non-random sampling or Non-probability sampling

Following are some of the common methods use in the non-probability sampling

(i) Judgment sampling or purposive sampling
(ii) Convenience sampling
(iii) Quota sampling
(iv) Snowball sampling or network sampling

(i) Judgment sampling

A sampling method, in which the choice of sample items depends entirely upon the judgment of the investigator, is called judgment sampling. In this method of sampling, the choice of sampling items depends exclusively on the judgment of the investigator. In other words, the investigator uses self-judgment in the choice and includes only those items of the universe which are conveniences to the investigator. It is the method for quick decision.

For instance, if we want to study of corruption in Nepalese society, we can select a sample of twenty of the senior professor of T.U. to give their opinion on the subject. We consider that the judgment of these professors is much superior to a convenience. Then we can get the desired information.

Merits:

- It is the simple method of sampling for quick decision.
- It gives the better result when sample size is small.

Demerits:

- It gives unreliable conclusion if the investigator is personally biased.
- Though simple, the method is not scientific and it is not in general use.
- Sampling error cannot be estimated because it is not based on random sampling.

(ii) Convenience sampling

A sampling method, in which the researcher selects the sample neither by probability nor by judgment but by convenience, is called convenience sampling. It is also called the chunk sampling. Selection of sampling units is totally based on the convenience of the researcher. The results obtained by this method can hardly be representative of the population. They are generally biased and unsatisfactory. However, convenience sampling is often used for making pilot studies. For instance, if any one wants to conduct 'man-on-the-street' interviews, he/she stands up in corner of the street and interviews the desired number passers-by. Then, required information can be obtained.

Merits:

- It is useful for making pilot studies.
- It is the simple method of sampling for quick decision.
- When both time and money are limited, convenience sampling is widely used i.e. it is less expensive and less time consuming.

Demerits:

- The results obtained by this method can hardly be representative of the population.
- Sampling error cannot be estimated because it is also not based on random sampling.

(iii) Quota sampling

A non-random sampling method in researcher are given quotas to be filled from the different strata and within pre-assigned quotas, the process of drawing selecting the required samples from these strata by judgment sampling is called quota sampling. Quota sampling is a type of judgment sampling. Sample quotas may be fixed according to some specified characteristics such as income group, sex, occupation, political or religious affiliation etc.

For instance, for the comment about the fiscal year budget in radio listening survey, the interviewer may select as a sample of 50 persons choosing from different areas such as 20 officials, 10 Professors, 10 businessmen, 5 farmers and 5 students. Here, interviewer is free to select the people to be interviewed for the comment.

Merits:

- It saves time and money rather than other sampling methods.
- It is stratified –cum-purposive so investigator enjoys the benefits of both.

Demerits:

- It may biased because of the personal believes and prejudices of investigator.
- Sampling error cannot be estimated because it is also not based on random sampling

(iv) Snowball sampling or Network sampling

Snowball sampling is a special type of non-probability sampling and is used when the desired sample characteristic is very rare. Therefore, this sampling design is widely used in applications where respondents are difficult to identify and are best located through referral networks. Hence, this sampling is also known as chain referral sampling or network sampling. In this sampling, an initial group is discovered and then subsequent respondents, possessing similar characteristics are identified based on referrals provided by the initial respondents.

This sampling is particularly used to study drug cultures (use), teenage gang activities, prostitution study, political activities, illegal activities etc.

Merits:

- This method is cheap, simple and cost–efficient.
- It is useful for rare populations for which no sampling frames are readily available.
- The chain referral process allows the researcher to reach populations that are difficult to sample when using other sampling methods.

Demerits:

- It is difficult to apply when the population is large.
- It does not ensure the inclusion of all elements in the list.

Example 1

A population consists of five numbers 1, 3, 5, 7 and 9

(i) Enumerate all possible samples of size two which can be drawn from them population without replacement.

(ii) Calculate the mean and variance of the population.

(iii) Show that the mean of the sampling distribution of the sample means is equal to the population mean.

(iv) Calculate the variance of the sampling distribution of sample mean.

(v) Calculate Standard error of mean

Solution.

Here, Population size, N = 5

Sample size, n = 2

(i) Number of possible samples of size 2 that can be drawn from the population without replacement, k = $N_{C_n} = 5_{C_2} = 10$

Thus , the possible samples are: (1, 3), (1, 5), (1, 7), (1, 9), (3, 5), (3, 7), (3, 9), (5, 7), (5, 9) & (7, 9)

(ii) Calculation of population mean and population variance:

Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
1	– 4	16
3	– 2	4
5	0	0
7	2	4
9	4	16
$\sum Y = 25$		$\sum(Y - \bar{Y})^2 = 40$

Population mean, $\mu = \bar{Y} = \frac{\sum Y}{N} = \frac{25}{5} = 5$

Population variance, $\sigma^2 = \frac{\sum(Y - \bar{Y})^2}{N} = \frac{40}{5} = 8$

(iii) Calculation of sample means and variance of the sampling distribution of means:

Sample no.	Sample values (X)	Sample means ($\bar{X}$)	$(X - \bar{X})$	$\sum(X - \bar{X})^2$
1	(1, 3)	2	– 3	9
2	(1, 5)	3	– 2	4
3	(1, 7)	4	– 1	1
4	(1, 9)	5	0	0
5	(3, 5)	4	– 1	1
6	(3, 7)	5	0	0
7	(3, 9)	6	1	1
8	(5, 7)	6	1	1
9	(5, 9)	7	2	4
10	(7, 9)	8	3	9
		$\sum \bar{X} = 50$		$\sum(X - X)^2 = 30$

Mean of the sample means or Grand mean, $(\bar{\bar{X}}) = \frac{\sum \bar{X}}{k} = \frac{50}{10} = 5$

Since, the mean of the sample means $(\bar{\bar{X}}) = 5$ equal to the population mean $\mu = 5$, we conclude that the mean of the sampling distribution of the sample means is equal to the population mean.

(ii) Variance of the sample means is

$$V(\bar{X}) = \frac{\sum(X - \bar{X})^2}{k} = \frac{30}{10} = 3$$

OR

$$V(\bar{X}) = \frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{8}{2}\frac{(5-2)}{(5-1)} = 3$$

(v) The standard error of sample mean is given by

S.E. $(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{3} = 1.732$

Example 2

A population consists of four numbers 2, 3, 5, and 6

(i) Enumerate all possible samples of size two which can be drawn from them population with replacement.

(ii) Calculate the mean and variance of the population.

(iii) Show that the mean of the sampling distribution of the sample means is equal to the population mean.

(iv) Calculate the variance of the sampling distribution of sample mean.

(v) Calculate Standard error of the sample mean.

Solution.

Here, Population size, N = 4

Sample size, n = 2

(i) Number of possible samples of size 2 that can be drawn from the population with replacement, k = $N^2 = 4^2 = 16$

Thus , the possible samples are: (2, 2), (2, 3), (2, 5), (2, 6), (3, 2), (3, 3), (3, 5), (3, 6), (5, 2), (5, 3), (5,5), (5,6), (6,2), (6,3), (6,5), (6,6)

(ii) Calculation of population mean and population variance:

Y	$Y - \bar{Y}$	$(Y - \bar{Y})^2$
2	– 2	4
3	– 1	1
5	1	1
6	2	4
$\sum Y = 16$		$\sum (Y - \bar{Y})^2 = 10$

Population mean, $\mu = \bar{Y} = \frac{\sum Y}{N} = \frac{16}{4} = 4$

Population variance, $\sigma^2 = \frac{\sum (Y - \bar{Y})^2}{N} = \frac{10}{4} = 2.5$

(iii) Calculation of sample means and variance of the sampling distribution of means:

Sample no.	Sample values (X)	Sample means ($\bar{X}$)	$(X - \bar{X})$	$\sum (X - \bar{X})^2$
1	(2, 2)	2	– 2	4
2	(2, 3)	2.5	– 1.5	2.25
3	(2, 5)	3.5	– 0.5	0.25
4	(2, 6)	4	0	0
5	(3, 2)	2.5	– 1.5	2.25
6	(3, 3)	3	– 1	1
7	(3, 5)	4	0	0
8	(3, 6)	4.5	0.5	0.25
9	(5, 2)	3.5	– 0.5	0.25
10	(5, 3)	4	0	0
11	(5,5)	5	1	1
12	(5,6)	5.5	1.5	2.25
13	(6,2)	4	0	0
14	(6,3)	4.5	0.5	0.25
15	(6,5)	5.5	1.5	2.25
16	(6,6)	6	2	4

		$\sum \bar{X} = 64$		$\sum (X - \bar{X})^2 = 20$

Mean of the sample means or Grand mean, $(\bar{\bar{X}}) = \frac{\sum \bar{X}}{k} = \frac{64}{16} = 4$

Since, the mean of the sample means $(\bar{\bar{X}}) = 4$ equal to the population mean $\mu = 4$, we conclude that the mean of the sampling distribution of the sample means is equal to the population mean.

(i) Variance of the sample means is

$$V(\bar{X}) = \frac{\sum (X - \bar{X})^2}{k} = \frac{20}{16} = 1.25$$

OR

$$V(\bar{X}) = \frac{\sigma^2}{n} = \frac{2.5}{2} = 1.25$$

(v) The standard error of sample mean is given by

S.E. $(\bar{X}) = \sqrt{V(\bar{X})} = \sqrt{1.25} = 1.118$

Probability Proportional to Size Sampling (PPS) (With replacement)

In case of SRS, the selection probabilities are equal for all the units of the population. If the units vary considerably in size, simple random sampling (SRS) may not be appropriate procedure for selecting the units. In such situation, the unequal probability of unit selection could be a better procedure.

One of the method of selecting the units using unequal probabilities is probability proportional to size (PPS) sampling. In order to get the idea about unequal probability of selecting the units, we will take the help of auxiliary variables. It should be noted that the selection of auxiliary variable should be positively correlated to the study variate. The probability of selection may be assigned proportional to size of the units. This type of sampling scheme in which the probability of selection is proportional to the size of the units is known as probability proportional to size (PPS) sampling.

The basic difference between SRS and PPS sampling is that, in SRS the probability of drawing any specified unit at any given draw is same, while in PPS sampling it differs from draw to draw. Therefore, the PPS sampling is more complex than SRS.

Techniques of selecting a sample

Generally, two different methods are used for selecting a desirable size of the sample in case of PPS.

(i) Cumulative total method

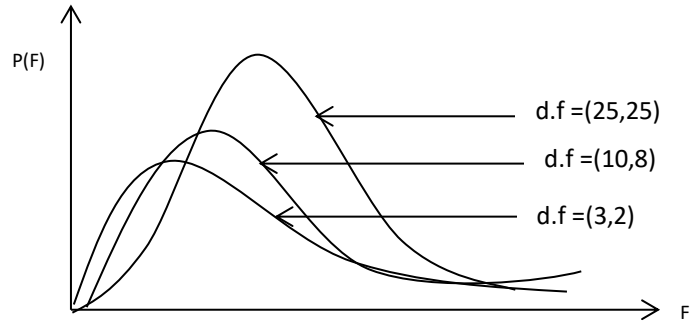
(ii) Lahiri's method

F-test (Variance ratio test)

If X and Y are two independent χ^2 −variates with $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$ degrees of freedom (d.f.) respectively, F-statistic is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and the ratio is given by

$$F = \frac{\frac{X}{\nu_1}}{\frac{Y}{\nu_2}} \sim F_{(\nu_1,\nu_2)}$$

is called F-statistic with (ν_1, ν_2) degrees of freedom (d.f.) and the distribution of F-statistic is called F-distribution. The F-statistic was developed by a great Statistician R.A Fisher. So, the distribution of F-statistic is also called Fisher's F-distribution. The sampling distribution of F-statistic does not involve any population parameters and depends only on the degrees of freedom ν_1 and ν_2.The range of values of F varies from 0 to ∞. F-distribution has a single mode. The shape of F-distribution depends on degrees of freedom in both numerator and denominator of F-ratio. However, F-distribution is generally skewed to the right and tends to become more symmetrical as the number of degrees of freedom in the numerator and denominator increase as shown in figure:



Application of F-distribution

F-distribution has a number of applications in statistics, some of which are given below:

(i) F-test for equality of two population variances.
(ii) F-test for testing the equality of several population means.
(iii) F-test for testing the significance of an observed sample multiple correlation coefficient.
(iv) F-test for testing the significance of an observed sample correlation ratio.
(v) F-test for testing the linearity of regression.

F-test (or variance ratio test) for equality of two population variances

Test statistic: Under H_o, the test statistic is

$$F = \frac{{S_1}^2}{{S_2}^2} \sim F(n_1 - 1, n_2 - 1) \text{ if } {S_1}^2 > {S_2}^2$$

$$F = \frac{{S_2}^2}{{S_1}^2} \sim F(n_2 - 1, n_1 - 1) \text{ if } {S_2}^2 > {S_1}^2$$

i.e. $F = \frac{\text{Greater variance}}{\text{Smaller variance}}$

Analysis of variance (ANOVA)

We performed Z-test (for large sample) and t-test (for small sample) for testing of hypothesis of significance of difference between two means. But if we need to test the homogeneity or equality of more than two means, z-test and t-test are inadequate. In such situation, the analysis of variance (Often abbreviated as ANOVA) is a powerful statistical tool to test the significance of difference among means of three or more than three populations. ANOVA technique is based on the F- distribution. The main objective of ANOVA technique is to test the equality or homogeneity of several population means. Therefore, it should be clearly understood that ANOVA technique is not designed to test the equality of several population variances.

The term 'Analysis of variance' was introduced by Prof. R.A. Fisher in 1920's to deal with the problem in the analysis of agronomical data. Variation is inherent in nature. The total variation in any set of numerical data is due to number of causes which may be classified as:

(i) Assignable causes & (ii) Chance causes.

The variation due to assignable causes can be detected and measured whereas the variation due to chance causes is beyond the control of human hand and cannot be determined and traced separately.

Definition. According to Prof. R.A. Fisher, Analysis of variance (ANOVA) is the "Separation of variance ascribable to one group of causes from the variance of ascribable to the other group"

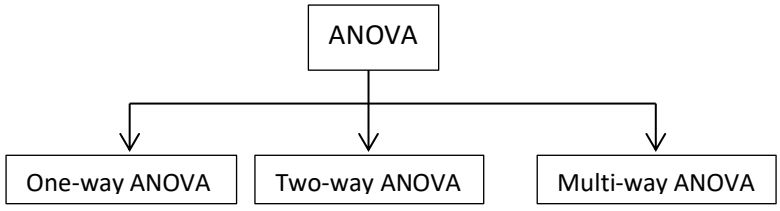
Thus, the analysis of variance (ANOVA) is a statistical technique specially designed to test whether the means of more than two quantitative populations are equal.

For example, to test the significance of difference in the average sales of three different salesmen in four districts, to find the significance of difference in the average yield of crop of five different varieties of seeds in four different plots of lands etc.

Classification of ANOVA

On the basis of number of factors (i.e. treatments) considered in the study, there are basically three types of ANOVA.

(a) One way ANOVA
(b) Two way ANOVA
(c) Multi-way ANOVA



But in this section, we will discuss only one way ANOVA and two-way ANOVA

Assumptions in ANOVA:

For the validity of the test in ANOVA following assumptions are made:

(i) All the sample observations are randomly selected and independent.
(ii) The parent population from which observations are taken is normal.
(iii) Various treatment and environmental effects (different effects) are additive in nature.
(iv) The population from which the samples are drawn have equal variances.

That is, $\sigma_1{}^2 = \sigma_2{}^2 = \sigma_3{}^2 = \ldots\ldots\ldots = \sigma_k{}^2$ for k populations.

Applications of ANOVA:

ANOVA is applicable

(i) To test homogeneity of several means (i.e. more than two means)
(ii) To test the linearity of the fitted regression line.
(iii) To test the significance of an observed sample correlation ratio.
(iv) To test the significance of observed multiple correlation coefficient.
(v) To test for the homogeneity of a group of regression coefficients in a multiple regression model.
(vi) To test the significance of equality of two population variances.

Computation of test statistic:

The technique of analysis of variance is to split up the total variation into two components variations: Between samples (or due to treatments) and within samples (or due to error). Each component gives unbiased estimate of population variance σ^2.Thus, total sum of square is broken up into: sum of squares due to independent factors and sum of squares due to random causes, called 'error'. The variance ratio is obtained by dividing the variance between the samples by the variance within samples. This ratio forms the test statistic known as F statistic.

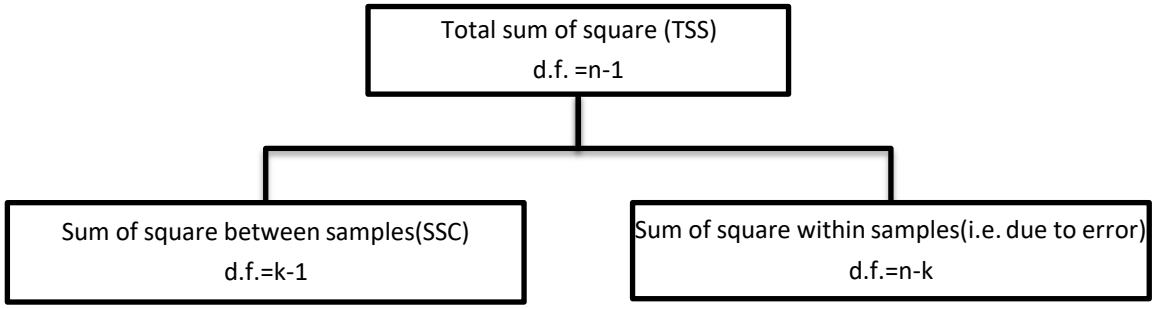
$$F_{\text{statistic}} = \frac{\text{Variance between samples}}{\text{Variance within samples}}$$

One Way ANOVA (One Factor Analysis of Variance)

In one way classification only one treatment is used in the specified Plots. That means the effect is studied on the basis of single treatment is known as one way ANOVA. In other words, under one way classification the influence of only one factor is considered at a time and we may conduct the experiment through a number of sample studies. For example, to compare the effect of fertilizers on yield of four different varieties of seeds, to compare the significance difference in prices of commodities in four cities, to compare the significance of difference in the scores of students in the four campuses, to test the significance of difference in the sales of tea in four seasons etc.

In one way ANOVA, the values of the response variable are affected (analyzed) by only one factor. The total variation or total sum of squares (TSS) among all observations can be splitted into two components:

- Sum of squares of variation between samples
- Sum of squares of variation within samples.



The following three methods are used for performing one way analysis of variance (One way ANOVA)

1. Actual mean method
2. Direct method
3. Short – cut method or Coding method

Note: In this chapter, we discuss only direct method and short-cut methods.

Direct Method

The following are the steps to be followed under direct method for one way ANOVA:

Step 1: Setting up of hypothesis

Null Hypothesis: H_0:$\mu_1 = \mu_2 = \mu_3 = \ldots\ldots = \mu_k$. That is, there is no significant difference between k population means. In other words, k population means are equal.

Alternative Hypothesis: H_1:$\mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots\ldots \neq \mu_k$. That is, there is significant difference between k population means. In other words, k populations means are not equal.

Step 2: Test statistic: Under H_0, test statistic is

$$F = \frac{\text{Mean Sum of Squares between samples}}{\text{Mean Sum of Square within samples}}$$

$$= \frac{MSC}{MSE}$$ with (k – 1) degree of freedom to MSC at numerator and (n – k) degree of freedom to MSE at denominator, i.e. it follows F – distribution with (k – 1, n – k) d.f.

Where,

MSC = Mean Sum of squares of variation between the samples.

$$= \frac{SSC}{k-1}$$

MSE = Mean Sum of squares of variation within the samples (errors).

$$= \frac{SSE}{n-k}$$

Calculation of TSS, SSC and SSE in order to find MSC and MSE

i. Find Grand Total i.e. the sum of the values of observations of all the k groups (or samples) denoted by T

Grand total (T) = $T_1 + T_2 + \ldots\ldots + T_k$

$= \sum X_1 + \sum X_2 + \ldots\ldots\ldots + \sum X_k$

ii. Find the correction factor (C.F.)

$$C.F. = \frac{T^2}{n},$$

Where n = total number of observations

$= n_1 + n_2 + \ldots\ldots + n_k$

iii. Find Row Sum of Squares (RSS)

R.S.S. = $\sum X_1^2 + \sum X_2^2 + \ldots\ldots\ldots + \sum X_k^2$

iv. Find total Sum of Squares:

SST = R.S.S. – C.F.

$= \sum X_1^2 + \sum X_2^2 + \ldots\ldots\ldots + \sum X_k^2 - C.F.$

v. Find Sum of Square of variation between the samples (SSC)

$SSC = \sum \frac{T_j^2}{n_j} - C.F.$

$= \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \ldots\ldots\ldots + \frac{T_k^2}{n_k}\right) - C.F.$

vi. Find Sum of square of variation within the samples(due to error)

SSE = SST – SSC

vii. Set up one way ANOVA table and find the calculated value of F as:

One way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F – ratio
Between the samples (Column)	SSC	k – 1	MSC = $\frac{SSC}{k-1}$	F = $\frac{MSC}{MSE}$
Within the samples (Error)	SSE	n – k	MSE = $\frac{SSE}{n-k}$	
Total	SST	n – 1	-	

Step 3: Level of significance(α) : Generally, level of significance, $\alpha = 5\%$ if it is not stated.

Step 4: Critical value: Obtain critical value (or significant value) of F as per α level of significance for (k – 1, n – k) d.f. From F –distribution table.

Step 5: Decision:

- If calculated value, $F_{cal} \leq F_{tab}$ (critical value (tabulated value), it is not significant . H_o is not rejected i.e. H_o is accepted.
- If calculated value, $F_{cal} > F_{tab,}$ critical value (tabulated value), , it is significant. H_o is rejected i.e. H_1 is accepted

Example 1 A certain medicine distributor wants to test whether three of his MR (Medical Representative), A, B and C in a given region make require amount of customer of his products during a given period of time. A record of four months showed the following results for the number of customer made by each MR for each month.

Month	Medical representative (MR)		
	A	B	C
1	7	5	13
2	8	7	11
3	10	9	17
4	11	3	7

Test at 95% confidence level; is there any significant difference in the average number of customers made by the three MR per month?

Solution: Since, in this problem the number of customers made is analyzed by using only one factor (i.e. due to medical representative), therefore this is the case of one way ANOVA.

Null Hypothesis: $H_0: \mu_A = \mu_B = \mu_C$. That is, there is no significant difference in the average number of customers made by three MR per month.

Alternative Hypothesis: $H_1: \mu_A \neq \mu_B \neq \mu_C$ (Two tailed test). That is, there is significant difference in the average number of customers made by three MR per month.

Test statistic: Under H_0, for one way ANOVA the test statistic is

$$F = \frac{MSC}{MSE}$$

Where,

MSC = Mean sum of squares of variation between medical representatives

$$= \frac{SSC}{k-1}$$

MSE = Mean sum of squares of variation within medical representatives (due to errors)

$$= \frac{SSE}{n-k}$$

Calculation of SST, SSC and SSE

Month	Medical Representative (MR)					
	A (X_1)	B (X_2)	C (X_3)	X_1^2	X_2^2	X_3^2
1	7	5	13	49	25	169
2	8	7	11	64	49	121
3	10	9	17	100	81	289
4	11	3	7	121	9	49
T_j	$T_1 = \sum X_1 = 36$	$T_2 = \sum X_2 = 24$	$T_3 = \sum X_3 = 48$	$\sum X_1^2 = 334$	$\sum X_2^2 = 164$	$\sum X_3^2 = 628$

Grand Total (T) = $T_1 + T_2 + T_3 = \sum X_1 + \sum X_2 + \sum X_3 = 36 + 24 + 48 = 108$

Correction Factor (C.F.) = $\frac{T^2}{n} = \frac{(108)^2}{12} = 972$

SST = Total sum of squares of variation

$= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C.F.$

$= 334 + 164 + 628 - 972$ Where, R.S.S. = $\sum X_1^2 + \sum X_2^2 + \sum X_3^2$

$= 154$

SSC = Total sum of squares of variation between medical representatives

$$= \sum \frac{T_j^2}{n_j} - C.F. = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3}\right) - C.F.$$

$$= \left(\frac{36^2}{4} + \frac{24^2}{4} + \frac{48^2}{4}\right) - 972 = 72$$

SSE = Sum of squares of variation within medical representatives (due to errors)

= SST – SSC = 154 – 72 = 82

One way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F – ratio
Between the samples (Column)	SSC = 72	k – 1 = 3 – 1 = 2	MSC = $\frac{SSC}{k-1} = \frac{72}{2}$ = 36	F = $\frac{MSC}{MSE}$ = $\frac{36}{9.11}$ F_{cal} = 3.95
Within the samples (Error)	SSE = 82	n – k = 12 – 3 = 9	MSE = $\frac{SSE}{n-k} = \frac{82}{9} = 9.11$	
Total	SST = 154	n – 1 = 12 – 1 = 11	-	

From ANOVA table,

F_{cal} = Calculated value of F (2, 9) = 3.95

Level of significance(α) = 5% (assume)

Critical value: Tabulated value of F at 5% level of significance for (2, 9) d.f. is 4.26 i.e. $F_{tab} = F_{0.05}$ (2,9) = 4.26

Decision: Since, calculated value of F is less than the tabulated value of F i.e. $F_{cal} < F_{tab}$, it is not significant and hence H_0 is accepted which means that there is no significant difference in the average number of customers made by three MR per month.

Example 2

The following table represents the sales of three salesmen in four different districts.

Districts	Sales persons (Sales figure in '000')		
	A	B	C
Nepalgunj	12	16	15
Surkhet	10	22	14
Kathmandu	9	18	8
Birgunj	11	20	10

Test whether there is any significant difference in the sales of different districts.

Solution: Since, in this problem the sales is analyzed by using only one factor (i.e. due to four different districts), therefore this is the case of one way ANOVA.

Null Hypothesis: H_0:$\mu_1 = \mu_2 = \mu_3 = \mu_4$. That is, there is no significant difference in the sales of four different districts.

Alternative Hypothesis: H_1: $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$.That is, there is significant difference in the sales of four different districts.

Test statistic: Under H_0, for one way ANOVA the test statistic is

$$F = \frac{MSC}{MSE}$$

Where,

MSC = Mean sum of squares of variation between districts

$= \frac{SSC}{k-1}$

MSE = Mean sum of squares of variation within districts (due to errors)

$= \frac{SSE}{n-k}$

Calculation of SST, SSC and SSE

Sales persons (Sales figure in'000')	Districts				X_1^2	X_2^2	X_3^2	X_4^2
	Nepalgunj (X_1)	Surkhet (X_2)	Kathmandu (X_3)	Birgunj (X_4)				
A	12	10	9	11	144	100	81	121
B	16	22	18	20	256	484	324	400
C	15	14	8	10	225	196	64	100
T_j	$T_1 = \sum X_1 = 43$	$T_2 = \sum X_2 = 46$	$T_3 = \sum X_3 = 35$	$T_4 = \sum X = 41$	$\sum X_1^2 = 625$	$\sum X_2^2 = 780$	$\sum X_3^2 = 469$	$\sum X_4^2 = 621$

Grand Total (T) = $T_1 + T_2 + T_3 + T_4 = \sum X_1 + \sum X_2 + \sum X_3 + \sum X_4 = 43 + 46 + 35 + 41 = 165$

Correction Factor (C.F.) = $\frac{T^2}{n} = \frac{(165)^2}{12} = 2268.75$

SST = Total sum of squares of variation

$= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - C.F.$ Where, R.S.S. $= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2$

$= 625+780+469+621 - 2268.75$

$= 226.25$

SSC = Total sum of squares of variation between districts

$= \sum \frac{T_j^2}{n_j} - C.F. = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4}\right) - C.F.$

$= \left(\frac{43^2}{3} + \frac{46^2}{3} + \frac{35^2}{3} + \frac{41^2}{3}\right) - 2268.75$

$= 2290.333 - 2268.75 = 21.583$

SSE = Sum of squares of variation within districts (due to errors)

= SST – SSC = 226.25 – 21.583 = 204.667

One way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F – ratio
Between districts	SSC = 21.583	k – 1 = 4 – 1 = 3	MSC $= \frac{SSC}{k-1} = \frac{21.583}{3} = 7.194$	$F = \frac{MSC}{MSE}$ $= \frac{7.194}{25.583}$ $F_{cal} = 0.281$
Within districts (Error)	SSE= 204.667	n – k = 12 – 4 = 8	MSE $= \frac{SSE}{n-k} = \frac{204.667}{8} =$ 25.583	
Total	SST = 226.25	n – 1 = 12 – 1 = 11	-	

From ANOVA table,

F_{cal} = Calculated value of F (3, 8) = 0.281

Level of significance(α) = 5% (assume)

Critical value: Tabulated value (critical value) of F at 5% level of significance for (3, 8) d.f. is 4.07 i.e. $F_{tab} = F_{0.05}$ (3, 8) = 4.07

Decision: Since, calculated value of F is less than the tabulated value of F i.e. $F_{cal} < F_{tab}$, it is not significant. Therefore, H_0 is accepted which means that there is no significant difference in the sales of four different districts.

Short- cut Method or Coding Method

If the magnitude of given observations are considerably large, then direct method also becomes time consuming and tedious at the time of calculations. To overcome this difficulty, we use short-cut method or coding method. This method refers the addition, subtraction, multiplication or division of the observations by a constant or common factor. It does not require any subsequent adjustment in the results after coding. Generally, subtraction or division by a common factor is done in order to get smaller values so that calculation becomes easier. After that, the remaining procedure has been done in the same way as direct method.

Example 3

The following table reveals the life of certain four batch of electric lamp in hours:

Batch I	1500	1510	1540	1550	1600	1650	1700
Batch II	1440	1450	1420	1520	1650		
Batch III	1400	1540	1510	1700	1750	1800	
Batch IV	1520	1560	1530	1540			

Setup analysis of variance to these data and test at 1% level of significance is there any significant difference in life of batch of electric lamps.

Solution:

Null Hypothesis, H_0:$\mu_1 = \mu_2 = \mu_3 = \mu_4$. That is , there is no significant difference in the life of four batch of electric lamps.

Alternative Hypothesis, H_1:$\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$. That is, there is significant difference in the life of four batch of electric lamps.

Test statistic: Under H_0, for one way ANOVA the test statistic is

$$F = \frac{MSC}{MSE}$$

Where,

MSC = Mean sum of squares of variation between batches

$$= \frac{SSC}{k-1}$$

MSE = Mean sum of squares of variation within batches (due to errors)

$$= \frac{SSE}{n-k}$$

Calculation of SST, SSC and SSE in order to find MSC and MSE

Now, the data can be coded by subtracting 1500 and dividing by 10, and then the above data reduced as follows:

Batch I (X_1)	Batch II (X_2)	Batch III (X_3)	Batch IV (X_4)	X_1^2	X_2^2	X_3^2	X_4^2
0	-6	-10	2	0	36	100	4
1	-5	4	6	1	25	16	36
4	-8	1	3	16	64	1	9
5	2	20	4	25	4	400	16
10	15	25		100	225	625	
15		30		225		900	
20				400			
$T_1 = \sum X_1 = 55$	$T_2 = \sum X_2 = -2$	$T_3 = \sum X_3 = 70$	$T_4 = \sum X_4 = 15$	$\sum X_1^2 = 767$	$\sum X_2^2 = 354$	$\sum X_3^2 = 2042$	$\sum X_4^2 = 65$

Grand Total (T) = $T_1 + T_2 + T_3 + T_4 = \sum X_1 + \sum X_2 + \sum X_3 + \sum X_4 = 55 - 2 + 70 + 15 = 138$

Correction Factor (C.F.) = $\frac{T^2}{n} = \frac{(138)^2}{22} = 865.64$

SST = Total sum of squares of variation

$= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2 - C.F$ Where, R.S.S. = $\sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2$

$= 767 + 354 + 2042 + 65 - 865.64$

$= 3228 - 865.64 = 2362.36$

SSC = Sum of squares of variation between batches

$= \sum \frac{T_j^2}{n_j} - C.F. = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4}\right) - C.F.$

$= \left(\frac{55^2}{7} + \frac{(-2)^2}{5} + \frac{70^2}{6} + \frac{15^2}{4}\right) - 865.64$

$= 440.22$

SSE = Sum of squares of variation within batches (due to errors)

$= SST - SSC = 2362.36 - 440.22 = 1922.14$

One way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F - ratio
Between the samples (Column)	SSC = 440.22	$k - 1 = 4 - 1 = 3$	MSC = $\frac{SSC}{k-1} = \frac{440.22}{3} = 146.74$	$F = \frac{MSC}{MSE} = \frac{146.74}{106.79}$ $F_{cal} = 1.374$
Within the samples (Error)	SSE = 1922.14	$n - k = 22 - 4 = 18$	MSE = $\frac{SSE}{n-k} = \frac{1922.14}{18} = 106.79$	
Total	SST = 2362.36	$n - 1 = 22 - 1 = 21$	-	

From ANOVA table,

F_{cal} = Calculated value of F (3, 18) = 1.374

Level of significance(α) = 1%

Critical value: Tabulated value (critical value) of F at 1% level of significance for (3, 18) d.f. is 5.09 i.e. $F_{tab} = F_{0.01}$ (3, 18) = 5.09

Decision: Since, calculated value of F is less than the tabulated value of F i.e. $F_{cal} < F_{tab}$, it is not significant. H_0 is accepted .Therefore; there is no significant difference in the life of four batches of electric lamps.

Example 4

Final grades obtained by three students in an advanced course of MBA are as follows:

A	3.2	3.1	2.9	2.7	3.3		
B	2.7	3.5	3.6	2.8	2.9	2.7	
C	2.9	3.7	3.7	3.5	3.4	3.5	3.6

Does the difference in sample means appear to be due to chance variation or is there sufficient evidence to indicate that not all population means are the same? Use 5% level of significance.

Solution:

Null Hypothesis: H_0:$\mu_A = \mu_B = \mu_C$.That is, there is no significant difference between all population means. In other words, all population means are same.

Alternative Hypothesis: H_1: $\mu_A \neq \mu_B \neq \mu_C$. That is, there is significant difference between all population means. In other words, all population means are not same.

Test statistic: Under H_0, for one way ANOVA, the test statistic is

$$F = \frac{MSC}{MSE}$$

Where,

MSC = Mean sum of squares of variation between students.

$$= \frac{SSC}{k-1}$$

MSE = Mean sum of squares of variation within students (due to errors)

$$= \frac{SSE}{n-k}$$

Calculation of SST, SSC and SSE in order to find MSC and MSE

Now, for easier calculation, the data can be coded by multiplying each observation by 10

A (X_1)	B (X_2)	C (X_3)	X_1^2	X_2^2	X_3^2
32	27	29	1024	729	841
31	35	37	961	1225	1369
29	36	37	841	1296	1369
27	28	35	729	784	1225
33	29	34	1089	841	1156
	27	35		729	1225
		36			1296
$T_1 = \sum X_1$ = 152	$T_2 = \sum X_2$ = 182	$T_3 = \sum X_3$ = 243	$\sum X_1^2$ = 4644	$\sum X_2^2$ = 5604	$\sum X_3^2$ = 8481

Grand Total (T) = $T_1 + T_2 + T_3 = \sum X_1 + \sum X_2 + \sum X_3 = 152 + 182 + 243 = 577$

Correction Factor (C.F.) = $\frac{T^2}{n} = \frac{(577)^2}{18}$ = 18496.06

SST = Total sum of squares of variation

= $\sum X_1^2 + \sum X_2^2 + \sum X_3^2$ – C. F Where, R.S.S. = $\sum X_1^2 + \sum X_2^2 + \sum X_3^2$

= 4644 + 5604 + 8481–18496.06

= 18729 – 18496.06 = 232.94

SSC = Sum of squares of variation between students

= $\sum \frac{T_j^2}{n_j}$ - C.F. = $\left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3}\right) - C.F.$

= $\left(\frac{152^2}{5} + \frac{182^2}{6} + \frac{243^2}{7}\right) - 18496.06 = 80.98$

SSE = Sum of squares of variation within students (due to errors)

= SST – SSC = 232.94 – 80.98=151.96

One way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F - ratio
Between the students (Column)	SSC = 80.98	k – 1 =3 – 1 = 2	MSC = $\frac{SSC}{k-1} = \frac{80.98}{2}$ = 40.49	F = $\frac{MSC}{MSE}$ = $\frac{40.49}{10.13}$ F_{cal} = 3.997
Within the students (Error)	SSE = 151.96	n – k = 18 – 3= 15	MSE = $\frac{SSE}{n-k} = \frac{151.96}{15}$ = 10.13	
Total	SST = 232.94	n – 1 = 18 – 1 = 17		

From ANOVA table,

F_{cal} = Calculated value of F (2, 15) = 3.997

Level of significance(α) = 5%

Critical value: Tabulated value (critical value) of F at 5% level of significance for (2, 15) d.f. is 3.68 i.e. F_{tab} = $F_{0.05}$ (2, 15) = 3.68

Decision: Since, calculated value of F is greater than the tabulated value of F i.e. $F_{cal} > F_{tab}$, it is significant. H_0 is rejected and hence H_1 is accepted which means that there is significant difference between all population means. In other words, all population means are not same.

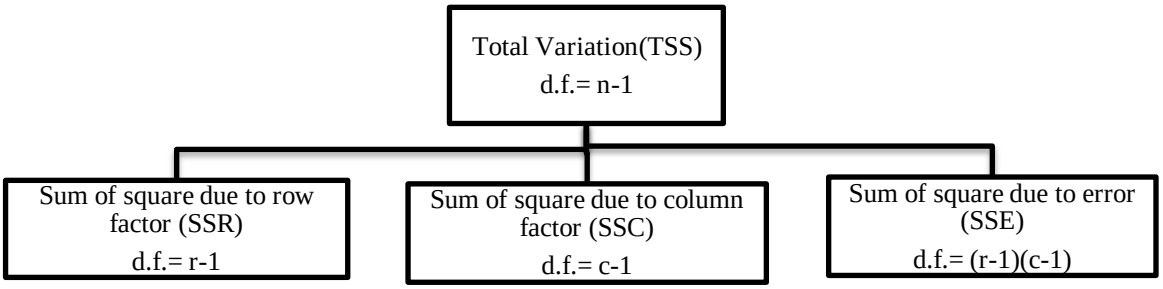
Two way ANOVA (Two Factors Analysis of Variance)

In the two way classification, the whole data is classified on the basis of two factors and test for the effect of two factors simultaneously. Such a test is called two way ANOVA. For example, sales of a product may be affected by different salesmen as well as different seasons, the yield of milk may be affected by rations as well as varieties of cows, to compare the significance difference in the yield of wheat due to fertilizers as well as variety of seeds, production of number of units may be affected by different workmen as well as different types of machines etc. All these factors are studied under a two factor analysis of variance i.e. Two way ANOVA. Thus, two way ANOVA is used to test the effect of two factors simultaneously on the values of response variable of our interest.

In two way ANOVA, the values of the response variable are affected (analyzed) by two factors simultaneously. In two way ANOVA, the total variation or total sum of squares (TSS) among all observations can be partioned into three components:

- Sum of square due to row factor or sum of squares of variation in rows (SSR)
- Sum of square due to column factor or sum of squares of variation in columns (SSC)
- Sum of squares of variation due to errors or residual (SSE).

Thus, SST = SSR+SSC+SSE



The steps in testing of hypothesis under two way ANOVA are as follows:

Step 1: Formulation of Hypothesis:

(i) For column factor:

Null hypothesis, $H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_c$ i.e. there is no significant difference in the c population means due to column factor. In other words, c populations means are equal.

Alternative hypothesis, $H_1 : \mu_1 \neq \mu_2 \neq \ldots\ldots \neq \mu_c$ i.e. there is significant difference in the c population means due to column factor. In other words, c populations means are not equal.

(ii) For row factor:

Null hypothesis, $H_0 : \mu_1 = \mu_2 = \ldots\ldots = \mu_r$ i.e. there is no significant difference in the r population means due to row factor. In other words, r population means are equal.

Alternative hypothesis, $H_1 : \mu_1 \neq \mu_2 \neq \ldots\ldots \neq \mu_r$ i.e. there is significant difference in the r population means due to row factor. In other words, r population means are not equal.

Step 2: Compute test statistic: Under H_0, test statistics for two way ANOVA are as follows:

$$F_c = \frac{MSC}{MSE}, \text{ with d.f. } [(c-1), (c-1)(r-1)]$$

$$F_r = \frac{MSR}{MSE}, \text{ with d.f. } [(r-1), (c-1)(r-1)]$$

Where, MSC = Mean sum of squares of variation between columns

MSR = Mean sum of squares of variation between rows

MSE = Mean sum of squares of variation due to error or residual

Calculation of TSS, SSC, SSR and SSE in order to find MSC, MSR and MSE

i. Find the sum of the values of observations (Grand Total) denoted by T:
$T = \sum T_i = \sum T_j$

ii. Find the correction factor (C.F.)
$C.F. = \frac{T^2}{n}$,
where $n = r \times c$

iii. Find Row Sum of Squares (RSS)
$R.S.S. = \sum X_{r1}{}^2 + \sum X_{r2}{}^2 + \dots\dots\dots + \sum X_{rc}{}^2$

iv. Find total Sum of Squares:

$SST = R.S.S. - C.F.$

v. Find Sum of Square of variation between column (SSC)
$SSC = \sum \frac{T_j^2}{n_j} - C.F. = \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \dots\dots\dots + \frac{T_c^2}{n_c}\right) - C.F.$
Where, T_j = each j^{th} column total
n_j = number of elements in j^{th} column

vi. Find Sum of Square of variation between row (SSR)
$SSR = \sum \frac{T_i^2}{n_i} - C.F.$
$= \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \dots\dots\dots + \frac{T_r^2}{n_r}\right) - C.F.$
Where, T_i = each ith row total
n_i = number of elements in i^{th} row

vii. Find Sum of square of variation due to error
$SSE = SST - SSC - SSR$

viii. Set up two way ANOVA table and find the calculated value of F as:

Two way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F – ratio
Between Column	SSC	c – 1	$MSC = \frac{SSC}{c-1}$	$F_c = \frac{MSC}{MSE}$
Between Row	SSR	r – 1	$MSR = \frac{SSR}{r-1}$	$F_r = \frac{MSR}{MSE}$
Residual or Error	SSE	(r – 1)(c – 1)	$MSE = \frac{SSE}{(r-1)(c-1)}$	
Total	SST	n – 1 = rc – 1		

Step 3: Level of significance(α) : Generally, level of significance, $\alpha = 5\%$ if it is not stated.

Step 4: Critical value: Obtain critical value (or significant value) of F as per α level of significance from F –distribution table at d.f. for:

a. Column factor: d.f. = [(c – 1), (c – 1) (r – 1)]. i.e. $F_\alpha[(c-1),(c-1)(r-1)]$
b. Row factor: d.f. = [(r – 1), (c – 1) (r – 1)]. i.e. $F_\alpha[(r-1),(c-1)(r-1)]$

Step 5: Decision:

a. If calculated value of F $\leq$ tabulated value of F, it is not significant and H_0 is accepted.
b. If calculated value of F > tabulated value of F, it is significant and H_1 is accepted.

Example 5

A tea company appoints four salesmen A, B, C and D, and observes their sales in three seasons- summer, winter and monsoon. The figures are given below

Season	Salesmen				
	A	B	C	D	Total
Summer	36	36	21	35	128
Winter	28	29	31	32	120
Monsoon	26	28	29	29	112
Total	90	93	81	96	360

Find out if there is difference in the sales of different salesman and seasons.

Solution: Since, there are two factors (Salesmen and seasons) affecting the sales i.e., sales are analyzed by using two factors Salesmen and seasons, therefore this is the case of two way ANOVA.

Null hypothesis:

(i) H_0: $\mu_A = \mu_B = \mu_C = \mu_D$.That is, there is no significant difference between the average sales made by four salesmen.
(ii) H_0: $\mu_S = \mu_W = \mu_M$.That is, there is no significant difference between the average sales made during three seasons.

Alternative hypothesis:

(i) H_1: $\mu_A \neq \mu_B \neq \mu_C \neq \mu_D$ That is, there is significant difference between the average sales made by four salesmen.
(ii) H_1: $\mu_S \neq \mu_W \neq \mu_M$. That is, there is significant difference between the average sales made during three seasons.

Test statistic: Under H_0, for two way ANOVA, the test statistics are as follows:

$$F_c = \frac{MSC}{MSE}, \text{ with d.f. } [(c-1), (c-1)(r-1)$$
$$F_r = \frac{MSR}{MSE}, \text{ with d.f. } [(r-1), (c-1)(r-1)]$$

Where, MSC = Mean sum of squares of variation between salesman = $\frac{SSC}{C-1}$

MSR = Mean sum of squares of variation between different season = $\frac{SSR}{r-1}$

MSE = Mean sum of squares of variation due to error = $\frac{SSE}{(c-1)(r-1)}$

Calculation of TSS, SSC, SSR and SSE in order to find MSC, MSR and MSE

Now, the given data are coded by subtracting 30 from each figure, we get

Season	A (X_1)	B (X_2)	C (X_3)	D (X_4)	Season's (Row) total (T_i)	X_1^2	X_2^2	X_3^2	X_4^2
Summer	6	6	-9	5	8	36	36	81	25
Winter	-2	-1	1	2	0	4	1	1	4
Monsoon	-4	-2	-1	-1	-8	16	4	1	1
Salesman's (column) total (T_j)	$\sum X_1 = 0$	$\sum X_2 = 3$	$\sum X_3 = -9$	$\sum X_4 = 6$	$T = \sum T_i = \sum T_j = 0$	$\sum X_1^2 = 56$	$\sum X_2^2 = 41$	$\sum X_3^2 = 83$	$\sum X_4^2 = 30$

n = 12

Grand total (T) = $\sum T_j = \sum T_i = 0$

Correction factor (C.F.) = $\frac{T^2}{n} = \frac{0^2}{12} = 0$

SST = Total sum of squares of variation

$= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 + \sum X_4^2$ - C.F.

$= 56 + 41 + 83 + 30 - 0$

$= 210$

SSC = Sum of squares of variation between salesmen

$= \sum \frac{T_j^2}{n_j}$ - C.F.

$= \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4}\right) - C.F.$

$= \left(\frac{0^2}{3} + \frac{3^2}{3} + \frac{(-9)^2}{3} + \frac{6^2}{3}\right) - 0 = 42$

SSR = Sum of squares of variation between seasons

$= \sum \frac{T_i^2}{n_i}$ - C.F.

$= \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3}\right) - C.F.$

$= \left(\frac{8^2}{4} + \frac{0^2}{4} + \frac{(-8)^2}{4}\right) - 0 = 32$

SSE = Sum of squares of variation due to errors or residuals

$=$ SST – SSC – SSR $= 210 - 42 - 32 = 136$

Two way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F – ratio
Between salesman	SSC =42	c – 1 = 4 – 1 = 3	MSC = $\frac{SSC}{c-1} = \frac{42}{3}$ =14	$F_c = \frac{MSC}{MSE}$ = $\frac{14}{22.67}$ = 0.62
Between seasons	SSR = 32	r – 1 = 3 – 1 = 2	MSR = $\frac{SSR}{r-1} = \frac{32}{2}$ = 16	$F_r = \frac{MSR}{MSE}$ = $\frac{16}{22.67}$ = 0.71
Residual or Error	SSE = 136	(r – 1)(c – 1) = 2× 3 = 6	MSE = $\frac{SSE}{(r-1)(c-1)}$ = $\frac{136}{6}$ =22.67	
Total	SST = 210	n – 1 = rc – 1 =3× 4 - 1 = 11	-	

From ANOVA table:

Calculated values: For salesmen: $F_{cal} = 0.62$

For seasons: $F_{cal} = 0.71$

Level of significance(α) = 5% (Assume)

Critical value: For salesman: tabulated, $F_{tab} = F_{0.05}(3,6) = 4.76$

For season: tabulated, $F_{tab} = F_{0.05}(2, 6) = 5.14$

Conclusion:

For Salesman: Since, the calculated value of F is less than the tabulated value of F, i.e. $F_{cal} < F_{tab}$, it is not significant and H_0 is accepted. Therefore, there is no significant difference between the average sales made by four salesmen.

For Season: Since, the calculated value of F is less than the tabulated value of F, i.e. $F_{cal} < F_{tab}$, it is not significant and H_0 is accepted. Therefore, there is no significant difference between the average sales made during three seasons.

Example 6

The following table represents the production of three varieties of wheat, obtained with four different kinds of fertilizers

Fertilizers	Varieties of wheat		
	X	Y	Z
A	8	3	7
B	10	4	8

C	6	5	6
D	8	4	7

Test at 5% level of significance whether there is significance difference in average production (i) due to variety of wheat (ii) due to fertilizers

Solution: Since, productions are analyzed by using two factors varieties of wheat and fertilizes, therefore this is the case of two way ANOVA.

Null hypothesis:

(i) H_0: $\mu_A = \mu_B = \mu_C = \mu_D$. That is, there is no significant difference in average productions due to fertilizer.

(ii) H_0: $\mu_X = \mu_Y = \mu_Z$ That is , there is no significant difference in average productions due to variety of wheat

Alternative hypothesis:

(iii) H_0: $\mu_A \neq \mu_B \neq \mu_C \neq \mu_D$. That is, there is significant difference in average productions due to fertilizer.

(iv) H_0: $\mu_X \neq \mu_Y \neq \mu_Z$(Two tailed test).That is, there is significant difference in average productions due to variety of wheat.

Test statistic: Under H_0, for two way ANOVA test statistics are as follows:

$F_c = \frac{MSC}{MSE}$, with d.f. [(c – 1), (c – 1)(r – 1)

$F_r = \frac{MSR}{MSE}$, with d.f. [(r – 1), (c – 1)(r – 1)]

Where, MSC = Mean sum of squares of variation between variety of wheat = $\frac{SSC}{C-1}$

MSR = Mean sum of squares of variation between fertilizers = $\frac{SSR}{r-1}$

MSE = Mean sum of squares of variation due to error = $\frac{SSE}{(c-1)(r-1)}$

Calculation of TSS, SSC, SSR and SSE in order to find MSC, MSR and MSE

Now the data are coded by subtracting 5 from each figure, we get

Calculation of MSC, MSR and MSE

Fertilizers	X (X_1)	Y (X_2)	Z (X_3)	Row total (T_i)	X_1^2	X_2^2	X_3^2
A	3	-2	2	3	9	4	4
B	5	-1	3	7	25	1	9
C	1	0	1	2	1	0	1
D	3	-1	2	4	9	1	4

column total (T_j)	$\sum X_1$ = 12	$\sum X_2$ = -4	$\sum X_3$ = 8	T = 16	$\sum X_1^2$ = 44	$\sum X_2^2$ = 6	$\sum X_3^2$ = 18

n = 12

Grand total (T) = $\sum T_j = \sum T_i = 16$ and

Correction factor (C.F.) = $\frac{T^2}{n} = \frac{16^2}{12} = 21.33$

SST = Total sum of squares of variation

$= \sum X_1^2 + \sum X_2^2 + \sum X_3^2 - C.F.$

$= 44 + 6 + 18 - 21.33 = 68 - 21.33$

$= 46.67$

SSC = Sum of squares of variation between variety of wheat

$= \sum \frac{T_j^2}{n_j} - C.F.$

$= \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3}\right) - C.F.$

$= \left(\frac{12^2}{4} + \frac{(-4)^2}{4} + \frac{8^2}{4}\right) - 21.33 = 34.67$

SSR = Sum of squares of variation between fertilizers

$= \sum \frac{T_i^2}{n_i} - C.F.$

$= \left(\frac{T_1^2}{n_1} + \frac{T_2^2}{n_2} + \frac{T_3^2}{n_3} + \frac{T_4^2}{n_4}\right) - C.F.$

$= \left(\frac{3^2}{3} + \frac{7^2}{3} + \frac{2^2}{3} + \frac{4^2}{3}\right) - 21.33 = 4.67$

SSE = Sum of squares of variation due to errors or residuals

= SST – SSC – SSR = 46.67 – 34.67 – 4.67 = 7.33

Two way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F – ratio
Between variety of wheat	SSC = 34.67	c – 1 = 3 – 1 = 2	MSC = $\frac{SSC}{c-1} = \frac{34.67}{2}$ = 17.335	$F_c = \frac{MSC}{MSE}$ = $\frac{17.335}{1.22}$ = 14.21
Between fertilizers	SSR = 4.67	r – 1 = 4 – 1 = 3	MSR = $\frac{SSR}{r-1} = \frac{4.67}{3}$ = 1.56	

Residual or Error	SSE = 7.33	(r – 1)(c – 1) = 3× 2 = 6	MSE = $\frac{SSE}{(r-1)(c-1)}$ = $\frac{7.33}{6}$ = 1.22	$F_r = \frac{MSR}{MSE}$ = $\frac{1.56}{1.22}$ = 1.28
Total	SST = 46.67	n – 1 = rc – 1 =4× 3 - 1 = 11		

From ANOVA table:

Calculated values: For variety of wheat: F_{cal} = 14.21

For fertilizers: F_{cal} = 1.28

Level of significance(α) = 5%

Critical value: For variety of wheat: tabulated, $F_{tab} = F_{0.05}(2,6) = 5.14$

For Fertilizers: tabulated, $F_{tab} = F_{0.05}(3, 6) = 4.76$

Conclusion:

For variety of wheat: Since, the calculated value of F is greater than the tabulated value of F, i.e. $F_{cal} > F_{tab}$, it is significant.H_0 is rejected and hence H_1 is accepted. Therefore, there is significant difference in the average productions due to variety of wheat.

For Fertilizer: Since, the calculated value of F is less than the tabulated value of F, i.e. $F_{cal} < F_{tab}$, it is not significant and H_0 is accepted. Therefore, there is no significant difference in the average productions due to fertilizer.

Example 7

Prepare two way ANOVA table and carry out the test for the significance difference in the average sales due to:

i. 4 salesmen
ii. 3 different cities, where the following information are given:
SST (Total sum of square) = 62
SSC (Sum of square between salesmen) = 6
SSR (Sum of square between cities) = 32

Solution:

Since, there are two factors (Salesmen and cities) affecting the sales i.e., sales are analyzed by using two factors Salesman and cities, therefore this is the case of two way ANOVA.

Null hypothesis:

(i) H_0: $\mu_A = \mu_B = \mu_C = \mu_D$.That is, there is no significant difference between the average sales due to 4 salesmen.

(ii) H_0: $\mu_X = \mu_Y = \mu_Z$: That is, there is no significant difference between the average sales due to 3 different cities.

Alternative hypothesis:

(iii) H_0: $\mu_A \neq \mu_B \neq \mu_C \neq \mu_D$.That is, there is significant difference between the average sales due to 4 salesmen.

(iv) H_0: $\mu_X \neq \mu_Y \neq \mu_Z$. That is, there is significant difference between the average sales due to 3 different cities.

Test statistic: Under H_0, for two way ANOVA test statistics are as follows:

$$F_c = \frac{MSC}{MSE}, \text{ with d.f. } [(c-1), (c-1)(r-1)$$

$$F_r = \frac{MSR}{MSE}, \text{ with d.f. } [(r-1), (c-1)(r-1)]$$

Where, MSC = Mean sum of squares of variation between average sales due to salesmen = $\frac{SSC}{C-1}$

MSR = Mean sum of squares of variation between sales due to different cities. = $\frac{SSR}{r-1}$

MSE = Mean sum of squares of variation due to error = $\frac{SSE}{(c-1)(r-1)}$

In order to find MSC, MSR and MSE, first we need to find SST, SSC, SSR and SSE

Given:

SST = 62, SSC = 6, SSR = 32

SSE = SST – SSC – SSR = 62 – 6 – 32 = 24

Two way ANOVA table

Source of variation	Sum of square (S.S.)	Degree of freedom (d.f.)	Mean Sum of Square (M.S.S.)	F – ratio
Between salesmen (Column)	SSC = 6	c – 1 = 4 – 1 = 3	MSC = $\frac{SSC}{c-1} = \frac{6}{3}$ = 2	$F_c = \frac{MSC}{MSE}$ $= \frac{2}{4}$ = 0.5
Between cities (Row)	SSR = 32	r – 1 = 3 – 1 = 2	MSR = $\frac{SSR}{r-1} = \frac{32}{2}$ = 16	$F_r = \frac{MSR}{MSE}$ $= \frac{16}{4}$ = 4
Residual or Error	SSE = 24	(r – 1)(c – 1) = 2× 3 = 6	MSE = $\frac{SSE}{(r-1)(c-1)}$ $= \frac{24}{6} = 4$	
Total	SST = 46.67	n – 1 = rc – 1 =3× 4 - 1 = 11		

From ANOVA table:

Calculated values: For salesmen: $F_{cal} = 0.5$

For cities: $F_{cal} = 4$

Level of significance(α) = 5%

Critical value: For salesmen: tabulated, $F_{tab} = F_{0.05}(3,6) = 4.76$

For cities: tabulated, $F_{tab} = F_{0.05}(2, 6) = 5.14$

Conclusion:

For salesmen: Since, the calculated value of F is less than the tabulated value of F, i.e. $F_{cal} < F_{tab}$, it is not significant and H_0 is accepted. Therefore, there is no significant difference between the average sales due to 4 salesmen.

For cities: Since, the calculated value of F is less than the tabulated value of F, i.e. $F_{cal} < F_{tab}$, it is not significant and H_0 is accepted. Therefore, there is no significant difference between the average sales due to 3 different cities.