MBA				**2019-2020**
				Semester – I
19MBAP106	**STATISTICS FOR DECISION MAKING**			**5H – 4C**
InstructionHours/week L:4T:1 P:0		**Marks:Internal:40**	**External:60**	**Total:100**
				End Semester Exam: 3 Hours

COURSE OBJECTIVES:
To make the students

1. To understand the classification and analysis of the data with statistical tools and techniques.
2. To know the descriptive and inferential statistics, and apply them to examine business and economic data.
3. To realize the applications of probability and distributions in the analytical decision making.
4. To conduct statistical estimation and hypothesis testing with statistical tools and techniques.
5. To understand the index number concepts and its applications.

COURSE OUTCOMES :
Learners should be able to

1. Understand the basic statistical tools and techniques and its application in business decision making.
2. Perform basic statistical estimation and hypothesis testing for interpret the results.
3. Know how to specify, estimate, and use statistical models to predict and obtain reliable forecasts.
4. Develop an ability to analyse and interpret the collected data to provide meaningful information in making management decisions
5. Demonstrate capabilities of problem-solving, critical thinking, and communication skills related to the discipline of statistics.

UNIT I Data and presentation of Data
Introduction to Statistics: Introduction to Statistics, Importance of Statistics in modern business environment. Classification, Tabulation and Presentation of Data: Introduction, Functions of Classification - Requisites of a good classification - Types of classification - Methods of classification, Tabulation - Basic difference between classification and tabulation -Parts of a table -Types of table , Frequency and Frequency Distribution - Derived frequency distributions - Bivariate and multivariate frequency distribution - Construction of frequency distribution, Presentation of Data – Diagrams, Graphical Presentation - Histogram - Frequency polygon - Frequency curve - Ogives

UNIT II Measures of Central Tendency and Dispersion

Measures of Central Tendency and Dispersion: Introduction, Objectives of statistical average, Requisites of a Good Average, Statistical Averages - Arithmetic mean - Properties of arithmetic mean - Merits and demerits of arithmetic mean, Median - Merits and demerits of median, Mode - Merits and demerits of mode, Geometric Mean, Harmonic Mean, Positional Averages, Dispersion – Range - Quartile deviations, Mean deviation ,Standard Deviation - Properties of standard deviation Coefficient of Variance

UNIT III Probability Distribution

Theory of Probability and Probability Distribution: Introduction - Definition of probability - Basic terminology used in probability theory, Approaches to probability, Rules of Probability - Addition rule - Multiplication rule, Conditional Probability, Steps Involved in Solving Problems on Probability, Bayes' Probability, Random Variables. Introduction - Random variables, Probability Distributions - Discrete probability distributions - Continuous probability distributions, Bernoulli Distribution - T, Binomial Distribution - Poisson Distribution - Normal Distribution

UNIT IV Hypothesis Testing

Testing of Hypothesis in Case of Large and Small Samples: Introduction – Large Samples – Assumptions, Testing Hypothesis - Null and alternate hypothesis - Selecting a Significance Level - Preference of type I error - Preference of type II error- Determine appropriate distribution, Two – Tailed Tests and One – Tailed Tests - Two – tailed tests. Classification of Test Statistics - Statistics used for testing of hypothesis - Test procedure - How to identify the right statistics for the test , Introduction – small samples, 't' Distribution, Uses of 't' test, Chi- Square - Applications of Chi-Square test - Tests for independence of attributes - Test of goodness of fit - Test for specified variance, F – Distribution and Analysis of Variance (ANOVA): Introduction, Analysis of Variance (ANOVA), Assumptions for F-test - Objectives of ANOVA - ANOVA table - Assumptions for study of ANOVA, Classification of ANOVA - ANOVA table in one-way ANOVA - Two way classifications. Simple Correlation and Regression: Introduction , Correlation - Causation and Correlation - Types of Correlation - Measures of Correlation - Scatter diagram - Karl Pearson's correlation coefficient - Spearman's Rank Correlation Coefficient. Regression - Regression analysis - Regression lines - Regression coefficient , Standard Error of Estimate , Multiple Regression Analysis , Reliability of Estimates , Application of Multiple Regressions

UNIT V Index Number

Index Numbers: Introduction, Definition of an Index Number – Relative - Classification of index numbers , Base year and current year - Chief characteristics of index numbers - Main steps in the construction of index numbers, Methods of Computation of Index Numbers – Un-weighted index numbers - Weighted index numbers, Tests for Adequacy of Index Number Formulae , Cost of Living Index Numbers of Consumer Price Index - Utility of consumer price index numbers - Assumptions of cost of living index number - Steps in construction of cost of living index numbers , Methods of Constructing Consumer Price Index - Aggregate expenditure method - Family budget method - Weight average of price relatives, Limitations of Index Numbers , Utility and Importance of Index Numbers

Note: Problems 60 Marks and Theory 40 Marks.

SUGGESTED READINGS:

1. Levin Richard , H. Siddiqui Masood, S. Rubin David, Rastogi Sanjay, (2017), *Statistics for Management*, 8th edition, pearson education, New Delhi.
2. Amir Aczel, Jayavel Sounderpandian, P Saravanan (2017), *Complete Business Statistics*, 7th edition, Mcgraw Hill Education, New Delhi.
3. Anderson et.al (2015), *Statistics for Business and Economics*, Cengage, New Delhi.
4. Ken Black (2012), *Applied Business Statistics,* 7th edition, Wiley, New Delhi.
5. SP Gupta (2012), *Statistical Methods*, S Chand Publishing, New Delhi.

UNIT – I

Introduction to Statistics: Introduction to Statistics, Importance of Statistics in modern business environment.Classification, Tabulation and Presentation of Data: Introduction , Functions of Classification - Requisites of a good classification - Types of classification - Methods of classification ,Tabulation - Basic difference between classification and tabulation -Parts of a table -Types of table , Frequency and Frequency Distribution - Derived frequency distributions - Bivariate and multivariate frequency distribution - Construction of frequency distribution , Presentation of Data – Diagrams, Graphical Presentation - Histogram - Frequency polygon -   Frequency curve - Ogives

KAHE

INTRODUCTION

ROLE OF MATHEMATICS AND STATISTICS IN BUSINESS DECISIONS:

Mathematics is used in most aspects of daily life. Many of the top jobs such as business consultants,computer consultants, airline pilots, company directors and a host of others require a solid understanding of basic mathematics, and in some cases require a quite detailed knowledge of mathematics. It also plays important role in business, like Business mathematics by commercial enterprises to record and manage business operations. Mathematics typically used in commerce includes elementary arithmetic, such as fractions, decimals, and percentages, elementary algebra, statistics and probability.

Business management can be made more effective in some cases by use of more advanced mathematics such as calculus, matrix algebra and linear programming. Commercial organizations use mathematics in accounting, inventory management, marketing, sales forecasting, and financial analysis. In Academia, "Business Mathematics" includes mathematics courses taken at an undergraduate level by business students. These courses are slightly less difficult and do not always go into the same depth as other mathematics courses for people majoring in mathematics or science fields. The two most common math courses taken in this form are Business Calculus and Business Statistics. Examples used for problems in these courses are usually real-life problems from the business world.

An example of the differences in coursework from a business mathematics course and a regular mathematics course would be calculus. In a regular calculus course, students would study trigonometric functions. Business calculus would not study trigonometric functions because it would be time- consuming and useless to most business students, except perhaps economics majors. Economics majors who plan to continue economics in graduate school are strongly encouraged to take regular calculus instead of business calculus, as well as linear algebra and other advanced math courses. Other subjects typically covered in business mathematics curriculum include: Matrix algebra Linear programming Probability theory Another meaning of business mathematics, sometimes called commercial math or consumer math, is a group of practical subjects used in commerce and everyday life. In schools, these subjects are often taught to students who are not planning a university education.

In the United States, they are typically offered in high schools and in schools that grant associate's degrees. A U.S. business math course might include a review of elementary arithmetic, including fractions, decimals, and percentages. Elementary algebra is often included as well, in the context of solving practical business problems. The practical applications typically include checking accounts, price discounts, markups and markdowns, payroll calculations, simple and compound interest, consumer and business credit, and mortgages.

KAHE

The emphasis in these courses is on computational skills and their practical application, with practical application predominating. For example, while computational formulas are covered in the material on interest and mortgages, the use of prepared tables based on those formulas is also presented and emphasized. Mathematics can provide powerful support for business decisions. In their later business careers, this will motivate them to consult with mathematicians and employ effective quantitative methods.

Mathematics provides many important tools for economics and other business fields. However, our discipline does not profit from this work when students (who later become part of the general public) are unaware of its existence. Presenting trivial mathematical applications only makes matters worse, since they are clearly recognizable as being of little importance. This actually diminishes our subject in the eyes of students. Using computers to bring the underlying structure of significant mathematics to undergraduates allows them to appreciate the role that our Subject can play in their academic work and later lives.

The recognition of its importance by many students each year will certainly strengthen the position of mathematics in our society. Why do business consultants and directors need to know math?" you may ask. Business is all about selling a product or service to make money. All transactions within a business have to be recorded in the Company accounts and quite often involve very large sums of money. So for example, you need to be able to estimate the effect of changing numbers in the accounts when trying to work out your expected performance for next year. Also businesses rely heavily on using percentages, in particular anyone who works as a sales person will need to be quick at mental arithmetic, approximation and in working out percentages. The more percentage discount you give a customer when you sell them a product, the less profit your company will make (and quite often the less you will be paid!) so it really does pay to know your math.

If you work as a sales assistant in many stores you now need to have the ability to calculate the cost of goods and change the customers require without using the till. Businesses like to know that you can cope if the machines break down and also they believe that you can give better customer service if you can respond to customers who know their mathematics. This is the stuff of letters which often appear in local newspapers as "… I bought 2 of the same item at Shop priced at $3.00, and gave the young sales assistant a $10 note and a $1 coin expecting to get a $5 note as change and do my bit to help prevent the store from running out of change in the till. To my amazement the sales assistant insisted that I had paid too much, I tried to explain to no avail but in the end reluctantly took back my $1 coin and was given 4 more $1 coins as change. Finally, there are jobs around where you can escape from using any math at all - refuse collector, builder's laborer, farm hand etc.

KAHE

However, when you invest your hard earned cash in the bank or building society or get a loan - how do you know that you are not being ripped off? You need to use math to calculate compound interest rates (to see how much your savings can grow). You also need to use math to understand the monthly percentages, which are added to your credit cards or bank loans, or you could end up paying $10,000 in 5 year's time for borrowing $2,000 today! This is a good reason to understand mathematics.

STATISTICS:

Statistics is a branch of mathematicsconcerned with the study of information that is expressed in numbers.

When census data cannot be collected, statisticians collect data by developing specific experiment designs and survey samples. Representative sampling assures that inferences and conclusions can reasonably extend from the sample to the population as a whole.
An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements. In contrast, an observational study does not involve experimental manipulation.

Two main statistical methods are used in data analysis: descriptive statistics, which summarize data from a sample using indexes such as the mean or standard deviation, and inferential statistics, which draw conclusions from data that are subject to random variation (e.g., observational errors, sampling variation).[3] Descriptive statistics are most often concerned with two sets of properties of a *distribution* (sample or population): *central tendency* (or *location*) seeks to characterize the distribution's central or typical value, while *dispersion* (or *variability*) characterizes the extent to which members of the distribution depart from its center and each other. Inferences on mathematical statistics are made under the framework of probability theory, which deals with the analysis of random phenomena.

SCOPE AND LIMITATIONS OF STATISTICS

Introduction

The term "statistics" is used in two senses : first in plural sense meaning a collection of numerical facts or estimates—the figure themselves. It is in this sense that the public usually think of statistics, e.g., figures relating to population, profits of different units in an industry etc. Secondly, as a singular noun, the term'statistics' denotes the various methods adopted for the collection, analysis and interpretation of the factsnumerically represented. In singular sense, the term 'statistics' is better described as statistical methods. In our study of the subject, we shall be more concerned with the second meaning of the word 'statistics'.

Definition

Statistics has been defined differently by different authors and each author has assigned new limits to the field which should be included in its scope.

We can do no better than give selected definitions of statistics by some authors and then come to the conclusion about the scope of the subject.

A.L. Bowley defines, **"Statistics may be called the science of counting".** At another place he defines, **"Statistics may be called the science of averages".** Both these definitions are narrow and throw light only on one aspect of Statistics.

According to King, **"The science of statistics is the method of judging collective, natural or social, phenomenon from the results obtained from the analysis or enumeration or collection of estimates".**

Many a time counting is not possible and estimates are required to be made. Therefore, Boddington defines it as "the science of estimates and probabilities". But this definition also does not cover the entire scope of statistics. The statistical methods are methods for the collection, analysis and interpretation of numerical data and form a basis for the analysis and comparison of the observed phenomena. In the words of Croxton &Cowden, "Statistics may be defined as the collection, presentation, analysis and interpretation of numericaldata".

Horace Secrist has given an exhaustive definition of the term satistics in the plural sense. According to him**"By statistics we mean aggregates of facts affected to a marked extent by a multiplicity of causesnumerically expressed, enumerated or estimated according to reasonable standards of accuracy collected in a systematic manner for a pre-determined purpose and placed in relation to each other".** This definition makes it quite clear that as numerical statement of facts, 'statistic' should possess the following characterics.

1.Statistics are aggregate of facts A single age of 20 or 30 years is not statistics, a series of ages are. Similarly, a single figure relating to production, sales, birth, death etc., would not be statistics although aggregates of such figures would be statistics because of their comparability and relationship.

2.Statistics are affected to a marked extent by a multiplicity of causes A number of causes affect statistics in a particular field of enquiry, e.g., in production statistics are affected by climate, soil, fertility, availability of raw materials and methods of quick transport.

KAHE

3.Statistics are numerically expressed,enumerated or estimated The subject of statistics is concerned essentially with facts expressed in numerical form—with theirquantitative details but not qualitative descriptions. Therefore, facts indicated by terms such as 'good', 'poor' are not statistics unless a numerical equivalent, is assigned to each expression. Also this may either beenumerated or estimated, where actual enumeration is either not possible or is very difficult.

4. Statistics are numerated or estimated according to reasonable standard of accuracy Personal bias and prejudices of the enumeration should not enter into the counting or estimation of figures, otherwise conclusions from the figures would not be accurate.

The figures should be counted or estimated according to reasonable standards of accuracy.Absolute accuracy is neither necessary nor sometimes possible in social sciences.But whatever standard of accuracy is once adopted, should be used throughout the process of collection or estimation.

5. Statistics should be collected in a systematic manner for a predetermined purpose The statistical methods to be applied on the purpose of enquiry since figures are always collected with some purpose. If there is no predetermined purpose, all the efforts in collecting the figures may prove to be wasteful. The purpose of a series of ages of husbands and wives may be to find whether young husbands have young wives and the old husbands have old wives.

6. Statistics should be capable of being placed in relation to each other The collected figure should be comparable and well-connected in the same department of inquiry. Ages of husbands are to be compared only with the corresponding ages of wives, and not with, say, heights of trees.

Functions of Statistics

The functions of statistics may be enumerated as follows :

(i) To present facts in a definite form : Without a statistical study our ideas are likely to be vague, indefinite and hazy, but figures helps as to represent things in their true perspective. For example, the statement that some students out of 1,400 who had appeared, for a certain examination, were declared successful would not give as much information as the one that 300 students out of 400 who took the examination were declared successful.

KAHE

(ii) To simplify unwieldy and complex data : It is not easy to treat large numbers and hence theyare simplified either by taking a few figures to serve as a representative sample or by taking average to give a bird's eye view of the large masses. For example, complex data may be simplified by presenting them in the form of a table, graph or diagram, or representing it through an average etc.

(iii) To use it as a technique for making comparisons : The significance of certain figures can be better appreciated when they are compared with others of the same type. The comparison between two different groups is best represented by certain statistical methods, such as average, coefficients, rates, ratios, etc.

(iv) To enlarge individual experience : An individual's knowledge is limited to what he can observe and see; and that is a very small part of the social organism. His knowledge is extended n various ways by studying certain conclusions and results, the basis of which are numerical investigations. For example, we all have general impression that the cost of living has increased.

But to know to what extent the increase has occurred, and how far the rise in prices has affected different income groups, it would be necessary to ascertain the rise in prices of articles consumed by them.

(v) To provide guidance in the formulation of policies : The purpose of statistics is to enable correct decisions, whether they are taken by a businessman or Government. In fact statistics is a great servant of business in management, governance and development. Sampling methods are employed in industry in tacking the problem of standardisation of products. Big business houses maintain a separate department for statistical intelligence, the work of which is to collect, compare and coordinate figures for formulating future policies of the firm regarding production and sales.

(vi) To enable measurement of the magnitude of a phenomenon : But for the development of the statistical science, it would not be possible to estimate the population of a country or to know the quantity of wheat, rice and other agricultural commodities produced in the country during any year.

KAHE

Importance of Statistics

These days statistical methods are applicable everywhere. There is no field of work in which statistical methods are not applied. According to A L. Bowley, ‘A knowledge of statistics is like a knowledge of foreign languages or of Algebra, it may prove of use at any time under any circumstances”. The importance of the statistical science is increasing in almost all spheres of knowledge, e g., astronomy, biology, meteorology, demography, economics and mathematics. Economic planning without statistics is bound to be baseless.

Statistics serve in administration, and facilitate the work of formulation of new policies. Financial institutions and investors utilise statistical data to summaries the past experience. Statistics are also helpful to an auditor, when he uses sampling techniques or test checking to audit the accounts of his client.

Limitations of Statistics

The scope of the science of statistic is restricted by certain limitations :

1. The use of statistics is limited numerical studies: Statistical methods cannot be applied to study the nature of all type of phenomena. Statistics deal with only such phenomena as are capable of being quantitatively measured and numerically expressed. For, example, the health, poverty and intelligence of a group of individuals, cannot be quantitatively measured, and thus are not suitable subjects for statistical study.

2. Statistical methods deal with population or aggregate of individuals rather than with individuals. When we say that the average height of an Indian is 1 metre 80 centimetres, it shows the height not of an individual but as found by the study of all individuals.

3. Statistical relies on estimates and approximations : Statistical laws are not exact laws like mathematical or chemical laws. They are derived by taking a majority of cases and are not true for every individual. Thus the statistical inferences are uncertain.

KAHE

4. Statistical results might lead to fallacious conclusions by deliberate manipulation of figures and unscientific handling. This is so because statistical results are represented by figures, which are liable to be manipulated. Also the data placed in the hands of an expert may lead to fallacious results. The figures may be stated without their context or may be applied to a fact other than the one to which they really relate. An interesting example is a survey made some years ago which reported that 33% of all the girl students at John Hopkins University had married University teachers. Whereas the University had only three girls student at that time and one of them married to a teacher.

Distrust of Statistics

Due to limitations of statistics an attitude of distrust towards it has been developed. There are some people who place statistics in the category of lying and maintain that, “there are three degrees of comparison in lying-lies, dammed lies and statistics”. But this attitude is not correct. The person who is handling statistics may be a liar or inexperienced. But that would be the fault not of statistics but of the person handling them.

The person using statistics should not take them at their face value. He should check the result from an independent source. Also only experts should handle the statistics otherwise they may be misused. It may be noted that the distrust of statistics is due more to insufficiency of knowledge regarding the nature, limitations and uses of statistics then to any fundamental inadequacy in the science of statistics. Medicines are meant for curing people, but if they are unscientifically handle by quacks, they may prove fatal to the patient. In both the cases, the medicine is the same; but its usefulness or harmfulness depends upon the man who handles it.
We cannot blame medicine for such a result. Similarly, if a child cuts his finger with a sharp knife, it is not a knife that is to be blamed, but the person who kept the knife at a place that the child could reach it. These examples help us in emphasising that if statistical facts are misused by some people it would be wrong to blame the statistics as such. It is the people who are to be blamed. In fact statistics are like clay which can be moulded in any way.

Collection of data

For studying a problem statistically first of all, the data relevant thereto must be collected. The numerical facts constitute the raw material of the statistical process. The interpretation of the ultimate conclusion and the decisions depend upon the accuracy with which the data are collected. Unless the data are collected with sufficient care and are as accurate as is necessary for the purposes of the inquiry, the result obtained cannot be expected to be valid or reliable.

Before starting the collection of the data, it is necessary to know the sources from which the data are to be collected.

Primary and Secondary Sources

The original compiler of the data is the primary source. For example, the office of the Registrar General will be the primary source of the decennial population census figures.

A secondary source is the one that furnishes the data that were originally compiled by someone else.

If the population census figures issued by the office of the Registrar-General are published in the Indian year Book, this publication will be the secondary source of the population data.

The source of data also are classified according to the character of the data yielded by them. Thus the data which are gathered from the primary source is known as primary data and the one gathered from the secondary source is known as secondary data. When an investigator is making use of figures which he has obtained by field enumeration, he is said to be using primary data and when he is making use of figures which he has obtained from some other source, he is said to be using secondary data.

Choice between Primary and Secondary Data

An investigator has to decide whether he will collect fresh (primary) data or he will compile data from the published sources. The former is reliable per se but the latter can be relied upon only by examining the following factors :—

(i) source from which they have been obtained;

(ii) their true significance;

(iii) completeness and

(iv) method to collection.

In addition to the above factors, there are other factors to be considered while making choice between the primary or secondary data :

KAHE

(i) Nature and scope of enquiry.

(ii) Availability of time and money.

(iii) Degree of accuracy required and

(iv) The status of the investigator i.e., individual, Pvt. Co., Govt. etc.

However, it may be pointed out that in certain investigations both primary and secondary data may have to be used, one may be supplement to the other.

Methods of Collection of Primary Data

The primary methods of collection of statistical information are the following :

1. Direct Personal Observation,

2. Indirect Personal Observation,

3. Schedules to be filled in by informants

4. Information from Correspondents, and

5. Questionnaires in charge of enumerators

The particular method that is decided to be adopted would depend upon the nature and availability of time, money and other facilities available to the investigation.

1. Direct Personal Observation

According to this method, the investigator obtains the data by personal observation. The method is adopted when the field of inquiry is small. Since the investigator is closely connected with the collection of data, it is bound to be more accurate.

KAHE

Thus, for example, if an inquiry is to be conducted into the family budgets and giving conditions of industrial labour, the investigation himself live in the industrial area as one of the industrial workers, mix with other residents and make patience and careful personal observation regarding how they spend, work and live.

2. Indirect Personal Observation

According to this method, the investigator interviews several persons who are either directly or indirectly in possession of the information sought to be collected. It may be distinguished form the first method in which information is collected directly from the persons who are involved in the inquiry. In the case of indirect personal observation, the persons from whom the information is being collected are known as witnesses or informants. However it should be ascertained that the informants really passes the knowledge and they are not prejudiced in favour of or against a particular view point. This method is adopted in the following situations :

1. Where the information to be collected is of a complete nature.

2. When investigation has to be made over a wide area.

3. Where the persons involved in the inquiry would be reluctant to part with the information.

This method is generally adopted by enquiry committee or commissions appointed by government.

3. Schedules to be filled in by the informants

Under this method properly drawn up schedules or blank forms are distributed among the persons from whom the necessary figure are to be obtained. The informants would fill in the forms and return them to the officer incharge of investigation. The Government of India issued slips for the special enumeration of scientific and technical personnel at the time of census. These slips are good examples of schedules to be filled in by the informants.

The merit of this method is its simplicity and lesser degree of trouble and pain for the investigator. Its greatest drawback is that the informants may not send back the schedules duly filled in.

KAHE

4. Information from Correspondents

Under this method certain correspondent are appointed in different parts of the field of enquiry, who submit their reports to the Central Office in their own manner. For example, estimates of agricultural wages may be periodically furnished to the Government by village school teachers.

The local correspondents being on the spot of the enquiry are capable of giving reliable information.

But it is not always advisable to place much reliance on correspondents, who have often got their own personal prejudices. However, by this method, a rough and approximate estimate is obtained at a very low cost. This method is also adopted by various departments of the government in such cases where regular information is to be collected from a wide area.

Questionnaire incharge of Enumerations

A questionnaire is a list of questions directly or indirectly connected with the work of the enquiry. The answers to these questions would provide all the information sought. The questionnaire is put in the charge of trained investigators whose duty is to go to all persons or selected persons connected with the enquiry. This method is usually adopted in case of large inquiries. The method of collecting data is relatively cheap. Also the information obtained is that of good quality.
The main drawback of this method is that the enumerator (i.e., investigator in charge of the questionnaire) may be a biased one and may not enter the answer given by the information. Where there are many enumerators, they may interpret various terms in questionnaire according to their whims. To that extent the information supplied may be either inaccurate or inadequate or not comparable. This drawback can be removed to a great extent by training the investigators before the enquiry begins. The meaning of different questions may be explained to them so that they do not interpret them according to their whims.

Drafting the Questionnaire

The success of questionnaire method of collecting information depends on the proper drafting of the questionnaire. It is a highly specialized job and requires great deal of skill and experience. However, the following general principle may be helpful in framing a questionnaire :

1. The number of the questions should be kept to the minimum fifteen to twenty five may be a fair number.

2. The questions must be arranged in a logical order so that a natural and spontaneous reply to each is induced.

3. The questions should be short, simple and easy to understand and they should convey one meaning.

4. As far as possible, quotation of a personal and pecuniary nature should not be asked.

5. As far as possible the questions should be such that they can be answered briefly in 'Yes' or 'No', or in terms of numbers, place, date, etc.

6. The questionnaire should provide necessary instructions to the Informants. For instance, if there is a question on weight. It should be specified as to whether weight is to be indicated in lbs or kilograms.

7. Questions should be objective type and capable of tabulation.

Specimen Questionnaire

We are giving below a specimen questionnaire of Expenditure Habits or Students residing in college Hostels.

Name of Student .... Class ....State and
District of origin ....

Age ....
1. How much amount do you get from your father/guardian p.m. ?
2. Do you get some scholarship ? If so, state the amount per month.
3. Is there any other source from which you get money regularly ?

(e.g. mother, brother or uncle).

KAHE

4. How much do you spend monthly on the following items :

	Rs.
College Tuition Fee	....
Hostel Food Expenses	....
Other hostel fees	....
Clothing	....
Entertainment	....
Smoking	....
Miscellaneous	....
Total	....

5. Do you smoke ?

If so what is the daily expenditure on it ?

6. Any other item on which you spend money ?

Sources of Secondary Data

There are number of sources from which secondary data may be obtained. They may be classified as follow. :

1. Published sources, and

2. Unpublished sources.

1. Published Sources The various sources of published data are :
1. Reports and official publications of-
(a) International bodies such as the International Monetary Fund, International Finance Corporation, and United Nations Organisation.

(b) Central and State Governments- such as the Report of the Patel Committee, etc.

2. Semi Official Publication. Various local bodies such as Municipal Corporation, and Districts Boards.
3. Private Publication of—
(a) Trade and professional bodies such as the Federation of India, Chamber of Commerce and Institute of Chartered Accountants of India.
(b) Financial and Economic Journals such as “Commerce”, ‘Capital’ etc.
(c) Annual Reports of Joint Stock Companies.
(d) Publication brought out by research agendas, research scholars, etc.

KAHE

2. Unpublished Sources

There are various sources of unpublished data such as records maintained by various government and private offices, studies made by research institutions, scholars, etc., such source can also be used where necessary.

Census and Sampling Techniques of Collection of Data

There are two important techniques of Data collection, (i) Census enquiry implies complete enumeration of each unit of the universe, (ii) In a sample survey, only a small part of the group, is considered, which is taken as representative. For example the population census in India implies the counting of each and every human being within the country.

In practice sometimes it is not possible to examine every item in the population. Also many a time it is possible to obtain sufficiently accurate results by studying only a part of the “population”. For example, if the marks obtained in statistics by 10 students in an examination are selected at random, say out of 100, then the average marks obtained by 10 students will be reasonably representative of the average marks obtained by all the 100 students. In such a case, the populations will be the marks of the entire group of 100 students and that of 10 students will be a sample.

Objects of Sampling

1. To get as much information as possible of the whole universe by examining only a part of it.
2. To determine the reliability of the estimates. This can be done by drawing successive samples from the some parent universe and comparing the results obtained from different samples.

Advantages of Census Method

1. As the entire ‘population’ is studied, the result obtained are most correct.
2. In a census, information is available for each individual item of the population which is not possible in the case of a sample. Thus no information is sacrificed under the census method.
3. If data are to be secured only from a small fraction of the aggregate, their completeness and accuracy can be ensured only by the census method, since greater attention thereby is given to each item.
4. The census mass of data being taken into consideration all the characteristics of the ‘population’ is maintained in original.

Disadvantages of Census Method

1. The cost of conducting enquiry by the census method is very high as the whole universe is to be investigated.
2. The census method is not practicable in very big enquiries due to the inconvenience of individual enumeration.
3. In the cases of very big enquiries, the census method can be resorted to by the government agencies only. The application of this method is limited to those who are having adequate financial resources and other facilities at their disposal.
4. As all the items in the universe are to be enumerated, there is a need for training of staff and investigators. Sometimes it becomes very difficult to maintain uniformity of standards, when many investigators are involved. Individual preferences and prejudices are there and it becomes very difficult to avoid bias in such type of enquiries.

Advantages of Sampling Method

1. Sample method is less costly since the sample is a small fraction of the total population.
2. Data can be collected and summarized more quickly. This is a vital consideration when the information is urgently needed.
3. A sample produces more accurate results than are ordinarily practicable on a complete enumeration.
4. Personnel of high quality can be employed and given intensive training as the number of much personnel would not be very large.
5. A sample method is not restricted to the Government agencies. Even private agencies can use this method as the financial burden is not heavy. It is much more economical than the census method.

Disadvantages of Sampling Method

1. In a census, information is available for each individual item of the population which is not possible in the case of a sample. Some information has to be sacrificed.

2. If data are to be secured only from a small fraction of the aggregate, their completeness and accuracy can be ensured only through the census method, since greater attention thereby is given to each item.

3. In using the technique of sampling, the investigator may not choose a representative sample. The aim of sampling is that it should afford a sufficiently accurate picture of a large group without the need for a complete enumeration of all the units of the group.

KAHE

If the sample chosen is not representative of the group, the very object of sampling is defeated.

4. The sampling technique is based upon the fundamental assumption that the population to be sampled is homogenous. It is not so, the sampling method should not be adopted unless the population is first divided into groups or “strata” before the selection of the sample is made.

Principle of sampling

There are two important principles on which the theory of sampling is based ;

1. Principle of Statistical Regularity,and
2. Principle of ‘Intertia of Large Numbers’

1. Principle of Statistical Regularity

This principle points out that if a sample is taken at random from a population. It is likely to possess almost the same characteristics as that of the population. By random selection, we mean a selection where each and every item of the population has an equal chance of being selected in the sample. In other words, the selection must not be made by deliberate exercise of one’s discretion. A sample selected in this manner would be representative of the population. For example, if one intends to make a study of the average weight of the students of Delhi University, it is not necessary to take the weight of each and every student. A few students may be selected at random from every college, their weights taken and the average weight of the University students in general may be inferred.

2. Principle of Intertia of Large Numbers

This principle is a corollary of the principle of statistical regularity. This principle is that, other things being equal, larger the size of the sample, more accurate the results are likely to be. This is because large numbers are more stable as compared to small ones. For example, if a coin is tossed 10 time we should expect an equal number of heads and tails, i.e., 5 each. But since the experiment is tried a small number of items it is likely that we may not get exactly 5 heads and 5 tails. The result may be a combination of 9 heads and 1 tail or 8 heads and 2 tails or 7 heads and 3 tails etc. If the same experiment is carried out 1,000 times the chance of getting 500 heads and 500 tails would be very higher. The basic reason for such likelihood is that the experiment has been carried out a sufficiently large number of time and possibility of variations in one direction compensates for others.

Method of Sampling

The various methods available for sampling are :

(i) Conscious or Deliberate or Purposive Sampling.

(ii) Random Sampling or Chance Selection.

(iii) Stratified Sampling.
(iv) Systematic Sampling.

(v) Multi-stage Sampling.

(i) Purposive Sampling

Purposive sampling is representative sampling by analyzing carefully the universe enquiry and selecting only those which seem to be most representatives of the characteristics of the universe. If economic conditions of people living in a state are to be studied according to this method, then a few villages and towns may be purposively selected so that intensive study on the principle that they shall be representative of the entire state.

Thus the purposive sampling is a purposive selection by the investigator that depends on the nature and purpose of the enquiry. This method is very much exposed to the dangers of personal prejudices. Also there is a possibility of certain wrong cases being included in the data under collection, consciously or unconsciously. However, it may be noted that this method gives a very representative sample data provided neither bias nor prejudices influence the process of data selection.

(ii) Random Sampling

In order to avoid the danger of personal bias and prejudices, a random sample is adopted. Under this method every item in the universe is given equal chance of being included in the sample. A random sample is the simplest type of sample. For obtaining such sample, a certain number of units are selected at random from the universe. But this sampling technique is based upon the fundamental assumption that the population to be swapped is homogenous. If it is not so, then the stratified sampling is adopted.

KAHE

(iii) Stratified Sampling

Under this method, the population is first sub-divided into groups or “strata” before the selection of the samples is made. This is done to achieve homogeneity within each group or “stratum”. A stratified sample is nothing but a set of random samples of a number of sub-populations, each representing a single group. The major advantage of such a stratification is that the several sub-divisions of the population which are relevant for purpose of inquiry are adequately represented.

(iv) Systematic Sampling

This method is used where complete list of the population from which sample is to be drawn is available. The method is to select every rth item*, from the list where ‘r’ refers to the sampling interval. The first item between the first and the rth is selected as random. For example, if a list of 500 students of a college is available and if we want to draw a sample of 100, we must select every fifth item (i.e., r = 5). The first item between one and five shall be selected at random. Suppose it comes out to be 4. Now we shall add five and obtain numbers of the desired sample. Thus the second item would be the 9th students; the third 14th students; the fourth 19 students; and so on.

$$\text{Sampling interval or } r = \frac{\text{size of the universe}}{\text{size of the sample}}$$

This method is more convenient to adopt than the random sampling or stratified sampling method. The time and work involved are relatively smaller. But the main drawback of this method is that systematic samples an not always random samples.

(v) Multi-Stage Sampling

As the name implies this method refers to a sampling procedure which is carried out in several stages. At first stage, the first stage units are sampled by some statistical method, such as random sampling. Then a sample of second stage units is selected from each of the selected first units. Further stages may be added as required. This method introduces flexibility in the sampling method which is lacking in the other methods. However, a multi-stage sample is less accurate than sample containing the same numbers of final stage units which have been selected by some suitable single stage process.

KAHE

DIAGRAMMATIC AND GRAPHIC REPRESENTATION OF DATA:

Importance of Diagrams

1. They are attractive and hence diagrams and graphs are commonly used in newspapers and magazines for the purpose of advertisements and campaign.

2. They give bird's eye view of the entire data at a glance.

3. They can be easily understood by common man.

4. They can be remembered for a longer period of time.

5. They facilitate comparison.

RULES FOR CONSTRUCTING DIAGRAMS AND GRAPHS

1. Serial number: Every diagram or graph must have a serial number. It is necessary to distinguish one from the other.

2. Title: Title must be given to every diagram or graph. From the title one can know the idea contained in it. The title should be brief and self-explanatory. It is usually placed at the top.

3. Proper size and scale: A diagram or graph should be of normal size and drawn with proper scale. The scale in a graphs specifies the size of the unit.

4. Cleanliness: Diagrams must be as simple as possible. Further they must be quite neat and clean. They should also be descent to look at.

5. Index: Every diagram or graph must be accompanied by an index. This illustrates different types of lines, shades or colors used in the diagram.

6. Footnote: Foot notes may be given at the bottom of a diagram if necessary. It clarifies certain points in the diagram.

KAHE

TYPES OF DIAGRAMS

Diagrams may be one-dimensional or two dimensional. In one-dimensional we have bar diagrams. In two dimensional we have pie diagram. Simple bar diagram, component bar diagram, sub-divided bar diagram and percentage bar diagram are different bar diagrams.

Simple bar diagram: Simple bar diagram is drawn when items are to be compared with respect to a single characteristic. A rectangular bar is constructed with height proportional to the magnitude of the items.

Multiple bar diagram: Multiple bar diagrams are drawn when we have two or more sets of comparable values. Component (sub-divided) bar diagram:

Component bar diagrams are used when two or more characteristics are observed on a unit. Each bar is proportionally subdivided.

Component pie diagram: It is drawn when data have magnitudes for two or more components.Circles with area proportional to magnitudes are drawn to represent the total magnitude. Then circles are divided sector-wise according to the magnitude of the components.

GRAPHS FOR PRESENTING FREQUENCY DISTRIBUTION

1. Histogram: The frequency distribution is represented by a set of rectangular bars with area proportional to class frequency. If the class intervals have equal width then the variable is taken along X-axis and frequency along Y-axis and a rectangle is constructed.

2. Frequency polygon: The mid values of class intervals are plotted against frequency of the class interval. These points are joined by straight lines and hence the frequency polygon is obtained.

3. Frequency curve: First we draw histogram for the given data. Then we join the mid points of the rectangles by a smooth curve. Total area under frequency curve represents total frequency. They are the most useful form of frequency distribution.

4. Ogives: Ogive is obtained by drawing the graph of a cumulative frequency distribution. Hence, ogives are also called as cumulative frequency curves. Since a cumulative frequency distribution can be of 'less than' or 'greater than' type, we have less than and greater than type of ogives. Since a cumulative frequency distribution can be of 'less than' or 'greater than' type, we have less than and greater than type of ogives.

KAHE

a) Less than ogive: Variables are taken along X-axis and less than cumulative frequencies are taken along Y-axis. Less than cumulative frequencies are plotted against upper limit of class interval and joined by a smooth-curve.

b) More than ogive: More than cumulative frequencies are plotted against lower limit of the class-interval and joined by a smooth-curve.

From the meeting point of these two ogives if we draw a perpendicular to X-axis, the point where it meets X-axis gives median of the distribution.

DIFFERENCE BETWEEN DIAGRAMS AND GRAPHS

DIAGRAMS

We are too well aware of the use of diagrams to explain information and facts that are presented in the form of text. If you need to explain the parts of a machine or the principle of its working, it becomes difficult to make one understand the concept through text only. This is where diagrams in the form of sketches come into play. Similarly, diagrams are made heavy use of in biology where students have to learn about different body parts and their functions. Visual representation of concepts through diagrams has better chances of retention in the memory of students than presenting them in the form of text.

Diagrams are resorted to right from the time a kid enters a school as even alphabets are presented to him in a more interesting and attractive manner with the help of diagrams.

GRAPHS

Whenever there are two variables in a set of information, it is better to present the information using graphs as it makes it easier to understand the data. For example, if one is trying to show how the prices of commodities have increased with respect to time, a simple line graph would be a more effective and interesting way rather than putting all this information in the form of text which is hard to remember whereas even a layman can see how prices have gone up or down in relation to time.

General Principles of Graphic Representation:

There are some algebraic principles which apply to all types of graphic representation of data. In a graph there are two lines called coordinate axes.

KAHE

One is vertical known as Y axis and the other is horizontal called X axis. These two lines are perpendicular to each other. Where these two lines intersect each other is called 'o' or the Origin. On the X axis the distances right to the origin have positive value and distances left to the origin have negative value. On the Y axis distances above the origin have a positive value and below the origin have a negative value.

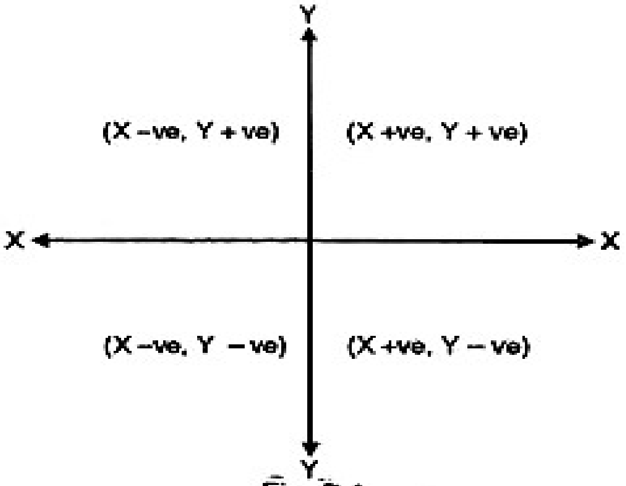


Fig. 7.1

Methods to Represent a Frequency Distribution:

Generally four methods are used to represent a frequency distribution graphically. These are Histogram, Smoothed frequency graph and Ogive or Cumulative frequency graph and pie diagram.

1. Histogram:

Histogram is a non-cumulative frequency graph, it is drawn on a natural scale in which the representative frequencies of the different class of values are represented through vertical rectangles drawn closed to each other. Measure of central tendency, mode can be easily determined with the help of this graph.

How to draw a Histogram:

Step—1:

Represent the class intervals of the variables along the X axis and their frequencies along the Y-axis on natural scale.

Step—2:

Start X axis with the lower limit of the lowest class interval. When the lower limit happens to be a distant score from the origin give a break in the X-axis n to indicate that the vertical axis has been moved in for convenience.

Step—3:

Now draw rectangular bars in parallel to Y axis above each of the class intervals with class units as base: The areas of rectangles must be proportional to the frequencies of the corresponding classes.

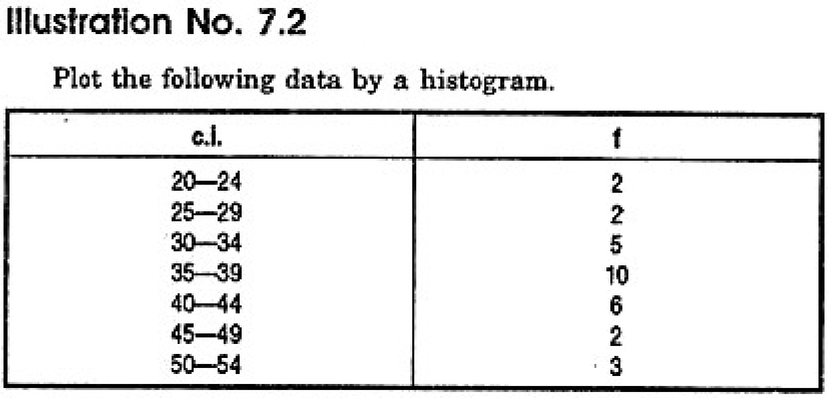


Illustration No. 7.2

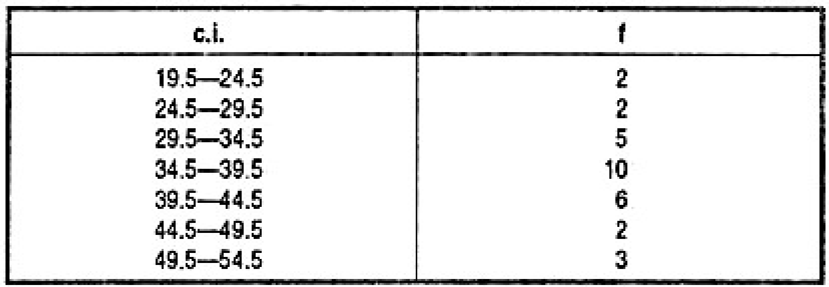
Plot the following data by a histogram.

c.i.	f
20—24	2
25—29	2
30—34	5
35—39	10
40—44	6
45—49	2
50—54	3

Solution:

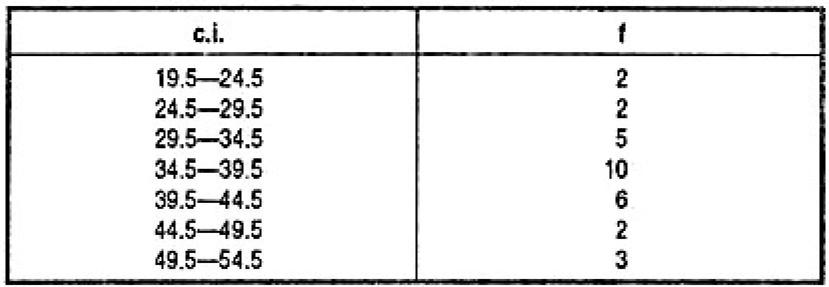
In this graph we shall take class intervals in the X axis and frequencies in the Y axis.

Before plotting the graph we have to convert the class into their exact limits.



c.i.	f
19.5—24.5	2
24.5—29.5	2
29.5—34.5	5
34.5—39.5	10
39.5—44.5	6
44.5—49.5	2
49.5—54.5	3

Histogram plotted from the data.



c.i.	f
19.5—24.5	2
24.5—29.5	2
29.5—34.5	5
34.5—39.5	10
39.5—44.5	6
44.5—49.5	2
49.5—54.5	3

Histogram plotted from the data.

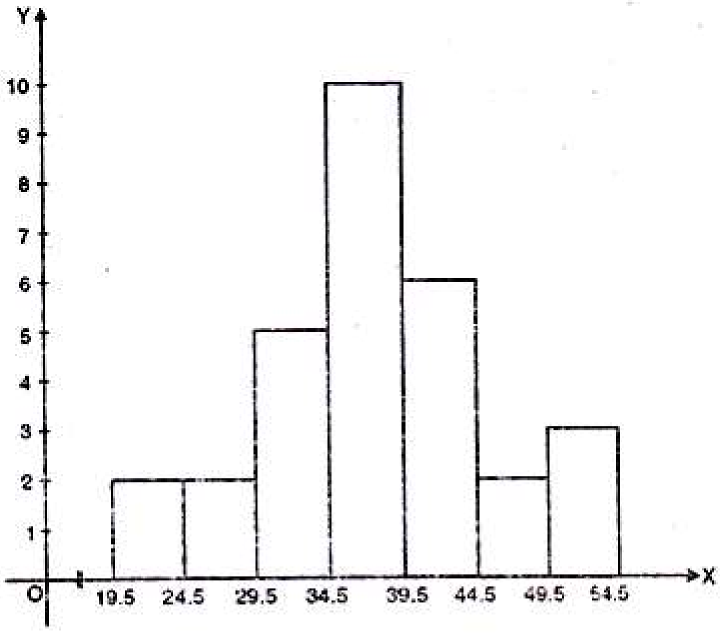


Fig. 7.2

Advantages of histogram:

1. It is easy to draw and simple to understand.

2. It helps us to understand the distribution easily and quickly.

3. It is more precise than the polygene.

Limitations of histogram:

1. It is not possible to plot more than one distribution on same axes as histogram.

2. Comparison of more than one frequency distribution on the same axes is not possible.

3. It is not possible to make it smooth.

Uses of histogram:

1. Represents the data in graphic form.

2. Provides the knowledge of how the scores in the group are distributed. Whether the scores are piled up at the lower or higher end of the distribution or are evenly and regularly distributed throughout the scale.

3. Frequency Polygon. The frequency polygon is a frequency graph which is drawn by joining the coordinating points of the mid-values of the class intervals and their corresponding frequencies.

Let us discuss how to draw a frequency polygon:

Step-1:

Draw a horizontal line at the bottom of graph paper named 'OX' axis. Mark off the exact limits of the class intervals along this axis. It is better to start with c.i. of lowest value. When the lowest score in the distribution is a large number we cannot show it graphically if we start with the origin. Therefore put a break in the X axis () to indicate that the vertical axis has been moved in for convenience. Two additional points may be added to the two extreme ends.

Step-2:

Draw a vertical line through the extreme end of the horizontal axis known as OY axis. Along this line mark off the units to represent the frequencies of the class intervals. The scale should be chosen in such a way that it will make the largest frequency (height) of the polygon approximately 75 percent of the width of the figure.

Step-3:

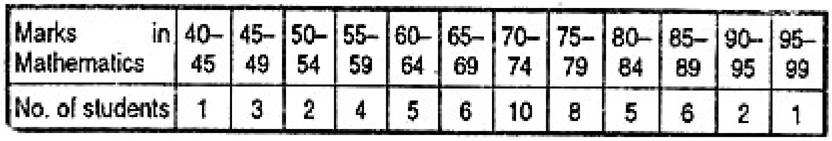
Plot the points at a height proportional to the frequencies directly above the point on the horizontal axis representing the mid-point of each class interval.

Step-4:

After plotting all the points on the graph join these points by a series of short straight lines to form the frequency polygon. In order to complete the figure two additional intervals at the high end and low end of the distribution should be included. The frequency of these two intervals will be zero.

Illustration:

Draw a frequency polygon from the following data:

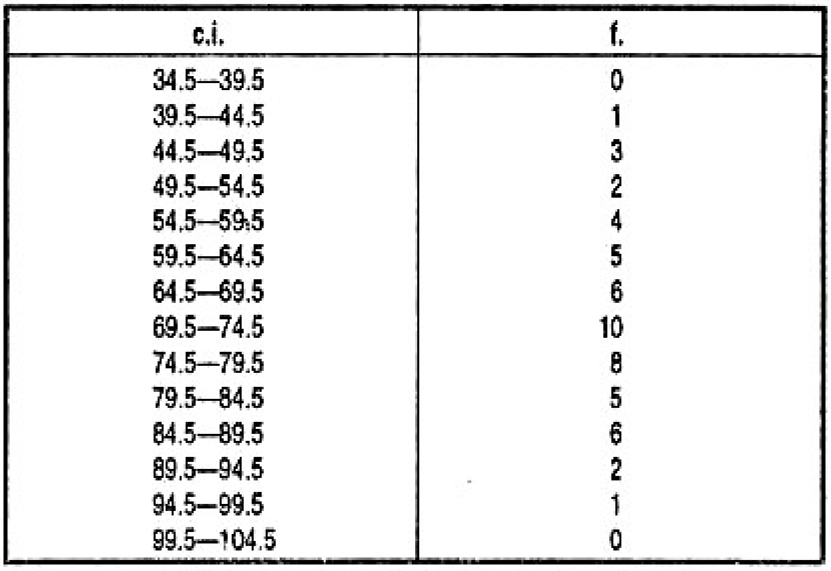


Marks in Mathematics	40–45	45–49	50–54	55–59	60–64	65–69	70–74	75–79	80–84	85–89	90–95	95–99
No. of students	1	3	2	4	5	6	10	8	5	6	2	1

Solution:

In this graph we shall take the class intervals (marks in mathematics) in X axis, and frequencies (Number of students) in the Y axis. Before plotting the graph we have to convert the c.i. into their exact limits and extend one c.i. in each end with a frequency of O.

Class intervals with exact limits:



c.i.	f.
34.5—39.5	0
39.5—44.5	1
44.5—49.5	3
49.5—54.5	2
54.5—59.5	4
59.5—64.5	5
64.5—69.5	6
69.5—74.5	10
74.5—79.5	8
79.5—84.5	5
84.5—89.5	6
89.5—94.5	2
94.5—99.5	1
99.5—104.5	0

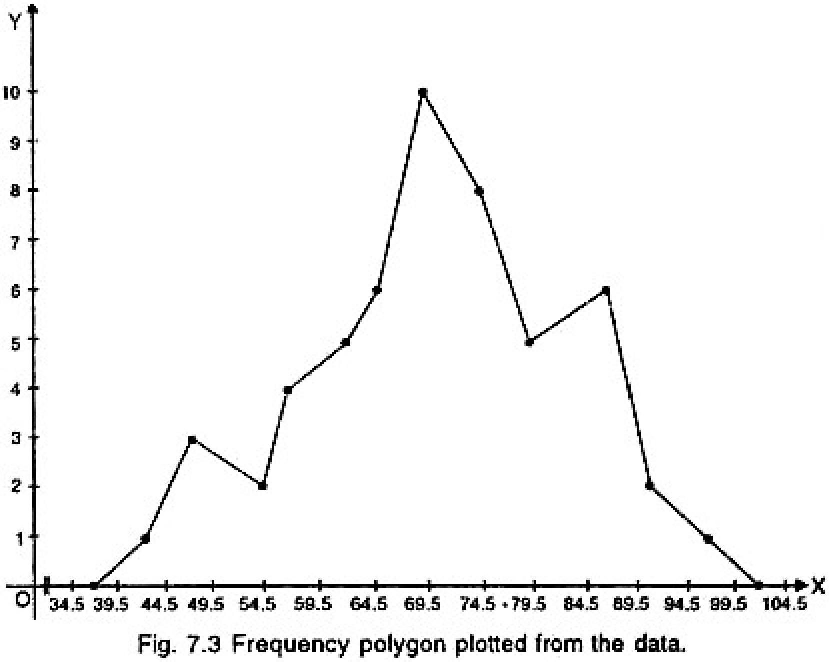


Fig. 7.3 Frequency polygon plotted from the data.

Advantages of frequency polygon:

1. It is easy to draw and simple to understand.

2. It is possible to plot two distributions at a time on same axes.

3. Comparison of two distributions can be made through frequency polygon.

4. It is possible to make it smooth.

Limitations of frequency polygon:

1. It is less precise.

2. It is not accurate in terms of area the frequency upon each interval.

Uses of frequency polygon:

1. When two or more distributions are to be compared the frequency polygon is used.

2. It represents the data in graphic form.

3. It provides knowledge of how the scores in one or more group are distributed. Whether the scores are piled up at the lower or higher end of the distribution or are evenly and regularly distributed throughout the scale.

2. Smoothed Frequency Polygon:

When the sample is very small and the frequency distribution is irregular the polygon is very jig-jag. In order to wipe out the irregularities and **"also get a better notion of how the figure might look if the data were more numerous, the frequency polygon may be smoothed."**

In this process to adjust the frequencies we take a series of 'moving' or 'running' averages. To get an adjusted or smoothed frequency we add the frequency of a class interval with the two adjacent intervals, just below and above the class interval. Then the sum is divided by 3. When these adjusted frequencies are plotted against the class intervals on a graph we get a smoothed frequency polygon.

Ogive or Cumulative Frequency Polygon:

Ogive is a cumulative frequency graphs drawn on natural scale to determine the values of certain factors like median, Quartile, Percentile etc. In these graphs the exact limits of the class intervals are shown along the X-axis and the cumulative frequencies are shown along the Y-axis. Below are given the steps to draw an ogive.

Step—1:

Get the cumulative frequency by adding the frequencies cumulatively, from the lower end (to get a less than ogive) or from the upper end (to get a more than ogive).

Step—2:

Mark off the class intervals in the X-axis.

Step—3:

Represent the cumulative frequencies along the Y-axis beginning with zero at the base.

Step—4:

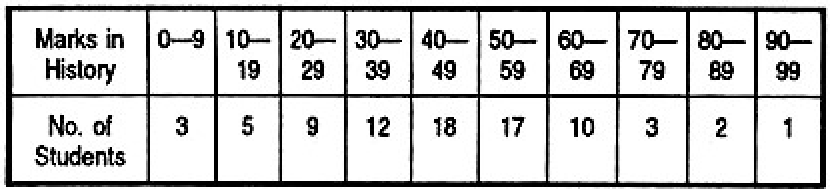
Put dots at each of the coordinating points of the upper limit and the corresponding frequencies.

Step—5:

Join all the dots with a line drawing smoothly. This will result in curve called ogive.

Illustration No. 7.5:

Draw an ogive from the data given below:

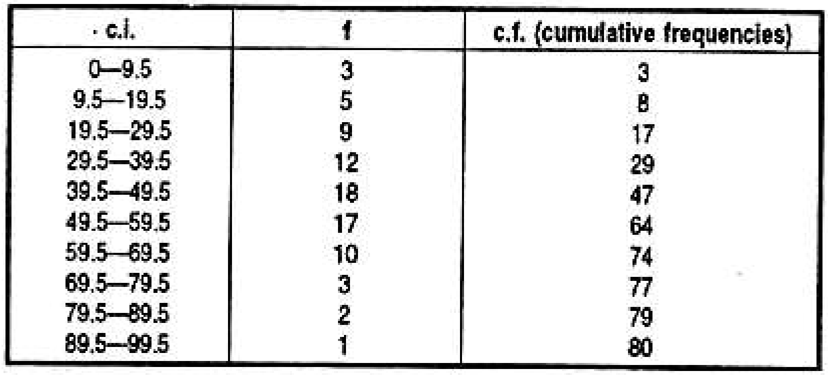


Marks in History	0—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79	80—89	90—99
No. of Students	3	5	9	12	18	17	10	3	2	1

Solution:

To plot this graph first we have to convert, the class intervals into their exact limits.

Then we have to calculate the cumulative frequencies of the distribution.



· c.i.	f	c.f. (cumulative frequencies)
0—9.5	3	3
9.5—19.5	5	8
19.5—29.5	9	17
29.5—39.5	12	29
39.5—49.5	18	47
49.5—59.5	17	64
59.5—69.5	10	74
69.5—79.5	3	77
79.5—89.5	2	79
89.5—99.5	1	80

Now we have to plot the cumulative frequencies in respect to their corresponding class-intervals.

Ogive plotted from the data given above:

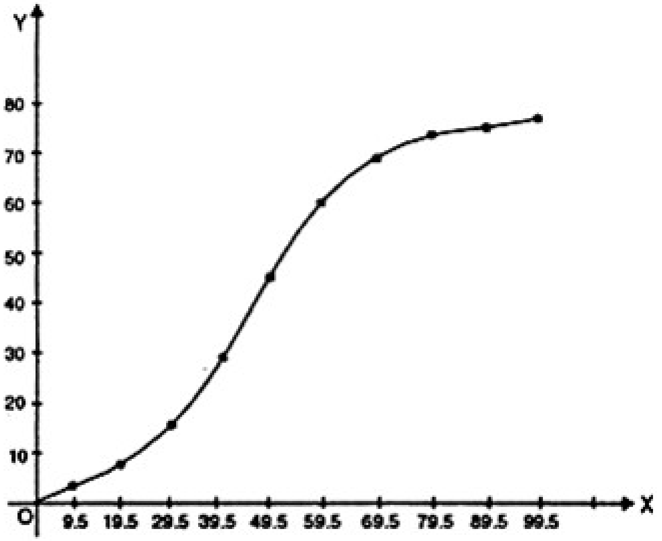


Fig. 7.5.

Uses of Ogive:

1. Ogive is useful to determine the number of students below and above a particular score.

2. When the median as a measure of central tendency is wanted.

3. When the quartiles, deciles and percentiles are wanted.

4. By plotting the scores of two groups on a same scale we can compare both the groups.

4. The Pie Diagram:

Figure given below shows the distribution of elementary pupils by their academic achievement in a school. Of the total, 60% are high achievers, 25% middle achievers and 15% low achievers. The construction of this pie diagram is quite simple.

There are 360 degree in the circle. Hence, 60% of 360′ or 216° are counted off as shown in the diagram; this sector represents the proportion of high achievers students.

Ninety degrees counted off for the middle achiever students (25%) and 54 degrees for low achiever students (15%). The pie-diagram is useful when one wishes to picture proportions of the total in a striking way. Numbers of degrees may be measured off **"by eye"** or more accurately with a protractor.

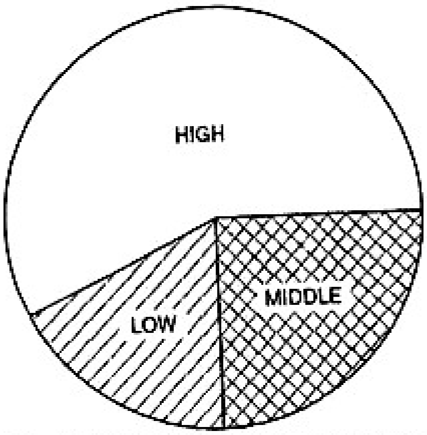


Fig. 7.6. Distribution by academic achievement of pupils in Class VI of a school.

Uses of Pie diagram:

1. Pie diagram is useful when one wants to picture proportions of the total in a striking way.

2. When a population is stratified and each strata is to be presented as a percentage at that time pie diagram is used.

HISTOGRAM

A **histogram** is a graph of the frequency distribution in which the vertical axis represents the count (frequency) and the horizontal axis represents the possible range of the data values. The **histogram** is widely used and needs little explanation.

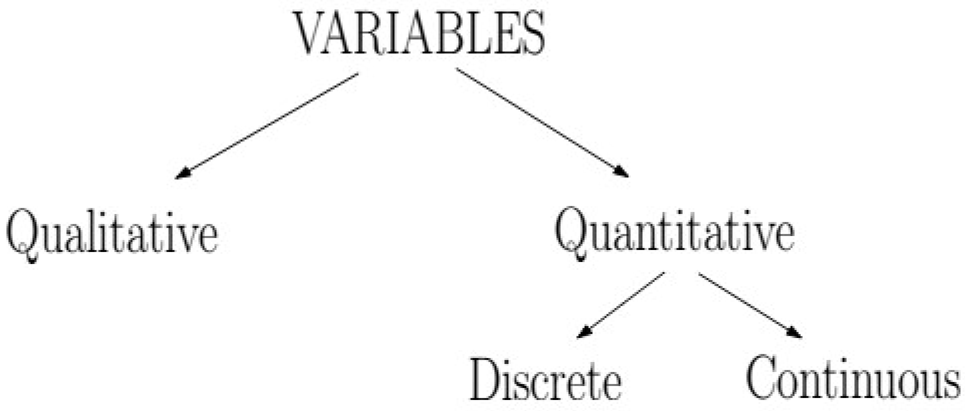
Once a study has been designed and data collected, researchers begin to SUMMARIZE their data. Data may be summarized by plotting figures and computing certain summary measures to obtain important information about the data.

STATISTIC : A summary measure computed from the data.

Recall : Every data point is the value of the response VARIABLE measured on a unit. So we should think of variable as the quantity that takes different values for different individuals.

Examples: gender, color of eyes, weight, bacteria count.

There are two types of variables, dependant upon on their possible values: qualitative (categorical) quantitative (numerical). Quantitative variables are further divided into discrete and continuous.



A qualitative variable places an individual into one of several groups or categories. Such variables are also called categorical variables.

The variable gender has two possible values male and female.

The variable major has numerous values such as Mathematics, Biology, Physics, Economics, Chemistry.

A quantitative variable takes numerical values for which arithmetic operations (such as adding and averaging) make sense. Quantitative variables are also called numerical variables.

NOTE: If unsure on how to classify a variable, question how it can be affected mathematically. We cannot average gender or major, therefore they are qualitative variables

Quantitative variables are divided into discrete and continuous:

Discrete quantitative variable takes on values which are spaced, i.e, for two adjacent values, there is no value that goes between them.

Continuous quantitative variable take values in a given interval. For ANY two values of the variable, we can always find another value that can go between the two.

Variables such as weight, time, and distance are continuous.

NOTE: The variable salary is continuous but essentially discrete if all salaries are rounded to the whole dollar

Classify each of the following variables as qualitative or quantitative (discrete or continuous):

Color of eyes. Qualitative

Blood pressure quantitative

Weight (in lb) quantitative

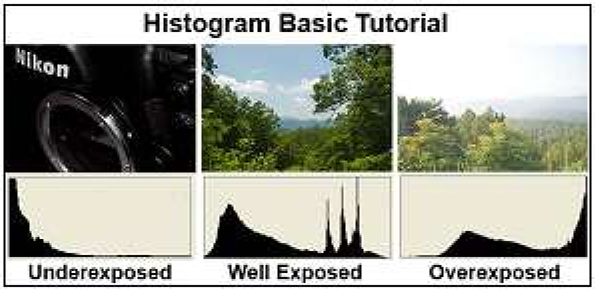
Residence (country) qualitative

Number of patients under a treatment quantitative

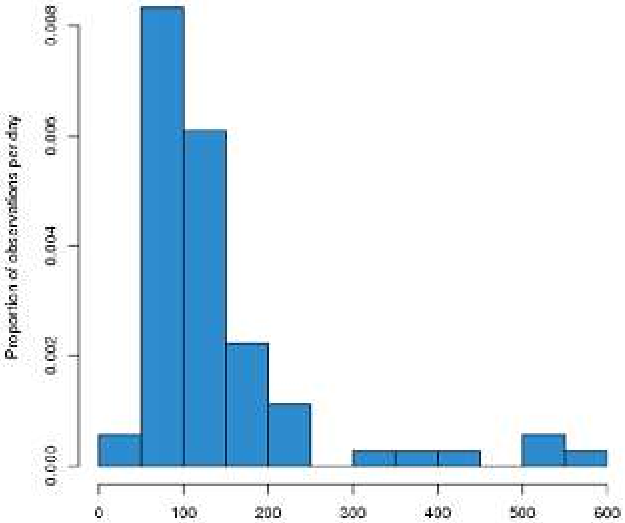
Zip code qualitative

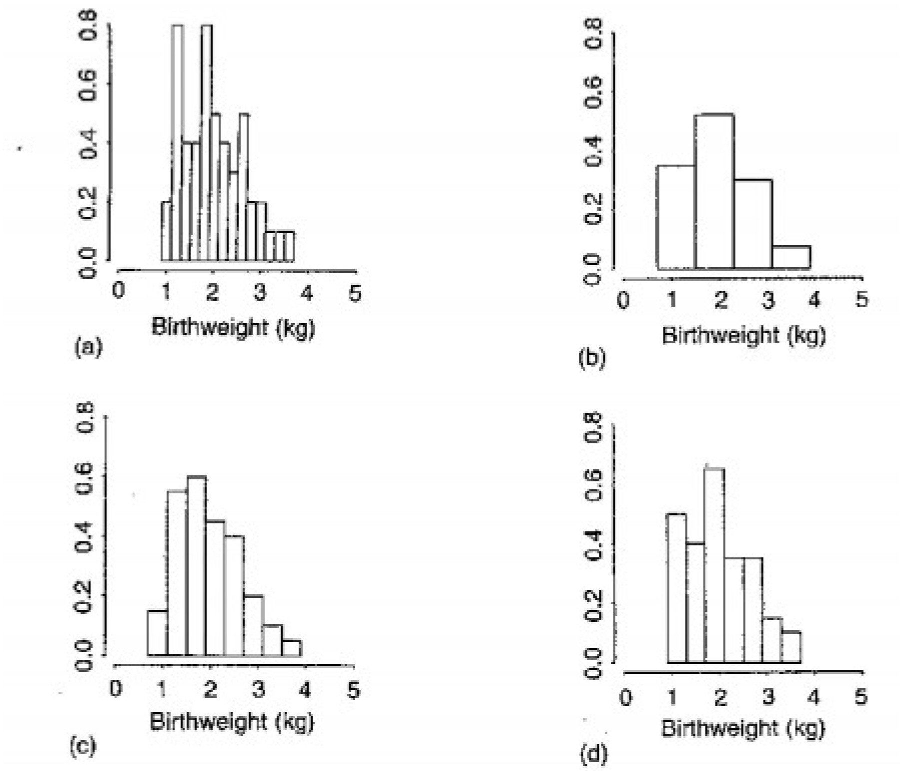
KAHE

An Introduction to the Histogram



- **Data** represents the values of the response variable measured from each unit.
- The **distribution of data** is a list summarizing the observed values of the response variable and how often they were observed.
- When the data is **quantitative**, whether discrete or continuous, a **histogram** may be used to display its distribution.

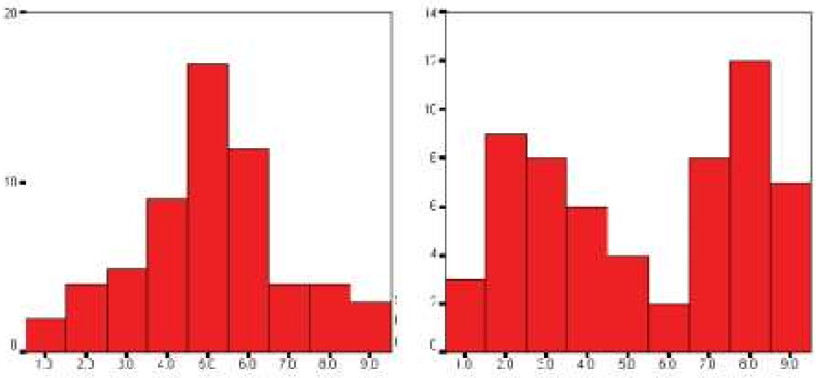




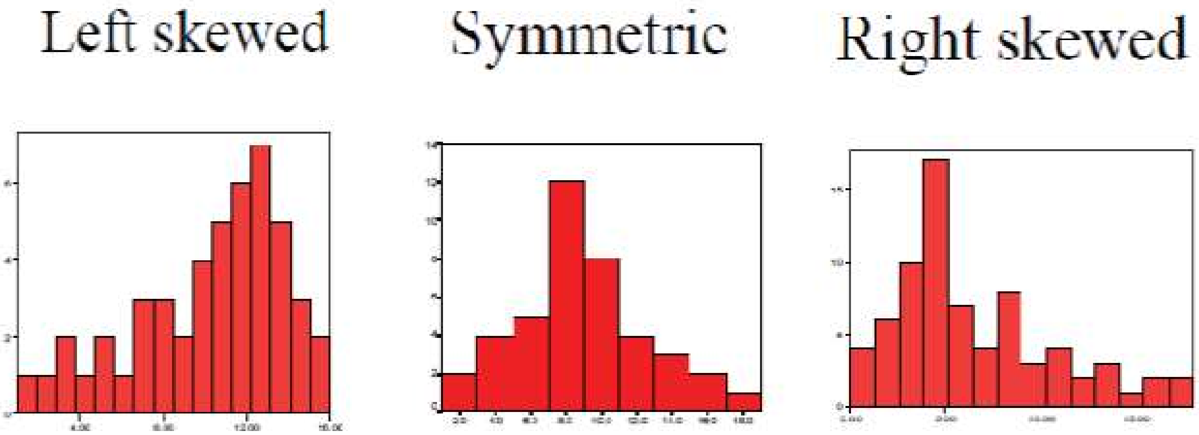
Analyzing the Histogram

Once you have plotted the histogram (data distribution), look for *outliers*, and the *overall pattern*, which is described by its **shape**, **center**, and **spread**. The shape of the distribution can be described

- by specifying the number of **modes**.
- as **symmetric** or **skewed**.



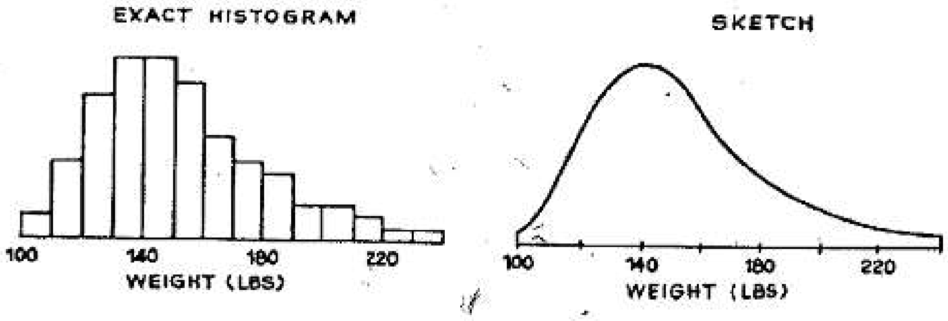
The Appearance of Symmetric and Skewed distributions


Left skewed

Symmetric

Right skewed

Histogram sketches are smooth curves drawn through the tops of the histogram bars and used to indicate the overall shape of a histogram.


EXACT HISTOGRAM

SKETCH

Figure : A Histogram and its Smooth Histogram Sketch.

OGIVE

An **ogive** graph is a plot used in statistics to show cumulative frequencies. It allows us to quickly estimate the number of observations that are less than or equal to a particular value.

Solved Example

Question: For the data given below, construct a less than cumulative frequency table and plot its ogive.

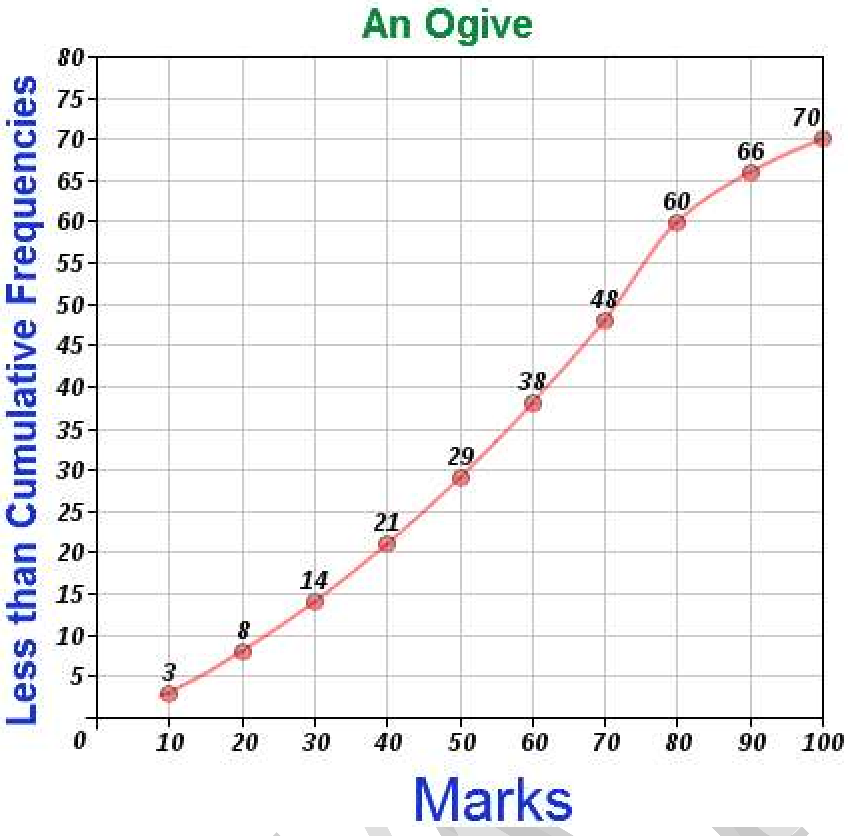
Marks	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70	70 - 80	80 - 90	90 -100
Frequency	3	5	6	7	8	9	10	12	6	4

Solution:

Marks	**Frequency**	**Less than cumulative frequency**
0 - 10	3	3
10 - 20	5	8
20 - 30	6	14
30 - 40	7	21
40 - 50	8	29
50 - 60	9	38
60 - 70	10	48
70 - 80	12	60
80 - 90	6	66
90 - 100	4	70

Plot the points having abscissa as upper limits and ordinates as the cumulative frequencies (10, 3), (20, 8), (30, 14), (40, 21), (50, 29), (60,38), (70, 48), (80, 60), (90, 66), (100, 70) and join the points by a smooth curve.

KAHE



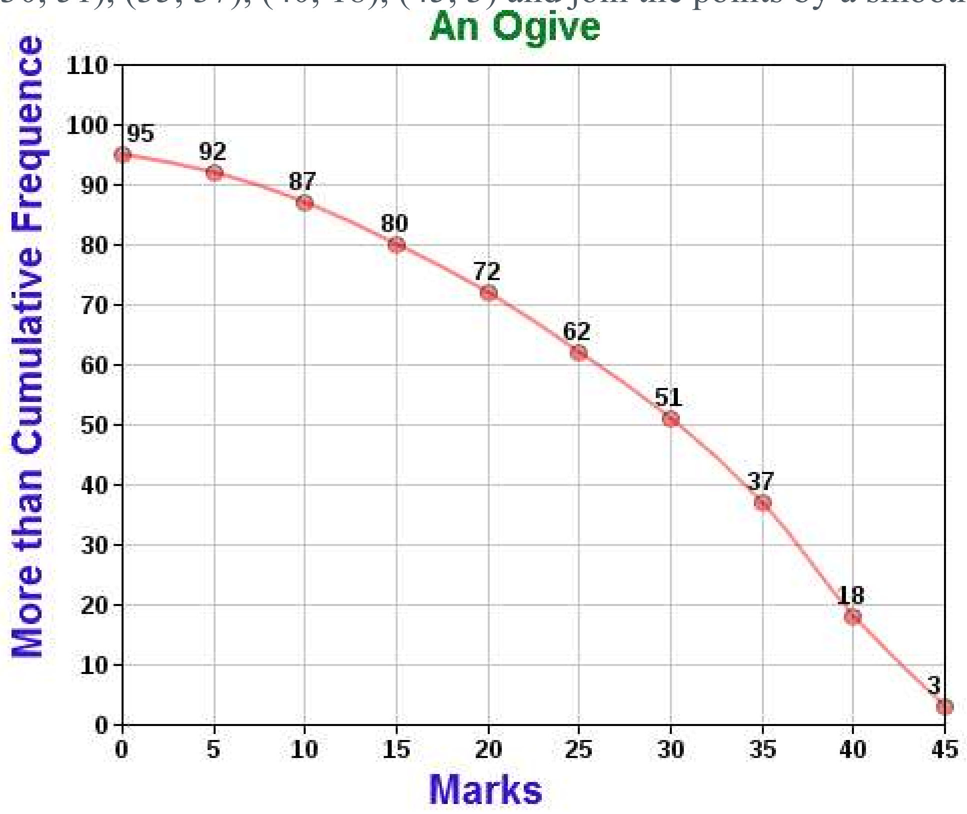
For the data given below, construct a more than cumulative frequency table and plot its ogive.

Marks	0 - 5	5 - 10	10 -15	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50
Frequency	3	5	7	8	10	11	14	19	15	13

Solution:

Marks	Frequency	More than cumulative frequency
0 - 5	3	95
5 - 10	5	95 - 3 = 92
10 -15	7	92 - 5 = 87
15 - 20	8	87 - 7 = 80
20 - 25	10	80 - 8 = 72
25 - 30	11	72 - 10 = 62
30 - 35	14	62 - 11 = 51
35 - 40	19	51 - 14 = 37
40 - 45	15	37 - 19 = 18
45 - 50	13	18 - 15 = 3

On the graph, plot the points (0, 95), (5, 92), (10, 87), (15, 80), (20, 72), (25, 62), (30, 51), (35, 37), (40, 18), (45, 3) and join the points by a smooth curve.



Question 2: Draw 'more than' and 'less than' ogive curves for the following data:

Class Interval	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50	50 - 55	55 - 60	60 - 65	65 - 70	70 - 75
Frequency	2	5	8	10	13	17	20	16	12	18	19	20

KAHE

Solution:
Let us calculate cumulative frequencies as follows:

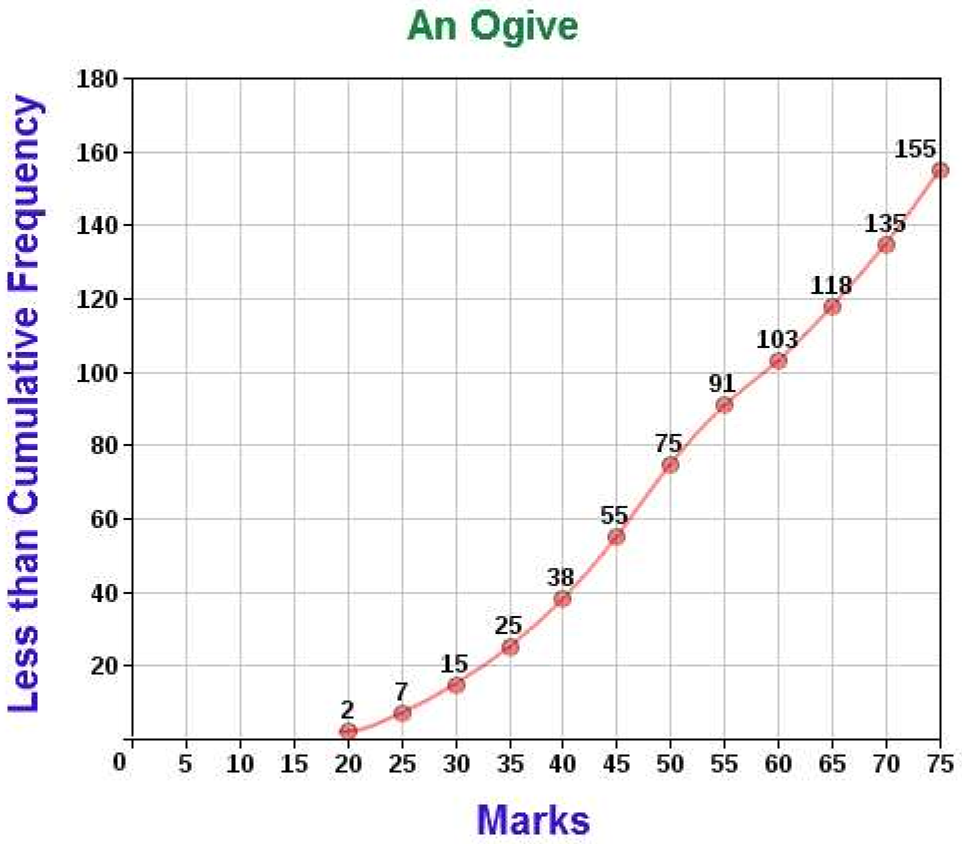
Class Interval	Frequency	Less than cumulative frequency	More than cumulative frequency
15 - 20	2	2	155
20 - 25	5	2 + 5 = 7	155 - 2 = 153
25 - 30	8	7 + 8 = 15	153 - 5 = 148
30 - 35	10	15 + 10 = 25	148 - 8 = 140
35 - 40	13	25 + 13 = 38	140 - 10 = 130
40 - 45	17	38 + 17 = 55	130 - 13 = 117
45 - 50	20	55 + 20 = 75	117 - 17 = 100
50 -55	16	75 + 16 = 91	100 - 20 = 80
55 - 60	12	91 + 12 = 103	80 -16 = 64
60 - 65	15	103 + 15 = 118	64 - 12 = 52
65 - 70	17	118 + 17 = 135	49 - 15 = 37
70 - 75	20	135 + 20 = 155	32 - 17 = 20

Less than ogive:

For **less than ogive**, plot the points, (20, 2), (25, 7), (30, 15), (35, 25), (40, 38), (45, 55), (50, 75), (55, 91), (60, 103), (65, 118), (70, 135), (75, 155) and join the points by a smooth curve.

Less than ogive plot for the given data is as follows:

KAHE

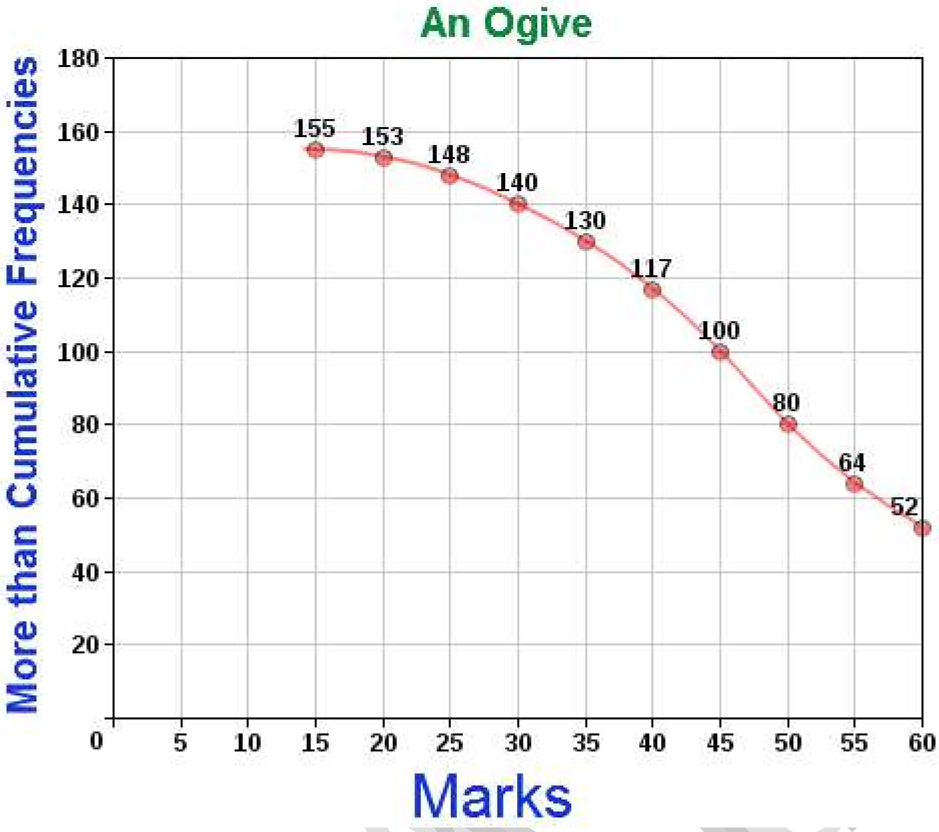


More than ogive:

For **more than ogive**, plot the points, (15, 155), (20, 153), (25, 148), (30, 140), (35, 130), (40, 117), (45, 100), (50, 80), (55, 64), (60, 52), (65, 37), (70, 20) and join the points by a smooth curve.

More than ogive plot for the given data is as follows:

KAHE


An Ogive
More than Cumulative Frequencies
180
160
140
120
100
80
60
40
20
0
5
10
15
20
25
30
35
40
45
50
55
60
155
153
148
140
130
117
100
80
64
52
Marks

KAHE

POSSIBLE QUESTIONS

PART-B(TWO MARKS)

1.Define classification.
2.Define Characteristics of classification.
3.Define Types of Diagrams.
4.Define objects of Classification.
5.Define Mutually exclusive.

PART-C (FIVE MARKS)

1.Define types of classification.
2.prepare a frequency table for the following data with width of each class interval as 10.Use exclusive method of classification.

57	44	80	75	00	18	45	14	04	64
72	51	69	34	22	83	70	20	57	28
96	56	50	47	10	34	61	66	80	46
22	10	84	50	47	73	42	33	48	65
10	34	66	53	75	90	58	46	39	69

3.Difference between classification and Tabulation.
4.The following data relate to the monthly expenditure (in rupees) of two families A and B:

	Expenditure (in Rs.)	
Items of Expenditure	Family A	Family B
Food	1600	1200
clothing	800	600
Rent	600	500
Light and Fuel	200	100
Miscellaneous	800	600

Represent the above data by a suitable percentage diagram
5.Describe about limitations of statistics.
6.A firm reported that its net worth in the year 1998-99 to 2002-03 was as follows:

Year	1998-99	1999-00	2000-01	2001-02	2002-03
Net worth	100	112	120	133	147

Plot the above data in the form of a semi graph.
7.Define Rules of Classification.

8.Draw a histogram for the following data :

Variable	100-110	110-120	120-130	130-140	140-150	150-160	160-170
Frequency	11	28	36	49	33	20	8

9.Describe the Types of Diagram.

10.Describe Diagrammatic and graphic representation of data.

PART-D(TEN MARKS)

1.Describe about general Rule of tabulation.

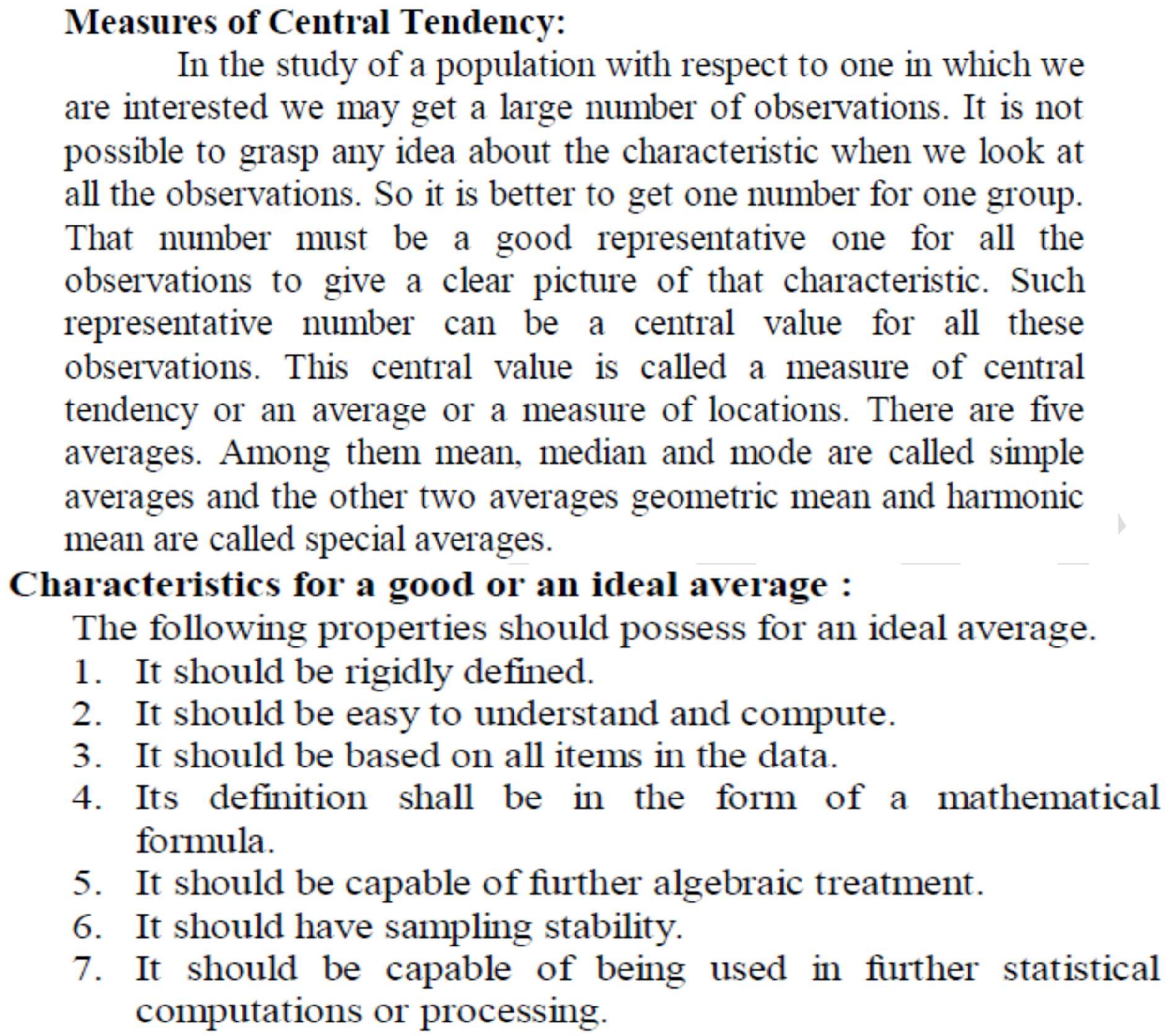
KAHE

UNIT – II

SYLLABUS

Measures of Central Tendency and Dispersion: Introduction, Objectives of statistical average, Requisites of a Good Average, Statistical Averages - Arithmetic mean - Properties of arithmetic mean - Merits and demerits of arithmetic mean ,Median - Merits and demerits of median , Mode - Merits and demerits of mode , Geometric Mean , Harmonic Mean , Positional Averages , Dispersion – Range - Quartile deviations, Mean deviation ,Standard Deviation -Properties of standard deviation Coefficient of Variance

KAHE



Measures of Central Tendency:

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

Characteristics for a good or an ideal average :

The following properties should possess for an ideal average.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. Its definition shall be in the form of a mathematical formula.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should be capable of being used in further statistical computations or processing.

Arithmetic mean or mean :

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

$$\bar{x} = \frac{x_1 + x_2 + x_3 + ... + x_n}{n}$$

$$= \frac{1}{n}\sum_{i=1}^{n} x_i$$

This formula is for the ungrouped or raw data.

Example 1 :

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

Solution:

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

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

Short-Cut method :

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

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

where, A = the assumed mean or any value in x
d = the deviation of each value from the assumed mean

Example 2 :

A student' s marks in 5 subjects are 75, 68, 80, 92, 56. Find his average mark.

Solution:

X	d=x-A
75	7
A $\boxed{68}$	0
80	12
92	24
56	-12
Total	31

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

$$= \quad 68 + \frac{31}{5}$$

$$= \ 68 + 6.2$$

$$= \ 74.2$$

Grouped Data :

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

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

where x = the mid-point of individual class
f = the frequency of individual class
N = the sum of the frequencies or total frequencies.

Short-cut method :

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

where $d = \dfrac{x - A}{c}$

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

Example 3:

Given the following frequency distribution, calculate the arithmetic mean

Marks	: 64	63	62	61	60	59
Number of Students	: 8	18	12	9	7	6

Solution:

X	F	fx	d=x-A	fd
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	–1	–9
60	7	420	–2	–14
59	6	354	–3	–18
	60	3713		- 7

Direct method

$$\bar{x} = \frac{\sum fx}{N} = \frac{3713}{60} = 61.88$$

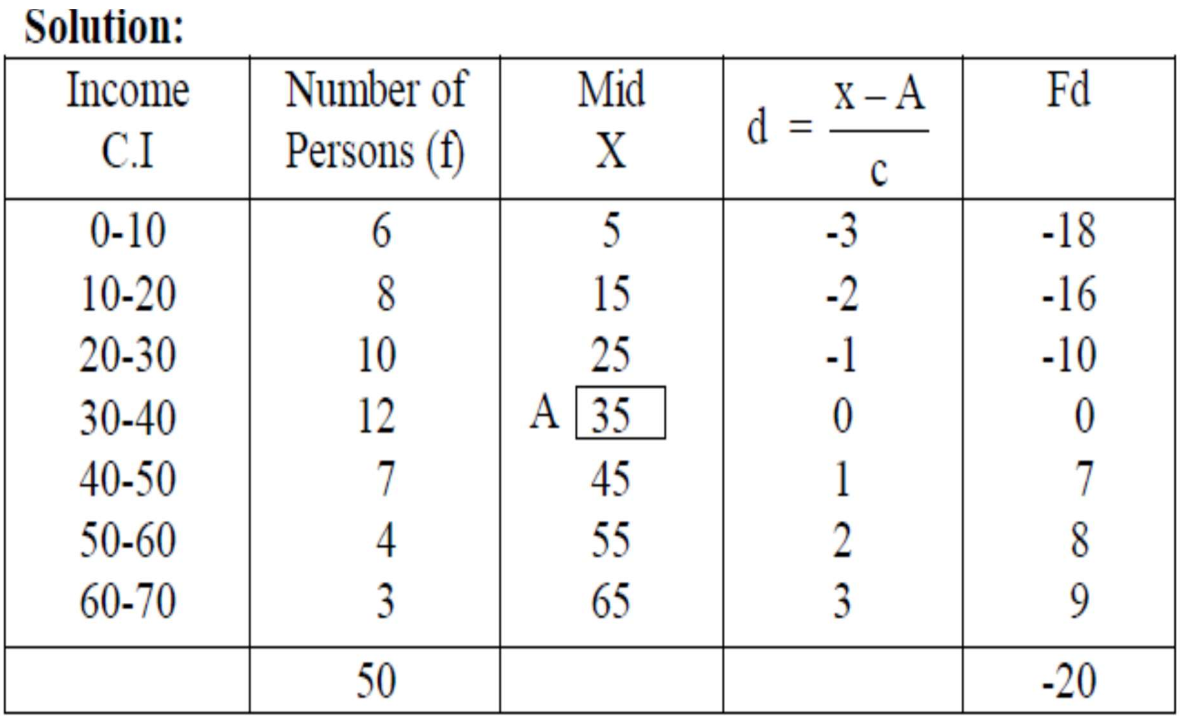
Short-cut method

$$\bar{x} = A + \frac{\sum fd}{N} = 62 - \frac{7}{60} = 61.88$$

Example 4 :

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

Income Rs(100)	0-10	10-20	20-30	30-40	40-50	50-60	60-70
Number of persons	6	8	10	12	7	4	3



Solution:

Income C.I	Number of Persons (f)	Mid X	$d = \frac{x - A}{c}$	Fd
0-10	6	5	-3	-18
10-20	8	15	-2	-16
20-30	10	25	-1	-10
30-40	12	A $\boxed{35}$	0	0
40-50	7	45	1	7
50-60	4	55	2	8
60-70	3	65	3	9
	50			-20

$$\text{Mean} = \bar{x} = A + \frac{\Sigma fd}{N}$$

$$= 35 - \frac{20}{50} \times 10$$

$$= 35 - 4$$
$$= 31$$

Merits and demerits of Arithmetic mean :

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series.
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

Demerits:

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given.

Median :

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

Ungrouped or Raw data :

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

By formula

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

Example 11:

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

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

Solution:

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

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

Using formula

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

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

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

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

$$= 25$$

Example 12 :

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

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

Solution:

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

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

$$= \left(\frac{10+12}{2}\right) = 11$$

$$\therefore \text{median} = 11.$$

Using the formula

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

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

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

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

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

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

$$= 10 + 1$$

$$= 11$$

Example 13:

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

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

Indicate in which subject is the level of knowledge higher ?

Solution:

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

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

Median = $\left(\frac{n+1}{2}\right)^{th}$ item = $\left(\frac{10+1}{2}\right)^{th}$ item =5.5th item

$$= \frac{Value\ of\ 5^{th}\ item + value\ of\ 6^{th}\ item}{2}$$

$$\text{Md (Statistics)} = \frac{46+47}{2} = 46.5$$

$$\text{Md (Accountancy)} = \frac{43+45}{2} = 44$$

There fore the level of knowledge in Statistics is higher than that in Accountancy.

Grouped Data:

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

Cumulative frequency : (cf)

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

Discrete Series:

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{N+1}{2}\right)$

Step3: See in the cumulative frequencies the value just greater than $\left(\frac{N+1}{2}\right)$

Step4: Then the corresponding value of x is median.

Example 14:

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

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

Solution:

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

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

$$= \text{size of}\left(\frac{60+1}{2}\right)^{\text{th}} \text{item}$$

$$= 30.5^{\text{th}} \text{ item}$$

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

Continuous Series:

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

Step1: Find cumulative frequencies.

Step2: Find $\left(\frac{N}{2}\right)$

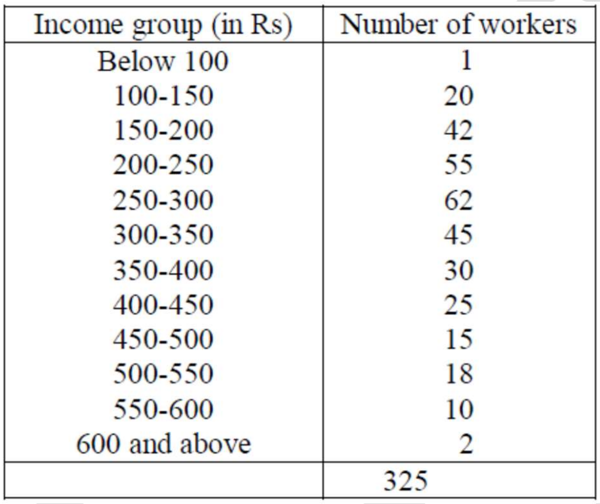
Step3: See in the cumulative frequency the value first greater than $\left(\frac{N}{2}\right)$, Then the corresponding class interval is called the Median class. Then apply the formula

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

Where l = Lower limit of the median class
m = cumulative frequency preceding the median
c = width of the median class
f =frequency in the median class.
N=Total frequency.

Example 15:

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



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

KAHE

Calculate median income

Solution:

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

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

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

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

$$= 250 + 35.89$$

$$= 285.89$$

Example 16*:*

Following are the daily wages of workers in a textile. Find the median.

Wages (in Rs.)	Number of workers
less than 100	5
less than 200	12
less than 300	20
less than 400	32
less than 500	40
less than 600	45
less than 700	52
less than 800	60
less than 900	68
less than 1000	75

Solution :

We are given upper limit and less than cumulative frequencies. First find the class-intervals and the frequencies. Since the values are increasing by 100, hence the width of the class interval equal to 100.

Class interval	f	c.f
0-100	5	5
100-200	7	12
200-300	8	20
300- 400	12	32
400-500	8	40
500-600	5	45
600-700	7	52
700-800	8	60
800-900	8	68
900-1000	7	75
	75	

$$\left(\frac{N}{2}\right) = \left(\frac{75}{2}\right) = 37.5$$

$$\text{Md} = l + \left(\frac{\frac{N}{2} - m}{f} \right) \times \text{c}$$

$$= 400 + \left(\frac{37.5 - 32}{8} \right) \times 100 = 400 + 68.75 = 468.75$$

Merits of Median :

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

Demerits of Median :

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

Mode :

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

Computation of the mode:

Ungrouped or Raw Data:

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

Example 29:

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

$\therefore$ Mode = $M_0 = 10$

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

Example 30:

1. 12, 10, 15, 24, 30 (no mode)
2. 7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10

$\therefore$ the modes are 7 and 10

Grouped Data:

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

Continuous distribution :

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

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

l = Lower limit of the model class

$$\triangle_1 = f_1 - f_0$$

$$\triangle_2 = f_1 - f_2$$

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

The above formula can also be written as

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

Example 31:

Calculate mode for the following :

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

Solution:

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

Here l=200,f_1=150,f_0=91, f_2=87, C=50

$$\text{Mode} = M_0 = 1 + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times c$$

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

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

$$= 200 + 24.18 = 224.18$$

Determination of Modal class :

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

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

Steps for Calculation :

We prepare a grouping table with 6 columns

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

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

Example 32:

Calculate mode for the following frequency distribution.

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

Grouping Table

C I	f	2	3	4	5	6
0- 5	9	21		36		
5-10	12		27		**43**	
10-15	15	31				**48**
15-20	16		**33**	**48**		
20-25	**17**	**32**			42	
25-30	15		25			38
30-35	10	23				
35-40	13					

Analysis Table

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

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

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

Here $l = 20$; $\Delta_1 = f_1 - f_0 = 17 - 16 = 1$

$\Delta_2 = f_1 - f_2 = 17 - 15 = 2$

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

$$= 20 + 1.67 = 21.67$$

MEASURES OF DISPERSION

Characteristics of a good measure of dispersion:

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

1.It should be rigidly defined

2. It should be based on all the items.

3. It should not be unduly affected by extreme items.

4. It should lend itself for algebraic manipulation.

5. It should be simple to understand and easy to calculate

Absolute and Relative Measures :

There are two kinds of measures of dispersion, namely

1.Absolute measure of dispersion

2.Relative measure of dispersion.

The various absolute and relative measures of dispersion are listed below.

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

7.3 Range and coefficient of Range:

7.3.1 Range:

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

In symbols, Range = L – S.

Where L = Largest value.

S = Smallest value.

In individual observations and discrete series, L and S are easily identified. In continuous series, the following two methods are followed.

Method 1:

L = Upper boundary of the highest class
S = Lower boundary of the lowest class.

Method 2:

L = Mid value of the highest class.
S = Mid value of the lowest class.

7.3.2 Co-efficient of Range :

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

Example1:

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

7, 9, 6, 8, 11, 10, 4

Solution:

L=11, S = 4.

Range = L – S = 11- 4 = 7

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

$$= \frac{11-4}{11+4}$$

$$= \frac{7}{15} = 0.4667$$

Example 2:

Calculate range and its co efficient from the following distribution.

Size:	60-63	63-66	66-69	69-72	72-75
Number:	5	18	42	27	8

Solution:

L = Upper boundary of the highest class.
= 75

S = Lower boundary of the lowest class.
= 60

Range = L – S = 75 – 60 = 15

$$\text{Co-efficient of Range} = \frac{L-S}{L+S} = \frac{75-60}{75+60} = \frac{15}{135} = 0.1111$$

7.3.3 Merits and Demerits of Range :

Merits:

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

Demerits:

1. It is very much affected by the extreme items.
2. It is based on only two extreme observations.
3. It cannot be calculated from open-end class intervals.
4. It is not suitable for mathematical treatment.
5. It is a very rarely used measure.

7.6 Standard Deviation and Coefficient of variation:

7.6.1 Standard Deviation :

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

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean.

The standard deviation is denoted by the Greek letter σ (sigma)

7.6.2 Calculation of Standard deviation-Individual Series :

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

a) Deviations taken from Actual mean

b) Deviation taken from Assumed mean

a) Deviation taken from Actual mean:

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

Steps:

1. Find out the actual mean of the series ($\bar{x}$)
2. Find out the deviation of each value from the mean $(x = X - \bar{X})$
3. Square the deviations and take the total of squared deviations $\sum x^2$

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

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

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

b) Deviations taken from assumed mean:

This method is adopted when the arithmetic mean is fractional value.

Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour, We apply short –cut method; deviations are taken from an assumed mean. The formula is:

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

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

Steps:

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

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

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

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

For the frequency distribution

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

KAHE

Example 9:

Calculate the standard deviation from the following data.

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

Solution:

Deviations from actual mean.

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

$$\overline{X} = \frac{120}{8} = 15$$

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

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

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

Example 10:

The table below gives the marks obtained by 10 students in statistics. Calculate standard deviation.

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

Solution: (Deviations from assumed mean)

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

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

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

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

$$= \sqrt{175.84} = 13.26$$

7.6.3 Calculation of standard deviation:

Discrete Series:

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

(a) Actual mean methods
(b) Assumed mean method
(c) Step-deviation method.

(a) Actual mean method:

Steps:

1. Calculate the mean of the series.
2. Find deviations for various items from the means i.e.,
   x- $\bar{x}$ = d.
3. Square the deviations (= d^2) and multiply by the respective frequencies(f) we get fd^2
4. Total to product (Σfd^2) Then apply the formula:

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

KA

(b) Assumed mean method:

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

Steps:

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

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

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

Example 11:

Calculate Standard deviation from the following data.

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

Solution:

Deviations from assumed mean

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

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

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

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

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

7.6.4 Calculation of Standard Deviation –Continuous series:

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

The formula is,

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

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

KAHE

Steps:

1. Find out the mid-value of each class.
2. Assume the center value as an assumed mean and denote it by A
3. Find out d $= \dfrac{m - A}{C}$
4. Multiply the deviations d by the respective frequencies and get Σfd
5. Square the deviations and get d^{2}
6. Multiply the squared deviations (d^{2}) by the respective frequencies and get Σfd^{2}
7. Substituting the values in the following formula to get the standard deviation

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

Example 13:

The daily temperature recorded in a city in Russia in a year is given below.

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

Calculate Standard Deviation.

Solution:

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

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

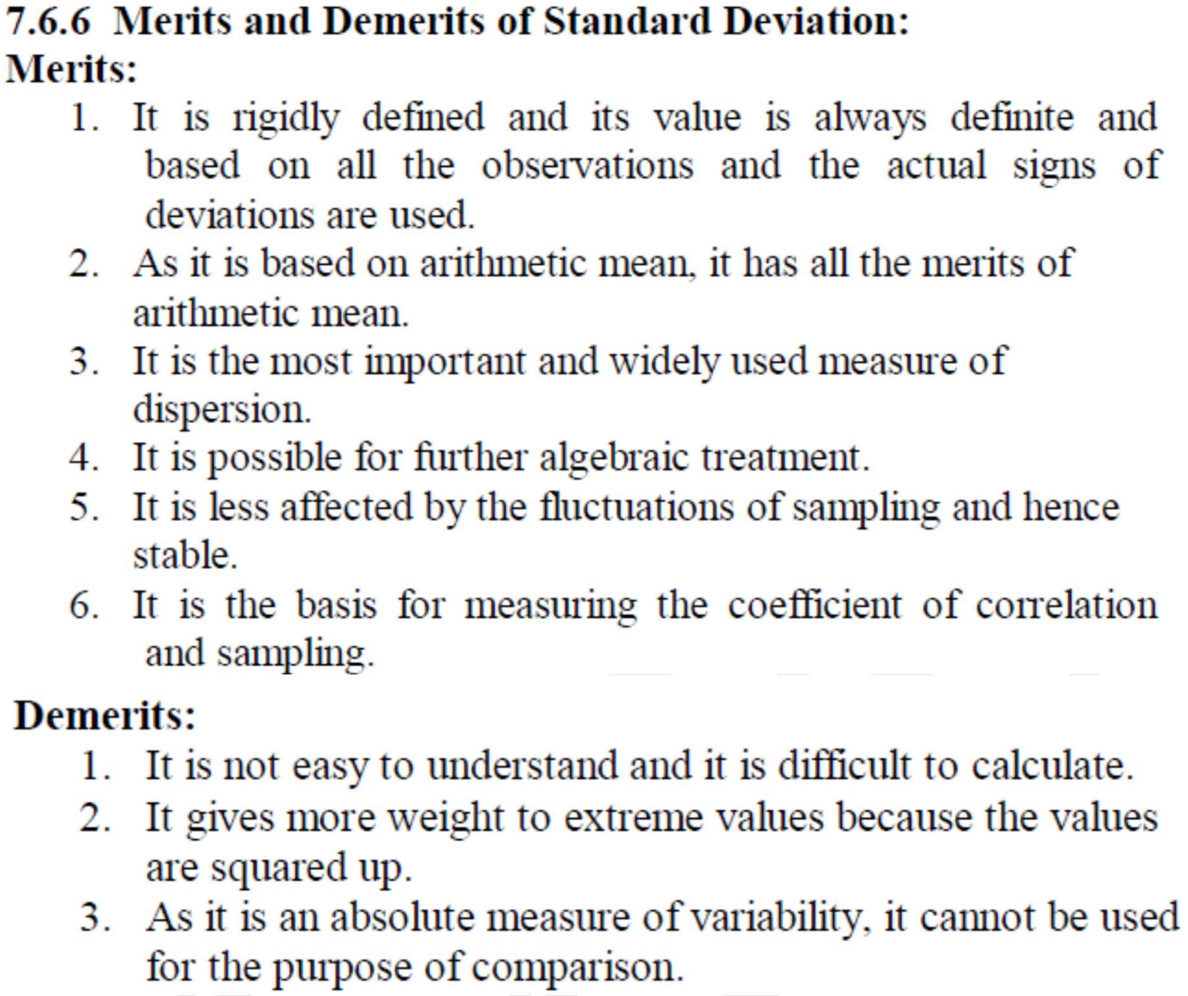
$$=\sqrt{\frac{1157}{365}-\left(\frac{389}{365}\right)^{2}}\times 10$$

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

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

$$=1.4262\times 10$$

$$=14.26^{\circ}c$$



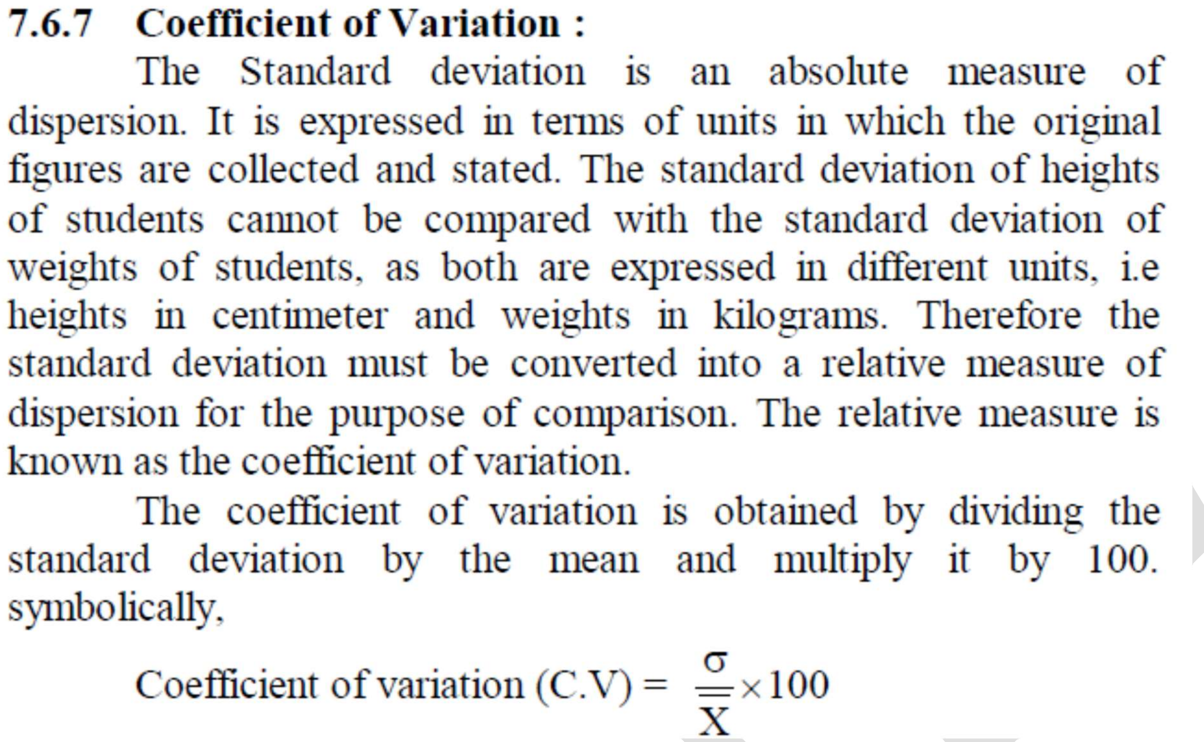
7.6.6 Merits and Demerits of Standard Deviation:

Merits:

1. It is rigidly defined and its value is always definite and based on all the observations and the actual signs of deviations are used.
2. As it is based on arithmetic mean, it has all the merits of arithmetic mean.
3. It is the most important and widely used measure of dispersion.
4. It is possible for further algebraic treatment.
5. It is less affected by the fluctuations of sampling and hence stable.
6. It is the basis for measuring the coefficient of correlation and sampling.

Demerits:

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



7.6.7 Coefficient of Variation :

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

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

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

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

Example 15:

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

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

1. Which factory A or B pays out a larger amount as weekly wages?
2. Which factory A or B has greater variability in individual wages?

Solution:

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

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

1. Total wages paid by factory A

$= 34.5 \times 476$

$=$ Rs.16.422

Total wages paid by factory B

$= 28.5 \times 524$

$=$ Rs.14,934.

Therefore factory A pays out larger amount as weekly wages.

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

KAHE

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

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

$$= 14.49$$

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

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

$$= 15.79$$

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

Example 16:

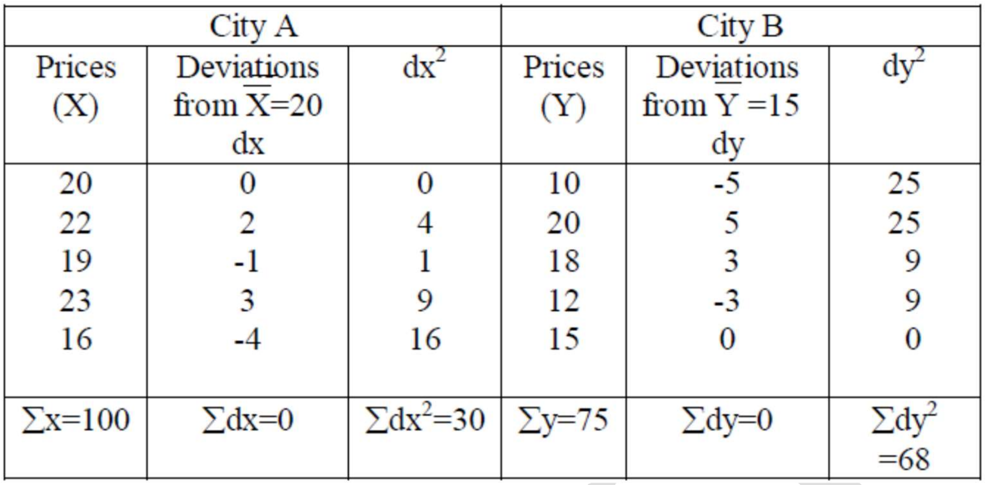
Prices of a particular commodity in five years in two cities are given below:

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

Which city has more stable prices?

Solution:

Actual mean method



City A			City B		
Prices (X)	Deviations from $\overline{X}$=20 dx	dx^2	Prices (Y)	Deviations from $\overline{Y}$ =15 dy	dy^2
20	0	0	10	-5	25
22	2	4	20	5	25
19	-1	1	18	3	9
23	3	9	12	-3	9
16	-4	16	15	0	0
$\sum x=100$	$\sum dx=0$	$\sum dx^2=30$	$\sum y=75$	$\sum dy=0$	$\sum dy^2$ =68

KAHE

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

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

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

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

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

$$= 12.25\ \%$$

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

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

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

$$\text{C.V.(y)} = \frac{\sigma_y}{\overline{y}} \text{ x } 100$$

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

$$= 24.6\ \%$$

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

POSSIBLE QUESTIONS

PART – B(TWO MARKS)

1.Defien standard Deviation.
2.Define Median.
3.Define Mode.
4.Find the range of weights of 7 students from the following 27,30,35,36,38,40,43.
5.Find the mode for the following Data.

850	750	600	825	850	725	600	850	640	530

PART – C(FIVE MARKS)

1.Calculate the geometric mean for the following data:

x :	12	13	14	15	16	17
f :	5	4	4	3	2	1

2.Find the standard deviation of the following distribution:

Age :	20-25	25-30	30-35	35-40	40-45	45-50
No of persons:	170	110	80	45	40	35

3.Calculate the Median for the following.

Marks	45-50	40-45	35-40	30-35	25-30	20-25	15-20	10-15	5-10
No.of students	10	15	26	30	42	31	24	15	7

4.Calculate the quartile deviation for the data given below

Daily Wages(Rs)	35-36	36-37	37-38	38-39	40-41	41-42	42-43
No.of wage earners	14	20	42	54	45	21	8

5.Calculate the Geometric Mean for the following Continuous Frequency Distribution.

Hourly Wages (in Rs.)	40 - 50	50 – 60	60 - 70	70 - 80	80 - 90	90 - 100
Number of Employees	10	20	15	30	15	10

6.Weekly Wages of a laborer are given below .Calculate Quartile Deviation and also the coefficient of Quartile Deviation

Marks	10	20	30	40	50	60	Total
No.of students	4	7	15	8	7	2	43

7.Calculate the mean and standard deviation for the following data.

X:	6	9	12	15	18
F:	7	12	13	10	8

8.Calculate mean and Standard Deviation of following frequency distribution of marks.

Marks	0-10	10-20	20-30	30-40	40-50	50-60	60-70
No.of students	5	12	30	45	50	37	21

9.Coefficient of variation of two series are 75% and 90% and their standard deviations are 15 and 18 respectively. Find their mean.

PART – D(TEN MARKS)

1.Find the interquartile range and the coefficient of quartile deviation from the following data.

Marks	Above 0	10	20	30	40	50	60	70	80
No.of students	150	140	100	80	80	70	30	14	0

KAHE

UNIT – III

Theory of Probability and Probability Distribution: Introduction - Definition of probability - Basic terminology used in probability theory, Approaches to probability , Rules of Probability - Addition rule - Multiplication rule , Conditional Probability, Steps Involved in Solving Problems on Probability , Bayes' Probability , Random Variables .

Introduction - Random variables , Probability Distributions - Discrete probability distributions - Continuous probability distributions , Bernoulli Distribution - t,Binomial Distribution - Poisson Distribution - Normal Distribution

KAHE

PROBABILITY

Introduction:

The theory of probability has its origin in the games of chance related to gambling such as tossing of a coin, throwing of a die, drawing cards from a pack of cards etc. Jerame Cardon, an Italian mathematician wrote ‘A book on games of chance’ which was published on 1663. Starting with games of chance, probability has become one of the basic tools of statistics. The knowledge of probability theory makes it possible to interpret statistical results, since many statistical procedures involve conclusions based on samples.

Probability theory is being applied in the solution of social, economic, business problems. Today the concept of probability has assumed greater importance and the mathematical theory of probability has become the basis for statistical applications in both social and decision-making research. Probability theory, in fact, is the foundation of statistical inferences.

Definitions and basic concepts:

The following definitions and terms are used in studying the theory of probability.

Trial:

Performing a random experiment is called a trial.

Random experiment:

Random experiment is one whose results depend on chance, that is the result cannot be predicted. Tossing of coins, throwing of dice are some examples of random experiments.

Outcomes:

The results of a random experiment are called its outcomes. When two coins are tossed the possible outcomes are HH, HT, TH, TT.

Event:

An outcome or a combination of outcomes of a random experiment is called an event. For example tossing of a coin is a random experiment and getting a head or tail is an event.

Sample space:

Each conceivable outcome of an experiment is called a sample point. The totality of all sample points is called a sample space and is denoted by **S**. For example, when a coin is tossed, the sample space is S = { H, T }. H and T are the sample points of the sample space S.

Equally likely events:

Two or more events are said to be equally likely if each one of them has an equal chance of occuring. For example in tossing of a coin, the event of getting a head and the event of getting a tail are equally likely events.

Mutually exclusive events:

Two or more events are said to be mutually exclusive, when the occurrence of any one event excludes the occurrence of the other event. Mutually exclusive events cannot occur simultaneously.

For example when a coin is tossed, either the head or the tail will come up. Therefore the occurrence of the head completely excludes the occurrence of the tail. Thus getting head or tail in tossing of a coin is a mutually exclusive event.

Exhaustive events:

Events are said to be exhaustive when their totality includes all the possible outcomes of a random experiment. For example, while throwing a die, the possible outcomes are {1, 2, 3, 4, 5, 6} and hence the number of cases is 6.

Complementary events:

The event ‘ A occurs’ and the event ‘ A does not occur’ are called complementary events to each other. The event ‘ A does not occur’ is denoted by A' or $\bar{A}$ or A^c. The event and its complements are mutually exclusive. For example in throwing a die, the event of getting odd numbers is { 1, 3, 5 } and getting even numbers is

{2, 4, 6}.These two events are mutually exclusive and complement to each other.

Independent events:

Events are said to be independent if the occurrence of one does not affect the others. In the experiment of tossing a fair coin, the occurrence of the event ' head' in the first toss is independent of the occurrence of the event ' head' in the second toss, third toss and subsequent tosses.

1.2 Definitions of Probability:

There are two types of probability. They are Mathematical probability and Statistical probability.

1.2.1 Mathematical Probability (or a priori probability):

If the probability of an event can be calculated even before the actual happening of the event, that is, even before conducting the experiment, it is called ***Mathematical probability.***

If the random experiments results in *'n'* exhaustive, mutually exclusive and equally likely cases, out of which *'m'* are favourable to the occurrence of an event A, then the ratio *m/n* is called the probability of occurrence of event A, denoted by P(A), is given by

$$P(A) = \frac{m}{n} = \frac{\text{Number of cases favourable to the event A}}{\text{Total number of exhaustive cases}}$$

Mathematical probability is often called ***classical probability*** or a ***priori probability*** because if we keep using the examples of tossing of fair coin, dice etc., we can state the answer in advance (*prior*), without tossing of coins or without rolling the dice etc..

The above definition of probability is widely used, but it cannot be applied under the following situations:

(1) If it is not possible to enumerate all the possible outcomes for an experiment.

(2) If the sample points(outcomes) are not mutually independent.

(3) If the total number of outcomes is infinite.

(4) If each and every outcome is not equally likely.

Some of the drawbacks of classical probability are removed in another definition given below:

1.2.2 Statistical Probability (or a posteriori probability):

If the probability of an event can be determined only after the actual happening of the event, it is called ***Statistical probability.***

If an event occurs ***m*** times out of ***n***, its relative frequency is ***m/n.***

In the limiting case, when ***n*** becomes sufficiently large it corresponds to a number which is called the probability of that event.

In symbol, $P(A) = \lim_{n\to\infty} (m/n)$

The above definition of probability involves a concept which has a long term consequence. This approach was initiated by the mathematician Von Mises .

If a coin is tossed 10 times we may get 6 heads and 4 tails or 4 heads and 6 tails or any other result. In these cases the probability of getting a head is **not 0.5** as we consider in Mathematical probability.

However, if the experiment is carried out a large number of times we should expect approximately equal number of heads and tails and we can see that the probability of getting head approaches 0.5. The Statistical probability calculated by conducting an actual experiment is also called a ***posteriori probability*** or ***empirical probability.***

1.2.3 Axiomatic approach to probability:

The modern approach to probability is purely axiomatic and it is based on the set theory. The axiomatic approach to probability was introduced by the Russian mathematician A.N. Kolmogorov in the year 1933.

Axioms of probability:

Let S be a sample space and A be an event in S and P(A) is the probability satisfying the following axioms:

(1) The probability of any event ranges from zero to one.
i.e $0 \leq P(A) \leq 1$

(2) The probability of the entire space is 1.
i.e $P(S) = 1$

(3) If A_1, A_2,…is a sequence of mutually exclusive events in S, then

$$P(A_1 \cup A_2 \cup ..) = P(A_1) + P(A_2) + \dots$$

Interpretation of statistical statements in terms of set theory:

$S \Rightarrow$ Sample space

$\bar{A} \Rightarrow$ A does not occur

$A \cup \bar{A} = S$

$A \cap B = \phi \Rightarrow$ A and B are mutually exclusive.

$A \cup B \Rightarrow$ Event A occurs or B occurs or both A and B occur. (at least one of the events A or B occurs)

$A \cap B \Rightarrow$ Both the events A and B occur.

$\bar{A} \cap \bar{B} \Rightarrow$ Neither A nor B occurs

$A \cap \bar{B} \Rightarrow$ Event A occurs and B does not occur

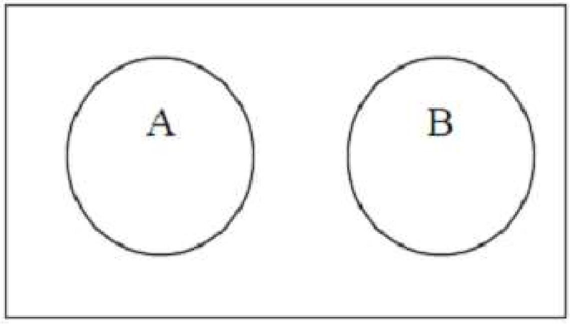
$\bar{A} \cap B \Rightarrow$ Event A does not occur and B occur.

1.3 Addition theorem on probabilities:

We shall discuss the addition theorem on probabilities for mutually exclusive events and not mutually exclusive events.

1.3.1 Addition theorem on probabilities for mutually exclusive events:

If two events A and B are mutually exclusive, the probability of the occurrence of either A or B is the sum of individual probabilities of A and B. ie $P(AUB) = P(A) + P(B)$
This is clearly stated in axioms of probability.



1.3.2 Addition theorem on probabilities for not-mutually exclusive events:

If two events A and B are not-mutually exclusive, the probability of the event that either A or B or both occur is given as

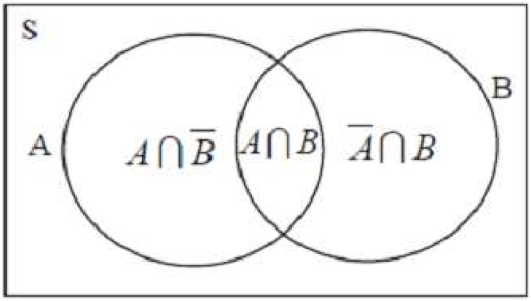
$$P(AUB) = P(A) + P(B) - P(A \cap B)$$

Proof:

Let us take a random experiment with a sample space S of N sample points.

Then by the definition of probability ,

$$P(AUB) = \frac{n(AUB)}{n(S)} = \frac{n(AUB)}{N}$$



From the diagram, using the axiom for the mutually exclusive events, we write

$$P(AUB) = \frac{n(A) + n(\overline{A} \cap B)}{N}$$

Adding and subtracting n($A \cap B$) in the numerator,

$$= \frac{n(A) + n(\overline{A} \cap B) + n(A \cap B) - n(A \cap B)}{N}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{N}$$

$$= \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N}$$

P(AUB) = P(A) + P(B) – P(A $\cap$ B)

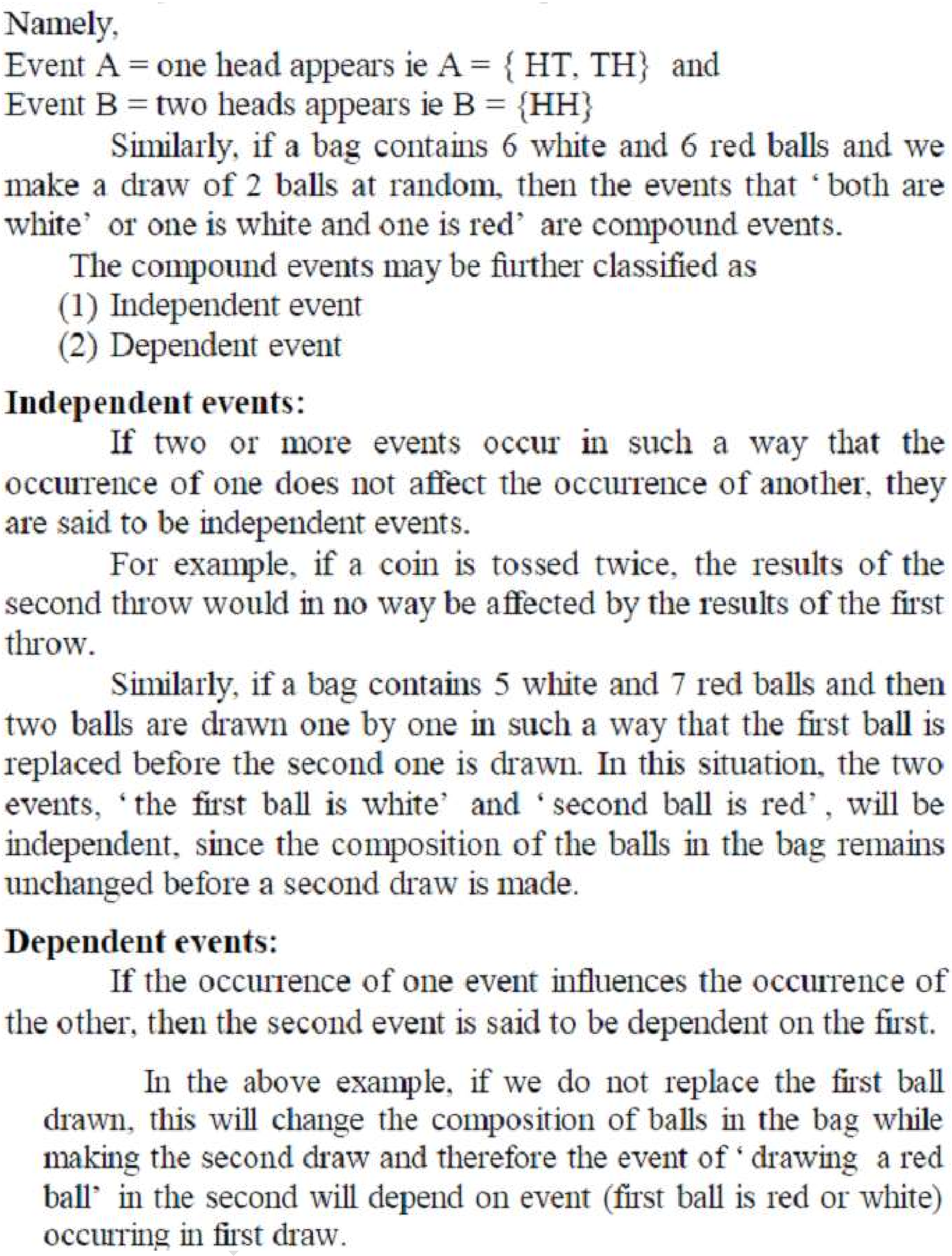
Note:

In the case of three events A,B,C, P(AUBUC) = P(A) + P(B) + P(C) – P($A \cap B$) – P($A \cap B$) – P($B \cap C$) + P ($A \cap B \cap C$)

Compound events:

The joint occurrence of two or more events is called compound events. Thus compound events imply the simultaneous occurrence of two or more simple events.

For example, in tossing of two fair coins simultaneously, the event of getting ‘ atleast one head’ is a compound event as it consists of joint occurrence of two simple events.



Namely,
Event A = one head appears ie A = { HT, TH} and
Event B = two heads appears ie B = {HH}

Similarly, if a bag contains 6 white and 6 red balls and we make a draw of 2 balls at random, then the events that ' both are white' or one is white and one is red' are compound events.

The compound events may be further classified as

(1) Independent event
(2) Dependent event

Independent events:

If two or more events occur in such a way that the occurrence of one does not affect the occurrence of another, they are said to be independent events.

For example, if a coin is tossed twice, the results of the second throw would in no way be affected by the results of the first throw.

Similarly, if a bag contains 5 white and 7 red balls and then two balls are drawn one by one in such a way that the first ball is replaced before the second one is drawn. In this situation, the two events, ' the first ball is white' and ' second ball is red' , will be independent, since the composition of the balls in the bag remains unchanged before a second draw is made.

Dependent events:

If the occurrence of one event influences the occurrence of the other, then the second event is said to be dependent on the first.

In the above example, if we do not replace the first ball drawn, this will change the composition of balls in the bag while making the second draw and therefore the event of ' drawing a red ball' in the second will depend on event (first ball is red or white) occurring in first draw.

Similarly, if a person draw a card from a full pack and does not replace it, the result of the draw made afterwards will be dependent on the first draw.

1.4 Conditional probability:

Let A be any event with p(A) >0. The probability that an event B occurs subject to the condition that A has already occurred is known as the conditional probability of occurrence of the event B on the assumption that the event A has already occurred and is denoted by the symbol P(B/A) or P(B|A) and is read as the probability of B given A.

The same definition can be given as follows also:

Two events A and B are said to be dependent when A can occur only when B is known to have occurred (or vice versa). The probability attached to such an event is called the **conditional probability** and is denoted by P(B/A) or, in other words, probability of B given that A has occurred.

If two events A and B are dependent, then the conditional probability of B given A is

$$P(B/A) = \frac{P(A \cap B)}{P(A)}$$

Similarly the conditional probability of A given B is given as

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Note:

If the events A and B are independent, that is the probability of occurrence of any one of them P(A/B) = P(A) and P(B/A) = P(B)

1.5 Multiplication theorem on probabilities:

We shall discuss multiplication theorem on probabilities for both independent and dependent events.

1.5.1 Multiplication theorem on probabilities for independent events:

If two events A and B are independent, the probability that both of them occur is equal to the product of their individual probabilities. i.e $P(A\cap B) = P(A) .P(B)$

Proof:

Out of n_1 possible cases let m_1 cases be favourable for the occurrence of the event A.

$$\therefore P(A) = \frac{m_1}{n_1}$$

Out of n_2 possible cases, let m_2 cases be favourable for the occurrence of the event B

$$\therefore P(B) = \frac{m_2}{n_2}$$

Each of n_1 possible cases can be associated with each of the n_2 possible cases.

Therefore the total number of possible cases for the occurrence of the event 'A' and 'B' is $n_1 \times n_2$. Similarly each of the m_1 favourable cases can be associated with each of the m_2 favourable cases. So the total number of favourable cases for the event 'A' and 'B' is $m_1 \times m_2$

$$\therefore P(A\cap B) = \frac{m_1\ m_2}{n_1\ n_2}$$

$$= \frac{m_1}{n_1} \cdot \frac{m_2}{n_2}$$

$$= P(A).P(B)$$

Note:

The theorem can be extended to three or more independent events. If A,B,C...... be independent events, then $P(A \cap B \cap C......) = P(A).P(B).P(C)......$

Note:

If A and B are independent then the complements of A and B are also independent. i.e $P(\overline{A} \cap \overline{B}) = P(\overline{A}) \, . \, P(\overline{B})$

1.5.2 Multiplication theorem for dependent events:

If A and B be two dependent events, i.e the occurrence of one event is affected by the occurrence of the other event, then the probability that both A and B will occur is

$$P(A \cap B) = P(A)\, P(B/A)$$

N
E
1
o
p

$$\therefore P(A \cap B) = \frac{m_1}{n}$$

$$= \frac{m_1}{n} \times \frac{m}{m} = \frac{m m_1}{n m}$$

$$= \frac{m}{n} \times \frac{m_1}{m}$$

$$\therefore P(A \cap B) = P(A)\ .\ P(B/A)$$

Note:

In the case of three events A, B, C, $P(A \cap B \cap C) = P(A)$. $P(B/A)$. $P(C/A \cap B)$. ie., the probability of occurrence of A, B and C is equal to the probability of A times the probability of B given that A has occurred, times the probability of C given that both A and B have occurred.

1.6 BAYES' Theorem:

The concept of conditional probability discussed earlier takes into account information about the occurrence of one event to

predict the probability of another event. This concept can be extended to revise probabilities based on new information and to determine the probability that a particular effect was due to specific cause. The procedure for revising these probabilities is known as Bayes theorem.

The Principle was given by Thomas Bayes in 1763. By this principle, assuming certain prior probabilities, the posteriori probabilities are obtained. That is why Bayes' probabilities are also called posteriori probabilities.

Bayes' Theorem or Rule (Statement only):

Let A_1, A_2, A_3, ...A_i, ...A_n be a set of n mutually exclusive and collectively exhaustive events and $P(A_1)$, $P(A_2)$. ., $P(A_n)$ are their corresponding probabilities. If B is another event such that P(B) is not zero and the priori probabilities $P(B|A_i)$ i =1,2. ., n are also known. Then

$$P(A_i | B) = \frac{P(B | A_i)\, P(A_i)}{\sum_{i=1}^{k} P(B | A_i)\, P(A_i)}$$

1.7 Basic principles of Permutation and Combination:

Factorial:

The consecutive product of first n natural numbers is known as ***factorial* n** and is denoted as **n!** or $\angle n$

That is $n! = 1 \times 2 \times 3 \times 4 \times 5 \times \ldots \times n$

$3! = 3 \times 2 \times 1$

$4! = 4 \times 3 \times 2 \times 1$

$5! = 5 \times 4 \times 3 \times 2 \times 1$

Also $5! = 5 \times (4 \times 3 \times 2 \times 1) = 5 \times (4!)$

Therefore this can be algebraically written as $n! = n \times (n - 1)!$

Note that $1! = 1$ and $0! = 1$.

Permutations:

Permutation means arrangement of things in different ways. Out of three things A, B, C taking two at a time, we can arrange them in the following manner.

A B BA

A C CA
B C CB

Here we find 6 arrangements. In these arrangements order of arrangement is considered. The arrangement AB and the other arrangement BA are different.

The number of arrangements of the above is given as the number of permutations of 3 things taken 2 at a time which gives the value 6. This is written symbolically, ${}_3P_2 = 6$

Thus the number of arrangements that can be made out of *n* things taken *r* at a time is known as the number of permutation of *n* things taken *r* at a time and is denoted as nPr.

The expansion of nPr is given below:

$nPr = n(n-1)(n-2)\ldots\ldots\ldots[n-(r-1)]$

The same can be written in factorial notation as follows:

$$nPr = \frac{n!}{(n-r)!}$$

For example, to find ${}_{10}P_3$ we write this as follows:

$$\begin{aligned} {}_{10}P_3 &= 10(10-1)(10-2) \\ &= 10 \times 9 \times 8 \\ &= 720 \end{aligned}$$

[To find ${}_{10}P_3$, Start with 10, write the product of 3 consecutive natural numbers in the descending order]

Simplifying ${}_{10}P_3$ using factorial notation:

$${}_{10}P_3 = \frac{10!}{(10-3)!} = \frac{10\times9\times8\times7\times6\times5\times4\times3\times2\times1}{7\times6\times5\times4\times3\times2\times1}$$

$$= 10 \times 9 \times 8$$
$$= 720$$

Note that $nPo = 1$, $nP_1 = n$, $nP_n = n!$

Combinations:

A combination is a selection of objects without considering the order of arrangements.

For example, out of three things A,B,C we have to select two things at a time.

This can be selected in three different ways as follows:
A B A C B C

Here the selection of the object A B and B A are one and the same. Hence the order of arrangement is not considered in combination. Here the number of combinations from 3 different things taken 2 at a time is 3.

This is written symbolically $_3C_2 = 3$

Thus the number of combination of n different things, taken r at a time is given by $nCr = \dfrac{n\,Pr}{r!}$

Or $nCr = \dfrac{n!}{(n-r)!r!}$

Note that $nC_0 = 1$, $nC_1 = n$, $nC_n = 1$

Find $_{10}C_3$. $_{10}C_3 = \dfrac{_{10}P_3}{3!} = \dfrac{10\times9\times8}{1\times2\times3} = 120$

Find $_8C_4$. $_8C_4 = \dfrac{8\times7\times6\times5}{1\times2\times3\times4} = 70$

[To find ${}_{8}C_{4}$: In the numerator, first write the product of 4 natural numbers starting with 8 in descending order and in the denominator write the factorial 4 and then simplify.]

Compare ${}_{10}C_{8}$ and ${}_{10}C_{2}$

$$ {}_{10}C_{8} = \frac{10\times9\times8\times7\times6\times5\times4\times3}{1\times2\times3\times4\times5\times6\times7\times8} = \frac{10\times9}{1\times2} = 45 $$

$$ {}_{10}C_{2} = \frac{10\times9}{1\times2} = 45 $$

From the above, we find ${}_{10}C_{8} = {}_{10}C_{2}$

This can be got by the following method also:

$${}_{10}C_{8} = {}_{10}C_{(10-8)} = {}_{10}C_{2}$$

This method is very useful, when the difference between n and r is very high in nCr.

This property of the combination is written as ${}_{n}C_{r} = {}_{n}C_{(n-r)}$.

To find ${}_{200}C_{198}$ we can use the above formula as follows:

$$ {}_{200}C_{198} = {}_{200}C_{(200-198)} = {}_{200}C_{2} = \frac{200\times199}{1\times2} = 19900. $$

Example:

Out of 13 players, 11 players are to be selected for a cricket team. In how many ways can this be done?

Out of 13 players, 11 players are selected in ${}_{13}C_{11}$ ways

$$ \text{i.e.} \quad {}_{13}C_{11} = {}_{13}C_{2} = \frac{13 \text{ x } 12}{1 \text{ x } 2} = 78. $$

Example 1:

Three coins are tossed simultaneously Find the probability that

(i) no head (ii) one head (iii) two heads
(iv) atleast two heads. (v) atmost two heads appear.

Solution:

The sample space for the 3 coins is

S = { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} ; n(S) = 8

(i) No head appear A = {TTT}; n(A) = 1

$$\therefore P(A) = \frac{1}{8}$$

(ii) One head appear B = {HTT, THT, TTH}; n (B) = 3

$$\therefore P(B) = \frac{3}{8}$$

(iii) Two heads appear C = {HHT, HTH, THH}; n(c)=3

$$\therefore P(C) = \frac{3}{8}$$

(iv) Atleast two heads appear
D = { HHT, HTH, THH, HHH}; n(D) = 4

$$\therefore P(D) = \frac{4}{8} = 1/2$$

(v) Atmost two heads appear E = { TTT, HTT, THT, TTH,HHT, HTH,THH}
n(E)= 7

$$\therefore P(E) = \frac{7}{8}$$

Example 2:

When two dice are thrown, find the probability of getting doublets (Same number on both dice)

Solution:

When two dice are thrown, the number of points in the sample space is n(S) = 36

Getting doublets: A = {(1,1) , (2,2) , (3,3) , (4,4) , (5,5) , (6,6)}

$$\therefore P(A) = \frac{6}{36} = \frac{1}{6}$$

Example 3:

A card is drawn at random from a well shuffled pack of 52 cards. What is the probability that it is (i) an ace (ii) a diamond card

Solution:

We know that the Pack contains 52 cards $\therefore$n(S)= 52

(i) There are 4 aces in a pack. n(A) = 4

$$\therefore P(A) = \frac{4}{52} = \frac{1}{13}$$

(ii) There are 13 diamonds in a pack $\therefore$n(B) = 13

$$\therefore P(B) = \frac{13}{52} = \frac{1}{4}$$

Example 4:

A ball is drawn at random from a box containing 5 green, 6 red, and 4 yellow balls. Determine the probability that the ball drawn is (i) green (ii) Red (iii) yellow (iv) Green or Red (v) not yellow.

Solution:

Total number of balls in the box = 5+6+4 = 15 balls

(i) Probability of drawing a green ball $= \frac{5}{15} = \frac{1}{3}$

(ii) Probability of drawing a red ball $= \frac{6}{15} = \frac{2}{5}$

(iii) Probability of drawing a yellow ball $= \frac{4}{15}$

(iv) Probability of drawing a Green or a Red ball $= \frac{5}{15} + \frac{6}{15} = \frac{11}{15}$

(v) Probability of getting not yellow = 1 – P (yellow)

$$= 1 - \frac{4}{15}$$

$$= \frac{11}{15}$$

Example 5:

Two dice are thrown, what is the probability of getting the sum being 8 or the sum being 10?

Solution:

Number of sample points in throwing two dice at a time is n(S)=36

Let A= {the sum being 8}

$\therefore$ A= {(6,2), (5,3) , (4,4), (3,5) , (2,6)}; $P(A) = \frac{5}{36}$

B = { the sum being 10}

$\therefore$ B = {(6,4), (5,5) (4,6)} ; $P(B) = \frac{3}{36}$

$A \cap B = \{ 0 \}$; $n(A \cap B) = 0$

$\therefore$ The two events are mutually exclusive

$$\therefore P(AUB) = P(A) + P(B)$$
$$= \frac{5}{36} + \frac{3}{36}$$
$$= \frac{8}{36} = \frac{2}{9}$$

Example 6 :

Two dice are thrown simultaneously. Find the probability that the sum being 6 or same number on both dice.

Solution:

n(S) = 36

The total is 6:

$\therefore$ A = {(5,1) , (4,2), (3,3) , (2,4) , (1,5)}; $P(A) = \frac{5}{36}$

Same number on both dice:

$\therefore B = \{(1,1)\ (2,2), (3,3), (4,4), (5,5), (6,6)\};\quad P(B) = \frac{6}{36}$

$A \cap B = \{(3,3)\}\ ;\qquad P(A \quad B) = \frac{1}{36}$

Here the events are not mutually exclusive.

$$\begin{aligned}
\therefore P(AUB) &= P(A) + P(B) - P(A \cap B) \\
&= \frac{5}{36} + \frac{6}{36} - \frac{1}{36} \\
&= \frac{5+6-1}{36} \\
&= \frac{11-1}{36} \\
&= \frac{10}{36} \\
&= \frac{5}{18}
\end{aligned}$$

Example 7:

Two persons A and B appeared for an interview for a job. The probability of selection of A is 1/3 and that of B is 1/2. Find the probability that

(i) both of them will be selected

(ii) only one of them will be selected

(iii) none of them will be selected

Solution:

$P(A)= \frac{1}{3}$, $P(B) = \frac{1}{2}$

$P(\overline{A}) = \frac{2}{3}$ and $P(\overline{B}) = \frac{1}{2}$

Selection or non-selection of any one of the candidate is not affecting the selection of the other. Therefore A and B are independent events.

(i) Probability of selecting both A and B

$$P(A \cap B) = P(A).P(B)$$
$$= \frac{1}{3} \times \frac{1}{2}$$
$$= \frac{1}{6}$$

(ii) Probability of selecting any one of them

= P (selecting A and not selecting B) + P(not selecting A and selecting B)

i.e $P\ (A \cap \overline{B}) + P\ (\overline{A} \cap B) = P(A).\ P(\overline{B}) + P(\overline{A}).\ P(B)$

$$= \frac{1}{3} \times \frac{1}{2} + \frac{2}{3} \times \frac{1}{2}$$
$$= \frac{1}{6} + \frac{2}{6}$$
$$= \frac{3}{6} \quad = \frac{1}{2}$$

(iii) Probability of not selecting both A and B

$$\text{i.e } P(\overline{A} \cap \overline{B})$$

$$= P(\overline{A}).\ P(\overline{B})$$

$$= \frac{2}{3} \cdot \frac{1}{2}$$

$$= \frac{1}{3}$$

Example 8:

There are three T.V programmes A , B and C which can be received in a city of 2000 families. The following information is available on the basis of survey.

1200 families listen to Programme A
1100 families listen to Programme B
800 families listen to Programme C
765 families listen to Programme A and B
450 families listen to Programme A and C
400 families listen to Programme B and C
100 families listen to Programme A, B and C

Find the probability that a family selected at random listens atleast one or more T.V Programmes.

Solution:

Total number of families n(S)= 2000

Let $n(A) = 1200$

$n(B) = 1100$

$n(C) = 800$

$$n(A \cap B) = 765$$
$$n(A \cap C) = 450$$
$$n(B \cap C) = 400$$
$$n(A \cap B \cap C) = 100$$

Let us first find $n(AUBUC)$.

$$n(AUBUC) = n(A) + n(B) + n(C) - n(A \cap B) - n(A \cap C) - n(B \cap C) + n(A \cap B \cap C)$$
$$= 1200 + 1100 + 800 - 765 - 450 - 400 + 100$$
$$n(AUBUC) = 1585$$

$$\text{now } P(AUBUC) = \frac{n(AUBUC)}{n(S)}$$
$$= \frac{1585}{2000} = 0.792$$

Therefore about 79% chance that a family selected at random listens to one or more T.V. Programmes.

Example 9:

A stockist has 20 items in a lot. Out of which 12 are non-defective and 8 are defective. A customer selects 3 items from the lot. What is the probability that out of these three items (i) three items are non-defective (ii) two are non defective and one is defective

Solution:

(i) Let the event, that all the three items are non-defective, be denoted by E_1. There are 12 non-defective items and out of them 3 can be selected in $12C_3$ ways ie $n(E_1)=12C_3$

Total number of ways in which 3 items can be selected are $20C_3$

i.e $n(S) = 20C_3$

$$\therefore\ P(E_1) = \frac{n(E_1)}{n(S)} = \frac{12C_3}{20C_3}$$

$$= \frac{12\times11\times10}{20\times19\times18}$$

$$= 0.193$$

ii) Let the event, that two items are non-defective and one is defective be denoted by E_2.

Two non-defective items out of 12 can be selected in$12C_2$ ways. One item out of 8 defective can be selected in $8C_1$ ways.

Thus $n(E_2) = 12C_2\ .\ 8C_1$

Then the probability $P(E_2) = \dfrac{n(E_2)}{n(S)} = \dfrac{12C_2\ .\ 8C_1}{20C_3}$

$$= \frac{12\times11\times8\times3}{20\times19\times18}$$

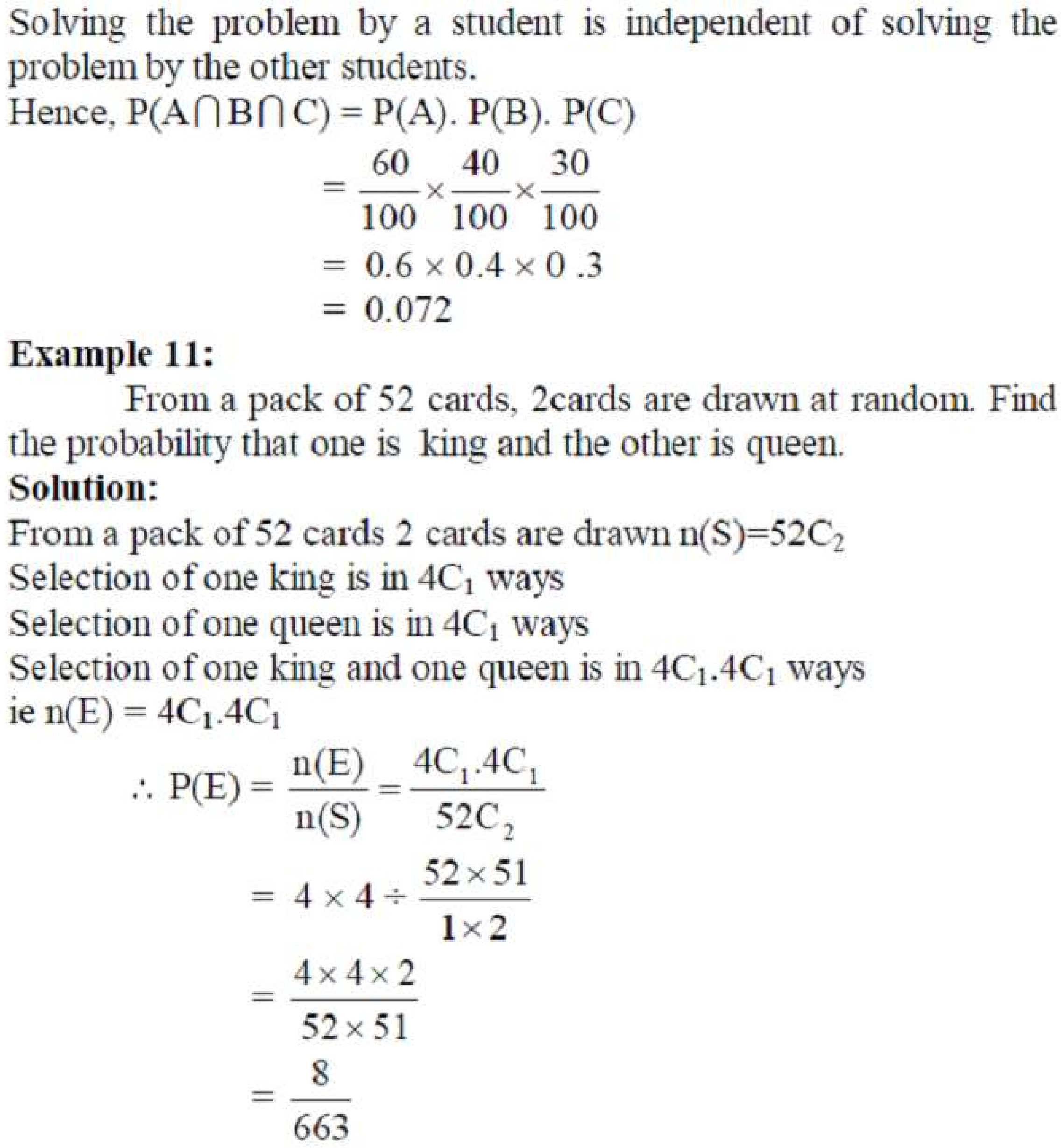
$$= 0.463$$

Example 10:

A test paper containing 10 problems is given to three students A,B,C. It is considered that student A can solve 60% problems, student B can solve 40% problems and student C can solve 30% problems. Find the probability that the problem chosen from the test paper will be solved by all the three students.

Solution:

Probability of solving the problem by A = 60%
Probability of solving the problem by B = 40%
Probability of solving the problem by C = 30%



Solving the problem by a student is independent of solving the problem by the other students.

Hence, $P(A\cap B\cap C) = P(A).\ P(B).\ P(C)$

$$= \frac{60}{100}\times\frac{40}{100}\times\frac{30}{100}$$

$$= 0.6\times 0.4\times 0\ .3$$

$$= 0.072$$

Example 11:

From a pack of 52 cards, 2cards are drawn at random. Find the probability that one is king and the other is queen.

Solution:

From a pack of 52 cards 2 cards are drawn $n(S)=52C_2$

Selection of one king is in $4C_1$ ways

Selection of one queen is in $4C_1$ ways

Selection of one king and one queen is in $4C_1.4C_1$ ways

ie $n(E) = 4C_1.4C_1$

$$\therefore\ P(E) = \frac{n(E)}{n(S)} = \frac{4C_1.4C_1}{52C_2}$$

$$= 4\times 4\div\frac{52\times 51}{1\times 2}$$

$$= \frac{4\times 4\times 2}{52\times 51}$$

$$= \frac{8}{663}$$

Example 12:

An urn contains 4 black balls and 6 white balls. If 3 balls are drawn at random, find the probability that (i) all are black (ii) all are white

Solution:

Total number of balls = 10

Total number ways of selecting 3 balls = $10C_3$

(i) Number of ways of drawing 3 black balls = $4C_3$

Probability of drawing 3 black balls = $\frac{4C_3}{10C_3}$

$$= \frac{4\times3\times2}{1\times2\times3} \div \frac{10\times9\times8}{1\times2\times3}$$

$$= \frac{4\times3\times2}{10\times9\times8}$$

$$= \frac{1}{30}$$

(ii) Number of ways of drawing 3 white balls = $6C_3$

Probability of drawing 3 white balls $= \frac{6C_3}{10C_3}$

$$= \frac{6\times5\times4}{10\times9\times8}$$

$$= \frac{1}{6}$$

Example 13:

A box containing 5 green balls and 3 red colour balls. Find the probability of selecting 3 green colour balls one by one (i) without replacement (ii) with replacement

Solution:

(i) Selection without replacement

Selecting 3 balls out of 8 balls = $8C_3$ ways

i.e $n(S) = 8C_3$

Selecting 3 green balls in $5C_3$ ways

$$\therefore \text{P(3 green balls)} = \frac{5C_3}{8C_3} = \frac{5\times4\times3}{8\times7\times6} = \frac{5}{28}$$

(ii) Selection with replacement

When a ball is drawn and replaced before the next draw, the number of balls in the box remains the same. Also the 3 events of drawing a green ball in each case is independent. $\therefore$Probability of drawing a green ball in each case is $\frac{5}{8}$

The event of selecting a green ball in the first, second and third event are same,

$\therefore$ Probability of drawing

$$\text{3 green balls} = \frac{5}{8} \times \frac{5}{8} \times \frac{5}{8} = \frac{125}{512}$$

Example 14:

A box contains 5 red and 4 white marbles. Two marbles are drawn successively from the box without replacement and it is noted that the second one is white. What is the probability that the first is also white?

Solution:

If w_1, w_2 are the events ' white on the first draw' , ' white on the second draw' respectively.

Now we are looking for $P(w_1/w_2)$

$$P(w_1/w_2) = \frac{P(w_1 \cap w_2)}{P(w_2)} = \frac{P(w_1).P(w_2)}{P(w_2)}$$

$$= \frac{(4/9)(3/8)}{(3/8)}$$

$$= \frac{4}{9}$$

Example 15:

A bag contains 6 red and 8 black balls. Another bag contains 7 red and 10 black balls. A bag is selected and a ball is drawn. Find the probability that it is a red ball.

Solution:

There are two bags

$$\therefore \text{ probability of selecting a bag} = \frac{1}{2}$$

Let A denote the first bag and B denote the second bag.

Then P(A) = P(B)= $\frac{1}{2}$

Bag ' A' contains 6 red and 8 black balls.

$\therefore$ Probability of drawing a red ball is $\frac{6}{14}$

Probability of selecting bag A and drawing a red ball from that bag is P(A). P(R/A) = $\frac{1}{2} \times \frac{6}{14} = \frac{3}{14}$

Similarly probability of selecting bag B and drawing a red ball from that bag is P(B). P(R/B) = $\frac{1}{2} \times \frac{7}{17} = \frac{7}{34}$

All these are mutually exclusive events

$\therefore$ Probability of drawing a red ball either from the bag A or B is

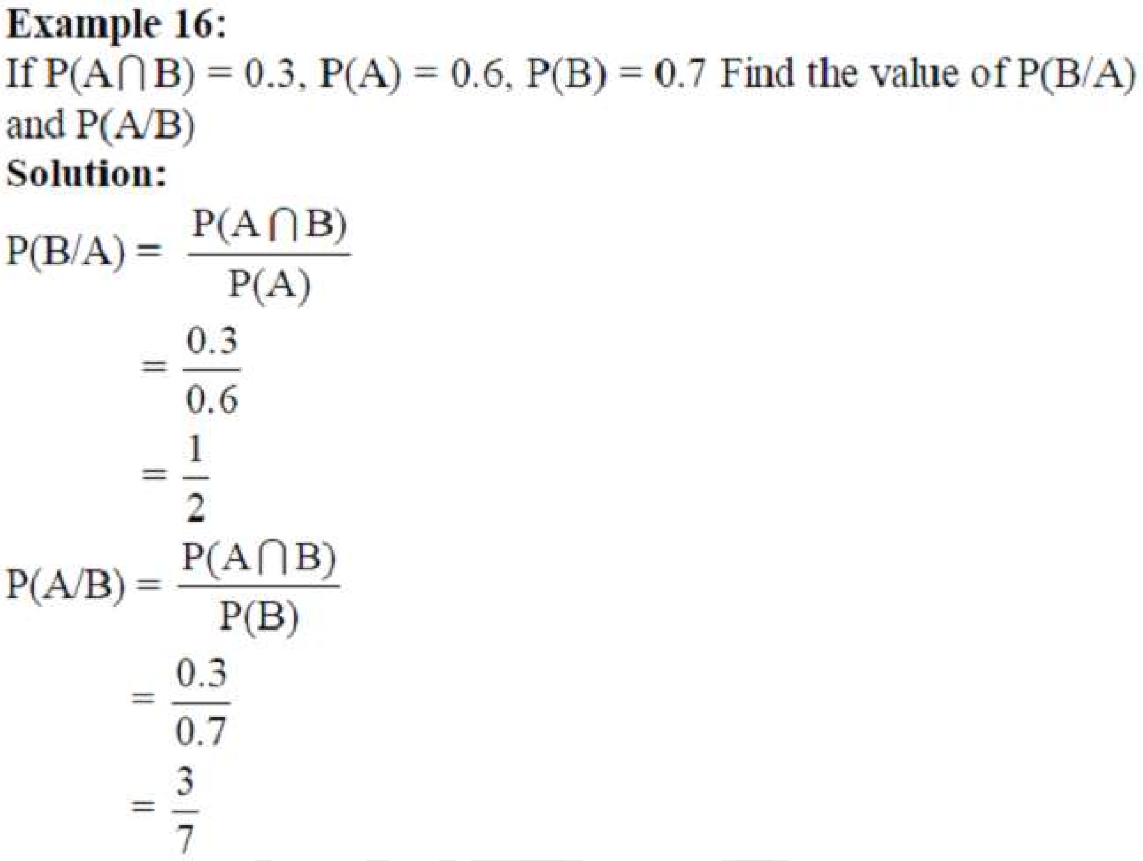
P(R) = P(A) P(R/A) + P(B) P(R/B)

$$= \frac{3}{14} + \frac{7}{34}$$

$$= \frac{17 \times 3 + 7 \times 7}{238}$$

$$= \frac{51 + 49}{238}$$

$$= \frac{100}{238} = \frac{50}{119}$$



Example 16:

If $P(A\cap B) = 0.3$, $P(A) = 0.6$, $P(B) = 0.7$ Find the value of $P(B/A)$ and $P(A/B)$

Solution:

$$P(B/A) = \frac{P(A\cap B)}{P(A)}$$

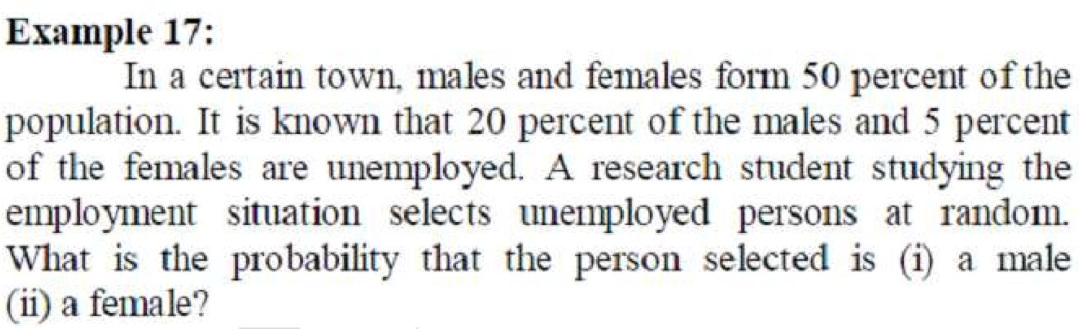
$$= \frac{0.3}{0.6}$$

$$= \frac{1}{2}$$

$$P(A/B) = \frac{P(A\cap B)}{P(B)}$$

$$= \frac{0.3}{0.7}$$

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



Example 17:

In a certain town, males and females form 50 percent of the population. It is known that 20 percent of the males and 5 percent of the females are unemployed. A research student studying the employment situation selects unemployed persons at random. What is the probability that the person selected is (i) a male (ii) a female?

Solution:

Out of 50% of the population 20% of the males are unemployed. i.e $\frac{50}{100} \times \frac{20}{100} = \frac{10}{100} = 0.10$

Out of 50% the population 5% of the females are unemployed.

$$\text{i.e } \frac{50}{100} \times \frac{5}{100} = \frac{25}{1000} = 0.025$$

Based on the above data we can form the table as follows:

	Employed	Unemployed	Total
Males	0.400	0.100	0.50
Females	0.475	0.025	0.50
Total	0.875	0.125	1.00

Let a male chosen be denoted by M and a female chosen be denoted by **F**

Let U denotes the number of unemployed persons then

(i) $P(M/U) = \frac{P(M \cap U)}{P(U)} = \frac{0.10}{0.125} = 0.80$

(ii) $P(F/U) = \frac{P(F \cap U)}{P(U)} = \frac{0.025}{0.125} = 0.20$

3.1 BINOMIAL DISTRIBUTION

3.1.0 Introduction:

In this chapter we will discuss the theoretical discrete distributions in which variables are distributed according to some definite probability law, which can be expressed mathematically. The Binomial distribution is a discrete distribution expressing the probability of a set of dichotomous alternative i.e., success or failure. This distribution has been used to describe a wide variety of process in business and social sciences as well as other areas.

3.1.1 Bernoulli Distribution:

A random variable X which takes two values 0 and 1 with probabilities q and p i.e., P(x=1) = p and P(x=0) = q, q = 1–p, is called a Bernoulli variate and is said to be a Bernoulli Distribution, where p and q takes the probabilities for success and failure respectively. It is discovered by Swiss Mathematician James Bernoulli (1654-1705).

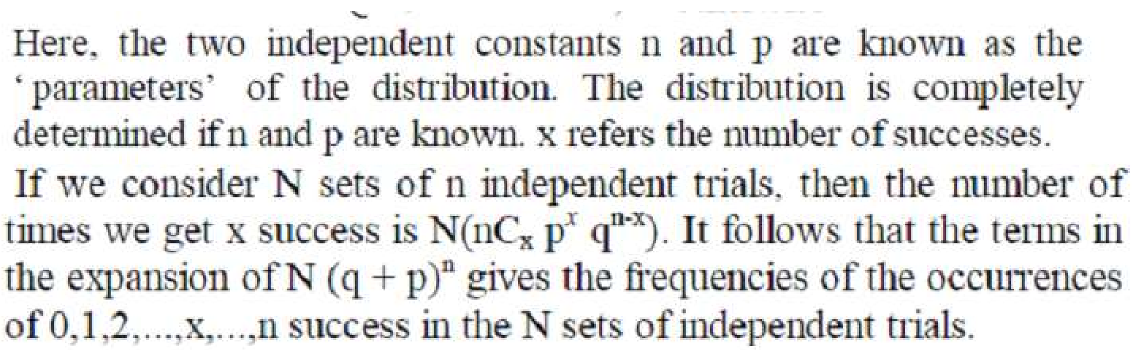
Examples of Bernoulli' s Trails are:

1) Toss of a coin (head or tail)
2) Throw of a die (even or odd number)
3) Performance of a student in an examination (pass or fail)

3.1.2 Binomial Distribution:

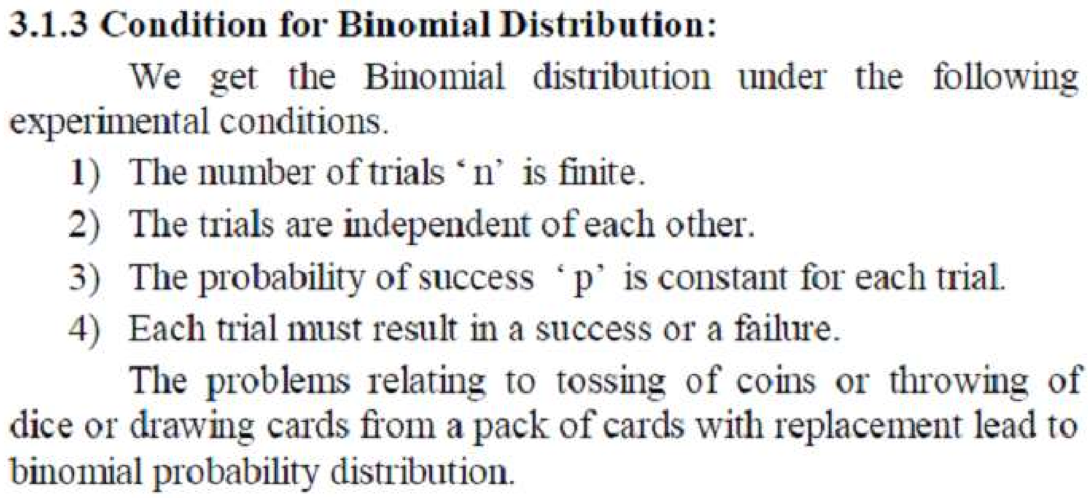
A random variable X is said to follow binomial distribution, if its probability mass function is given by

$$P\,(X = x) = P(x) = \begin{cases} nC_x\ p^x\, q^{n\text{-}x} & ;\ x = 0, 1,2, ..,n \\ 0 & ;\ \text{otherwise} \end{cases}$$



Here, the two independent constants n and p are known as the 'parameters' of the distribution. The distribution is completely determined if n and p are known. x refers the number of successes.

If we consider N sets of n independent trials, then the number of times we get x success is $N(nC_x\ p^x\ q^{n-x})$. It follows that the terms in the expansion of $N\ (q + p)^n$ gives the frequencies of the occurrences of 0,1,2,...,x,...,n success in the N sets of independent trials.

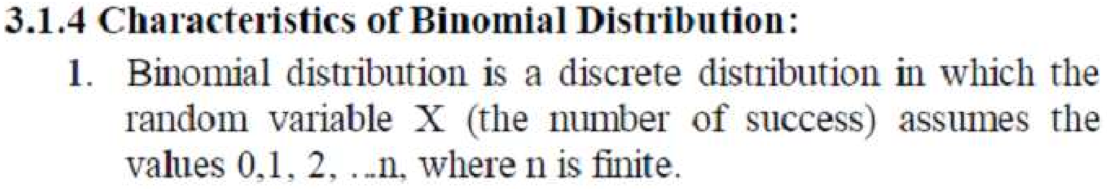


3.1.3 Condition for Binomial Distribution:

We get the Binomial distribution under the following experimental conditions.

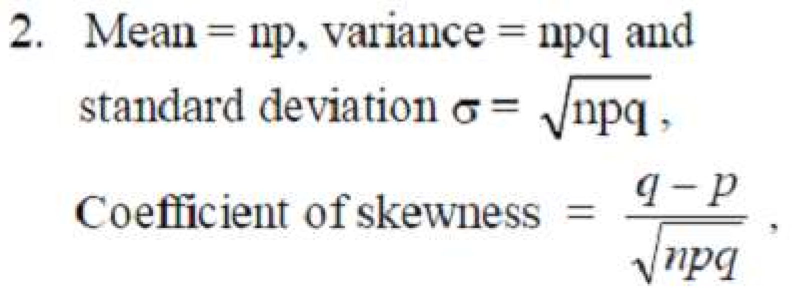
1) The number of trials 'n' is finite.
2) The trials are independent of each other.
3) The probability of success 'p' is constant for each trial.
4) Each trial must result in a success or a failure.

The problems relating to tossing of coins or throwing of dice or drawing cards from a pack of cards with replacement lead to binomial probability distribution.



3.1.4 Characteristics of Binomial Distribution:

1. Binomial distribution is a discrete distribution in which the random variable X (the number of success) assumes the values 0,1, 2, ..n, where n is finite.



2. Mean = np, variance = npq and

standard deviation $\sigma = \sqrt{npq}$,

Coefficient of skewness $= \dfrac{q-p}{\sqrt{npq}}$, Coefficient of kurtosis = $\frac{1 - 6pq}{npq}$, clearly each of the probabilities is non-negative and sum of all probabilities is 1 ($p < 1$, $q < 1$ and $p + q = 1$, $q = 1 - p$).

3. The mode of the binomial distribution is that value of the variable which occurs with the largest probability. It may have either one or two modes.
4. If two independent random variables X and Y follow binomial distribution with parameter (n_1, p) and (n_2, p) respectively, then their sum (X+Y) also follows Binomial distribution with parameter $(n_1 + n_2, p)$
5. If n independent trials are repeated N times, N sets of n trials are obtained and the expected frequency of x success is $N(nC_x\ p^x\ q^{n-x})$. The expected frequencies of 0,1,2…n success are the successive terms of the binomial distribution of $N(q + p)^n$.

Example 1:

Comment on the following: " The mean of a binomial distribution is 5 and its variance is 9"

Solution:

The parameters of the binomial distribution are n and p

We have mean $\Rightarrow np = 5$

Variance $\Rightarrow npq = 9$

$$\therefore q = \frac{npq}{np} = \frac{9}{5}$$

$$q = \frac{9}{5} > 1$$

Which is not admissible since q cannot exceed unity. Hence the given statement is wrong.

Example 2:

Eight coins are tossed simultaneously. Find the probability of getting atleast six heads.

Solution:

Here number of trials, n = 8, p denotes the probability of getting a head.

$$\therefore p = \frac{1}{2} \text{ and } q = \frac{1}{2}$$

If the random variable X denotes the number of heads, then the probability of a success in n trials is given by

$$P(X = x) = nc_x\, p^x q^{n\text{-}x}, \quad x = 0, 1, 2, \ldots, n$$

$$= 8C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{8-x} = 8C_x \left(\frac{1}{2}\right)^8$$

$$= \frac{1}{2^8}\, 8C_x$$

Probability of getting atleast six heads is given by

$$P(x \geq 6) = P(x = 6) + P(x = 7) + P(x = 8)$$

$$= \frac{1}{2^8}\, 8C_6 + \frac{1}{2^8}\, 8C_7 + \frac{1}{2^8}\, 8C_8$$

$$= \frac{1}{2^8}\,[\, 8C_6 + 8C_7 + 8C_8]$$

$$= \frac{1}{2^8}\,[\, 28 + 8 + 1] = \frac{37}{256}$$

Example 3:

Ten coins are tossed simultaneously. Find the probability of getting (i) atleast seven heads (ii) exactly seven heads (iii) atmost seven heads

Solution:

$$p = \text{Probability of getting a head} = \frac{1}{2}$$

$$q = \text{Probability of not getting a head} = \frac{1}{2}$$

The probability of getting x heads throwing 10 coins simultaneously is given by

$P(X = x) = nC_x\, p^x\, q^{n-x}.\quad ,\quad x = 0, 1, 2, \dots, n$

$$= 10C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{10-x} = \frac{1}{2^{10}}\, 10C_x$$

i) Probability of getting atleast seven heads

$$P(x \geq 7) = P(x = 7) + P(x = 8) + P(x = 9) + P(x = 10)$$

$$= \frac{1}{2^{10}} [\, 10C_7 + 10C_8 + 10C_9 + 10C_{10}]$$

$$= \frac{1}{1024} [\, 120 + 45 + 10 + 1] = \frac{176}{1024}$$

ii) Probability of getting exactly 7 heads

$$P(\, x = 7) = \frac{1}{2^{10}}\, 10C_7 = \frac{1}{2^{10}} (120)$$

$$= \frac{120}{1024}$$

iii) Probability of getting atmost 7 heads

$$P(x \le 7) = 1 - P(x > 7)$$
$$= 1 - \{P(x = 8) + P(x = 9) + P(x = 10)\}$$
$$= 1 - \frac{1}{2^{10}}\{10C_8 + 10C_9 + 10C_{10}\}$$
$$= 1 - \frac{1}{2^{10}}[45 + 10 + 1]$$
$$= 1 - \frac{56}{1024}$$
$$= \frac{968}{1024}$$

Example 4:

20 wrist watches in a box of 100 are defective. If 10 watches are selected at random, find the probability that (i) 10 are defective (ii) 10 are good (iii) at least one watch is defective (iv) at most 3 are defective.

Solution:

20 out of 100 wrist watches are defective

Probability of defective wrist watch , $p = \frac{20}{100} = \frac{1}{5}$

$$\therefore q = 1 - p = \frac{4}{5}$$

Since 10 watches are selected at random, n =10

$$P(X = x) = nC_x\, p^x\, q^{n-x} \quad , \quad x = 0 , 1, 2, \ldots,10$$

$$= 10C_x \left(\frac{1}{5}\right)^x \left(\frac{4}{5}\right)^{10-x}$$

i) Probability of selecting 10 defective watches

$$P(x =10) = 10C_{10} \left(\frac{1}{5}\right)^{10} \left(\frac{4}{5}\right)^{0}$$

$$= 1.\ \frac{1}{5^{10}}.1 \quad = \quad \frac{1}{5^{10}}$$

ii) Probability of selecting 10 good watches (i.e. no defective)

$$P(x = 0) = 10C_0 \left(\frac{1}{5}\right)^{0} \left(\frac{4}{5}\right)^{10}$$

$$= 1.1. \left(\frac{4}{5}\right)^{10} = \left(\frac{4}{5}\right)^{10}$$

iii) Probability of selecting at least one defective watch

$$P(x \geq 1) = 1 - P(x < 1)$$

$$= 1 - P(x = 0)$$

$$= 1 - 10C_0 \left(\frac{1}{5}\right)^{0} \left(\frac{4}{5}\right)^{10}$$

$$= 1 - \left(\frac{4}{5}\right)^{10}$$

iv) Probability of selecting at most 3 defective watches

$$P\ (x \le 3) = P\ (x = 0) + P(x = 1) + P(x = 2) + P(x = 3)$$

$$= 10C_0\left(\frac{1}{5}\right)^0 \left(\frac{4}{5}\right)^{10} + 10C_1\left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 + 10C_2\left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8$$

$$+ 10C_3 \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7$$

$$= 1.1. \left(\frac{4}{5}\right)^{10} + 10\left(\frac{1}{5}\right)^1 \left(\frac{4}{5}\right)^9 + \frac{10.9}{1.2} \left(\frac{1}{5}\right)^2 \left(\frac{4}{5}\right)^8$$

$$+ \frac{10.9.8}{1.2.3} \left(\frac{1}{5}\right)^3 \left(\frac{4}{5}\right)^7$$

$$= 1.\ (0.107) + 10\ (0.026) + 45\ (0.0062) + 120\ (0.0016)$$

$$= 0.859 \text{ (approx)}$$

Example 5:

With the usual notation find p for binomial random variable X if n = 6 and 9P(X = 4) = P(X = 2)

Solution:

The probability mass function of binomial random variable X is given by

$P(X = x) = nC_x\, p^x\, q^{n-x}.$, $x = 0, 1, 2, \ldots, n$

Here n = 6 $\therefore$ $P(X = x) = 6C_x\, p^x\, q^{6-x}$

$$P\ (x = 4) = 6C_4\, p^4\, q^2$$

$$P\ (x = 2) = 6C_2\, p^2\, q^4$$

Given that,

$$9.\ P(x = 4) = P(x = 2)$$

$$9.\ 6C_4\, p^4 q^2 = 6C_2\, p^2 q^4$$

$$\Rightarrow 9 \times 15p^2 = 15q^2$$

$$9p^2 = q^2$$

Taking positive square root on both sides we get,

$$\begin{aligned} 3p &= q \\ &= 1-p \\ 4p &= 1 \\ \therefore p &= \frac{1}{4} = 0.25 \end{aligned}$$

3.1.5 Fitting of Binomial Distribution:

When a binomial distribution is to be fitted to an observed data, the following procedure is adopted.

1. Find Mean $= \bar{x} = \dfrac{\Sigma fx}{\Sigma f} = np$

   $$\Rightarrow p = \frac{\bar{x}}{n} \text{ where n is number of trials}$$

2. Determine the value, $q = 1 - p$.
3. The probability function is $P(x) = {}_nC_x\, p^x q^{n-x}$ put $x = 0$, we set $P(0) = q^n$ and $f(0) = N \times P(0)$
4. The other expected frequencies are obtained by using the recurrence formula is given by

   $$f(x+1) = \frac{n-x}{x+1}\,\frac{p}{q}\,f(x)$$

Example 6:

A set of three similar coins are tossed 100 times with the following results

Number of heads :	0	1	2	3
Frequency :	36	40	22	2

Solution:

X	f	fx
0	36	0
1	40	40
2	22	44
3	2	6
	$\Sigma f = 100$	$\Sigma fx =$ 90

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{90}{100} = 0.9$$

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

$$= \frac{0.9}{3} = 0.3$$

$$q = 1 - 0.3$$

$$= 0.7$$

The probability function is $P(x) = nC_x\, p^x\, q^{n-x}$

Here $n = 3,\ p = 0.3\ q = 0.7$

$$\therefore P(x) = 3C_x\,(0.3)^x\,(0.7)^{3-x}$$

$$P(0) = 3C_0\,(0.3)^0\,(0.7)^3$$

$$= (0.7)^3 \quad = 0.343$$

$\therefore\ f(0) = N \times P(0) = 0.343 \times 100 = 34.3$

The other frequencies are obtained by using the recurrence formula

$f(x+1) = \dfrac{n-x}{x+1}\left(\dfrac{p}{q}\right) f(x)$. By putting $x = 0, 1, 2$ the expected frequencies are calculated as follows.

$$f(1) = \frac{3-0}{0+1}\left(\frac{p}{q}\right) \times 34.3$$

$$= 3 \times (0.43) \times 34.3 \; = \; 44.247$$

$$f(2) = \frac{3-1}{1+1}\left(\frac{p}{q}\right) \; f(1)$$

$$= \frac{2}{2}\,(0.43) \times 44.247$$

$$= 19.03$$

$$f(3) = \frac{3-2}{2+1}\left(\frac{p}{q}\right) f(2)$$

$$= \frac{1}{3}(0.43) \times 19.03$$

$$= 2.727$$

The observed and theoretical (expected) frequencies are tabulated below:

					Total
Observed frequencies	36	40	22	2	100
Expected frequencies	34	44	19	3	100

Example 7:

4 coins are tossed and number of heads noted. The experiment is repeated 200 times and the following distribution is obtained .

x: Number of heads	0	1	2	3	4
f: frequencies	62	85	40	11	2

Solution:

X	0	1	2	3	4	Total
f	62	85	40	11	2	200
fx	0	85	80	33	8	206

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{206}{200} = 1.03$$

$$p = \frac{\bar{x}}{n} = \frac{1.03}{4} = 0.2575$$

$$\therefore q = 1 - 0.2575 = 0.7425$$

Here $n = 4$, $p = 0.2575$; $q = 0.7425$

The probability function of binomial distribution is

$$P(x) = nC_x \; p^x q^{n-x}$$

The binomial probability function is

$$P(x) = 4C_x \; (0.2575)^x (0.7425)^{4-x}$$

$$P(0) = (0.7425)^4$$

$$= 0.3039$$

$$\therefore \; f(0) = NP(0)$$

$$= 200 \times 0.3039$$

$$= 60.78$$

The other frequencies are calculated using the recurrence formula $f(x+1) = \frac{n-x}{x+1}\left(\frac{p}{q}\right) f(x)$. By putting x = 0,1, 2, 3 then the expected frequencies are calculated as follows:

Put x = 0, we get

$$f(1) = \frac{4-0}{0+1}(0.3468)\,(60.78)$$
$$= 84.3140$$

$$f(2) = \frac{4-1}{1+1}\,(0.3468)\,(84.3140)$$
$$= 43.8601$$

$$f(3) = \frac{4-2}{2+1}\,(0.3468)\,(43.8601)$$
$$= 10.1394$$

$$f(4) = \frac{4-3}{3+1}\,(0.3468)\,(10.1394)$$
$$= 0.8791$$

The theoretical and expected frequencies are tabulated below:

						Total
Observed frequencies	62	85	40	11	2	200
Expected frequencies	61	84	44	10	1	200

3.2 POISSON DISTRIBUTION:

3.2.0 Introduction:

Poisson distribution was discovered by a French Mathematician-cum-Physicist Simeon Denis Poisson in 1837. Poisson distribution is also a discrete distribution. He derived it as a limiting case of Binomial distribution. For n-trials the binomial distribution is $(q + p)^n$; the probability of x successes is given by $P(X=x) = nC_x p^x q^{n-x}$. If the number of trials n is very large and the probability of success ' p' is very small so that the product np = m is non – negative and finite.

The probability of *x* success is given by

$$P(X = x) = \begin{cases} \dfrac{e^{-m} m^x}{x!} & \text{for } x = 0,1,2, \ldots \\ 0 & ; \text{ otherwise} \end{cases}$$

Here m is known as parameter of the distribution so that m >0

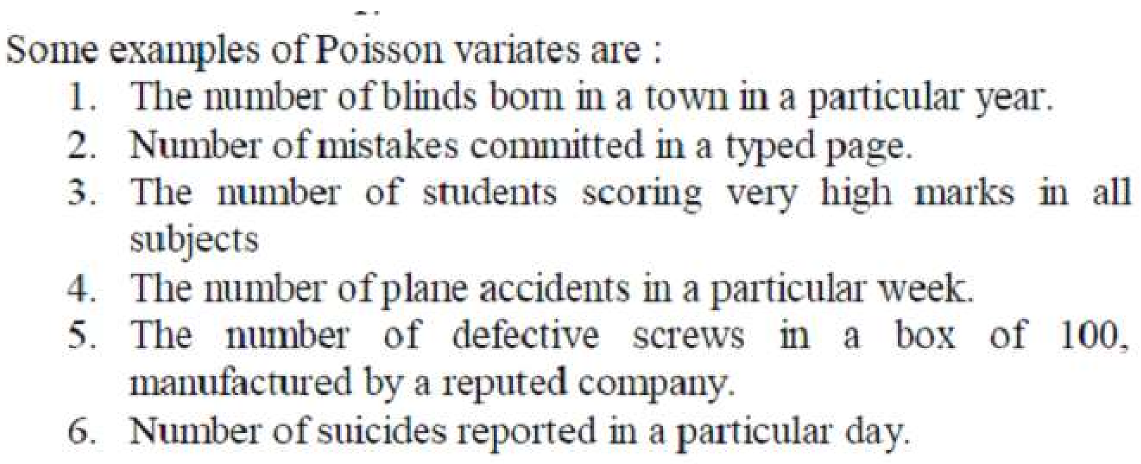
Since number of trials is very large and the probability of success p is very small, it is clear that the event is a rare event. Therefore Poisson distribution relates to rare events.

Note:

1) e is given by $e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots = 2.71828$

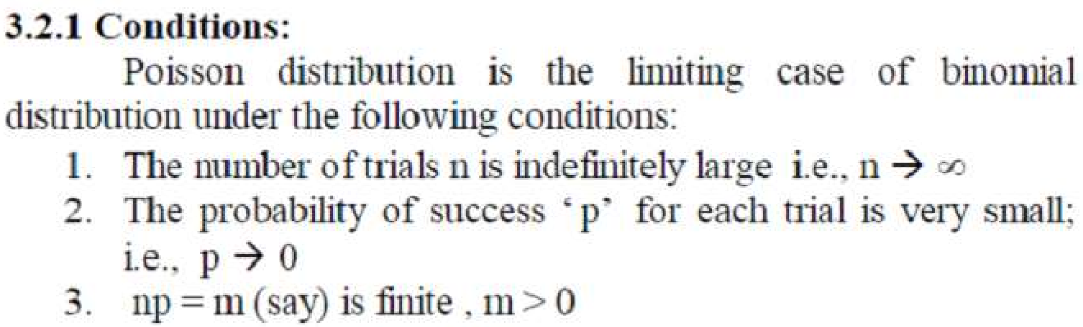
2) $P(X=0) = \dfrac{e^{-m} m^0}{0!}$, $0! = 1$ and $1! = 1$

3) $P(X=1) = \dfrac{e^{-m} m^1}{1!}$



Some examples of Poisson variates are :

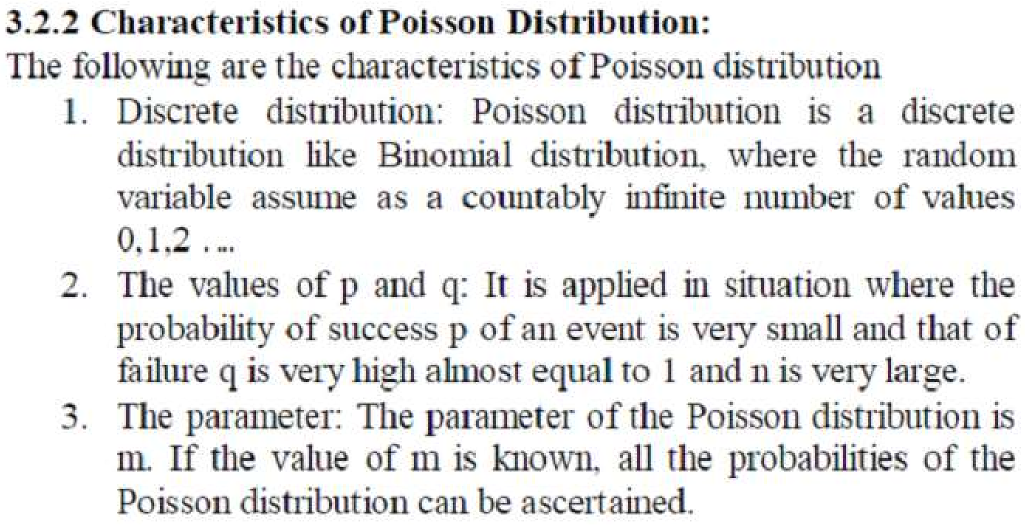
1. The number of blinds born in a town in a particular year.
2. Number of mistakes committed in a typed page.
3. The number of students scoring very high marks in all subjects
4. The number of plane accidents in a particular week.
5. The number of defective screws in a box of 100, manufactured by a reputed company.
6. Number of suicides reported in a particular day.



3.2.1 Conditions:

Poisson distribution is the limiting case of binomial distribution under the following conditions:

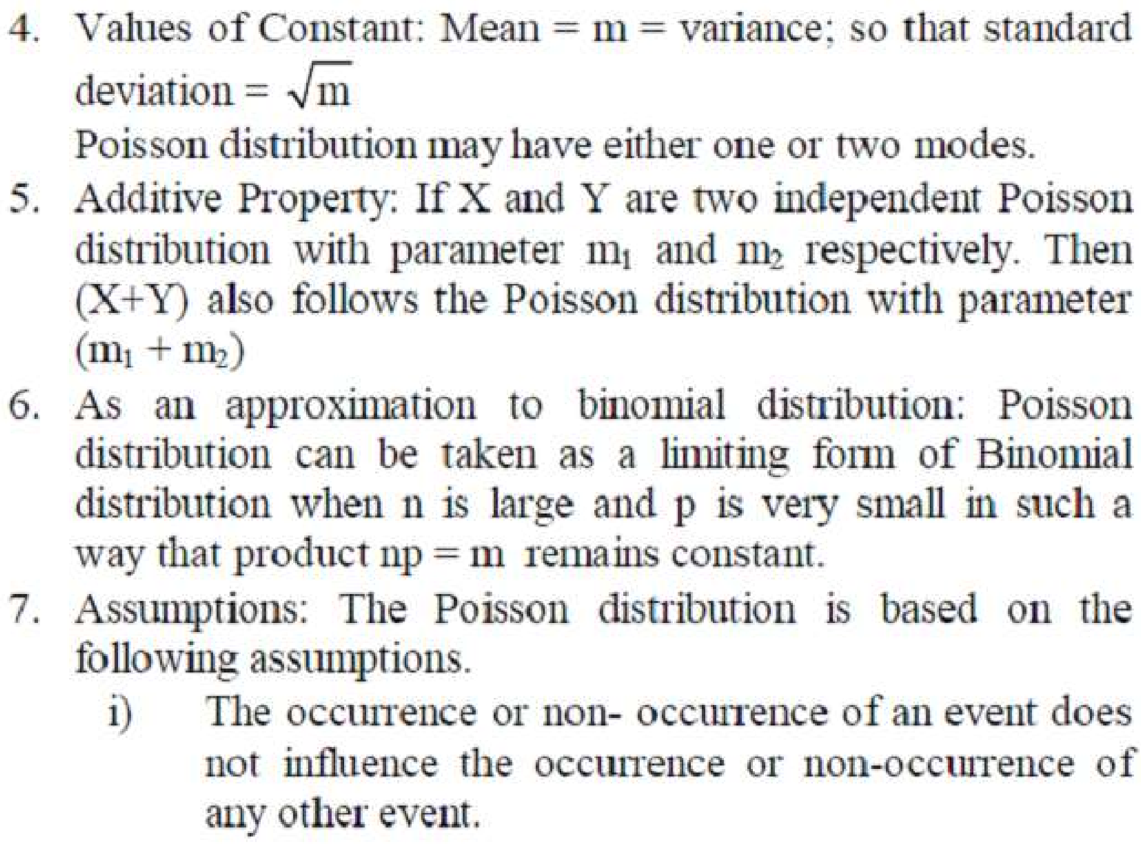
1. The number of trials n is indefinitely large i.e., $n \rightarrow \infty$
2. The probability of success 'p' for each trial is very small; i.e., $p \rightarrow 0$
3. $np = m$ (say) is finite , $m > 0$



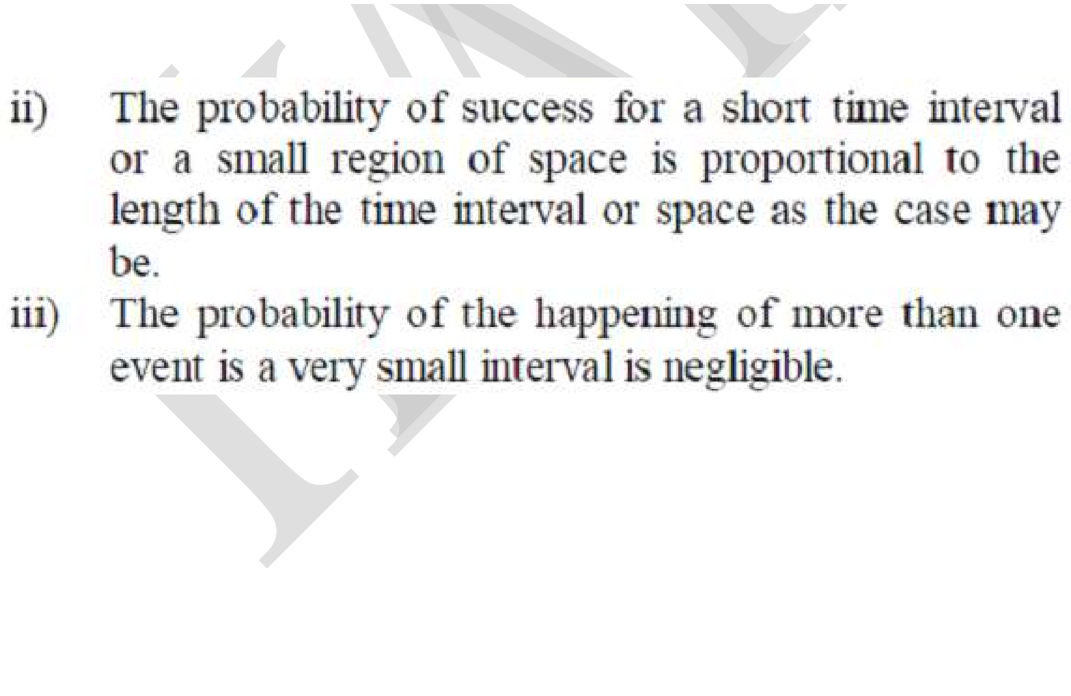
3.2.2 Characteristics of Poisson Distribution:

The following are the characteristics of Poisson distribution

1. Discrete distribution: Poisson distribution is a discrete distribution like Binomial distribution, where the random variable assume as a countably infinite number of values 0,1,2 ....
2. The values of p and q: It is applied in situation where the probability of success p of an event is very small and that of failure q is very high almost equal to 1 and n is very large.
3. The parameter: The parameter of the Poisson distribution is m. If the value of m is known, all the probabilities of the Poisson distribution can be ascertained.



4. Values of Constant: Mean = m = variance; so that standard deviation = $\sqrt{m}$
   Poisson distribution may have either one or two modes.
5. Additive Property: If X and Y are two independent Poisson distribution with parameter m_1 and m_2 respectively. Then (X+Y) also follows the Poisson distribution with parameter $(m_1 + m_2)$
6. As an approximation to binomial distribution: Poisson distribution can be taken as a limiting form of Binomial distribution when n is large and p is very small in such a way that product np = m remains constant.
7. Assumptions: The Poisson distribution is based on the following assumptions.
   i) The occurrence or non- occurrence of an event does not influence the occurrence or non-occurrence of any other event.



ii) The probability of success for a short time interval or a small region of space is proportional to the length of the time interval or space as the case may be.

iii) The probability of the happening of more than one event is a very small interval is negligible.

Example 8:

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? [given that $e^{-2} = 0.13534$]

Mean, $\bar{x} = np$, $n = 2000$ and $p = \dfrac{1}{1000}$

$$= 2000 \times \frac{1}{1000}$$

$$m = 2$$

The Poisson distribution is

$$P(X=x) = \frac{e^{-m} m^{x}}{x!}$$

$$\therefore P(X=5) = \frac{e^{-2} 2^{5}}{5!}$$

$$= \frac{(0.13534) \times 32}{120}$$

$$= 0.036$$

(Note: The values of e^{-m} are given in Appendix)

Example 9:

In a Poisson distribution 3P(X=2) = P(X=4) Find the parameter ‘ m’ .

Solution:

Poisson distribution is given by $P(X=x) = \dfrac{e^{-m} m^{x}}{x!}$

Given that 3P(x=2) = P(x= 4)

$$3.\ \frac{e^{-m} m^2}{2!} = \frac{e^{-m} m^4}{4!}$$

$$m^2 = \frac{3\times 4!}{2!}$$

$$\therefore \quad m = \pm 6$$

Since mean is always positive $\therefore$ m = 6

Example 10:

If 2% of electric bulbs manufactured by a certain company are defective. Find the probability that in a sample of 200 bulbs i) less than 2 bulbs ii) more than 3 bulbs are defective.[$e^{-4} = 0.0183$]

Solution:

The probability of a defective bulb = $p = \frac{2}{100} = 0.02$

Given that n = 200 since p is small and n is large

We use the Poisson distribution

mean, $m = np = 200 \times 0.02 = 4$

Now, Poisson Probability function, $P(X = x) = \frac{e^{-m} m^x}{x!}$

i) Probability of less than 2 bulbs are defective

$$= P(X<2)$$

$$= P(x = 0) + P(x = 1)$$

$$= \frac{e^{-4} 4^0}{0!} + \frac{e^{-4} 4^1}{1!}$$

$$= e^{-4} + e^{-4}\,(4)$$
$$= e^{-4}\,(1 + 4) = 0.0183 \times 5$$
$$= 0.0915$$

ii) Probability of getting more than 3 defective bulbs

$$P(x > 3) \quad = 1 - P(x \le 3)$$
$$= 1 - \{P(x = 0) + P(x = 1) + P(x=2) + P(x=3)\}$$
$$= 1 - e^{-4}\ \{1 + 4 + \frac{4^2}{2!} + \frac{4^3}{3!}\}$$
$$= 1 - \{0.0183 \times (1 + 4 + 8 + 10.67)\}$$
$$= 0.567$$

3.2.3 Fitting of Poisson Distribution:

The process of fitting of Poisson distribution for the probabilities of x = 0, 1,2,... success are given below :

i) First we have to calculate the mean = $\bar{x} = \frac{\sum fx}{\sum f} = m$

ii) The value of e^{-m} is obtained from the table (see Appendix)

iii) By using the formula $P(X=x) = \frac{e^{-m}.m^x}{x!}$

Substituting x = 0, $P(0) = e^{-m}$

Then $f(0) = N \times P(0)$

The other expected frequencies will be obtained by using the recurrence formula

$$f(x+1) = \frac{m}{x+1}\ f(x)\ ;\ x = 0,1,2,\ \dots$$

Example 11:

The following mistakes per page were observed in a book.

Number of mistakes (per page)	0	1	2	3	4
Number of pages	211	90	19	5	0

Fit a Poisson distribution to the above data.

Solution:

x_i	f_i	$f_i x_i$
0	211	0
1	90	90
2	19	38
3	5	15
4	0	0
	N = 325	Σfx = 143

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{N}$$

$$= \frac{143}{325} = 0\ .44 = m$$

$$\text{Then } e^{-m} \Rightarrow e^{-0.44} = 0.6440$$

Probability mass function of Poisson distribution is

$$P(x) = e^{-m}\frac{m^x}{x!}$$

Put x = 0, $P(0) = e^{-0.44}\ \frac{44^0}{0!}$

$$= e^{-0.44}$$
$$= 0.6440$$
$$\therefore f(0) = \mathrm{N\,P}(0)$$
$$= 325 \times 0.6440$$
$$= 209.43$$

The other expected frequencies will be obtained by using the recurrence formula

$\mathrm{f}(\mathrm{x}+1) = \dfrac{\mathrm{m}}{\mathrm{x}+1}\mathrm{f}(\mathrm{x})$. By putting x = 0,1,2,3 we get the expected frequencies and are calculated as follows.

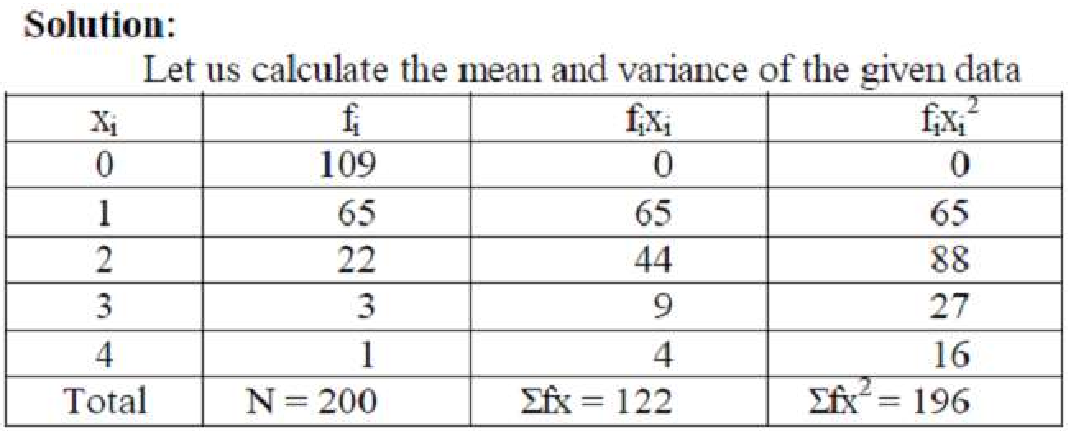
$$\mathrm{f}(1) = 0.44 \times 209.43 \quad = 92.15$$
$$\mathrm{f}(2) = \frac{0.44}{2} \times 92.15 \quad = 20.27$$
$$\mathrm{f}(3) = \frac{0.44}{3} \times 20.27 \quad = 2.97$$
$$\mathrm{f}(4) = \frac{0.44}{4} \times 2.97 \quad = 0.33$$

						Total
Observed frequencies	211	90	19	5	0	325
Expected frequencies	210	92	20	3	0	325

Example 12:

Find mean and variance to the following data which gives the frequency of the number of deaths due to horse kick in 10 corps per army per annum over twenty years.

X	0	1	2	3	4	Total
F	109	65	22	3	1	200



Solution:

Let us calculate the mean and variance of the given data

x_i	f_i	$f_i x_i$	$f_i x_i^2$
0	109	0	0
1	65	65	65
2	22	44	88
3	3	9	27
4	1	4	16
Total	N = 200	$\Sigma fx = 122$	$\Sigma fx^2 = 196$

$$\text{Mean} = \bar{x} = \frac{\sum_i f_i x}{N}$$

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

$$= 0.61$$

$$\text{Variance} = \sigma^2 = \frac{\sum_i f_i^2 x}{N} - \left(\bar{x}\right)^2$$

$$= \frac{196}{200} - (0.61)^2$$

$$= 0.61$$

Hence, mean = variance = 0.61

Example 13:

100 car radios are inspected as they come off the production line and number of defects per set is recorded below

No. of defects	0	1	2	3	4
No. of sets	79	18	2	1	0

Fit a Poisson distribution and find expected frequencies

Solution:

x	f	fx
0	79	0
1	18	18
2	2	4
3	1	3
4	0	0
	N = 100	$\Sigma fx = 25$

$$\text{Mean} = \bar{x} = \frac{\Sigma fx}{N}$$

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

$$\therefore m = 0.25$$

Then $e^{-m} = e^{-0.25} = 0.7788 = 0.779$

Poisson probability function is given by

$$P(x) = \frac{e^{-m} m^{x}}{x!}$$

$$P(0) = \frac{e^{-0.25}(0.25)^0}{0!} = (0.779)$$

$\therefore$ $f(0) = N.P(0) = 100 \times (0.779) = 77.9$

Other frequencies are calculated using the recurrence formula

$$f(x+1) = \frac{m}{x+1} f(x).$$

By putting x = 0,1,2,3, we get the expected frequencies and are calculated as follows.

$$f(1) = f(0+1) = \frac{m}{0+1} f(0)$$

$$f(1) = \frac{0.25}{1}(77.9)$$

$$= 19.46$$

$$f(2) = \frac{0.25}{2}(19.46)$$

$$= 2.43$$

$$f(3) = \frac{0.25}{3}(2.43)$$

$$= 0.203$$

$$f(4) = \frac{0.25}{4}(0.203)$$

$$= 0.013$$

Observed frequencies	79	18	2	1	0	100
Expected frequencies	78	20	2	0	0	100

Example 14:

Assuming that one in 80 births in a case of twins, calculate the probability of 2 or more sets of twins on a day when 30 births occurs. Compare the results obtained by using (i) the binomial and (ii) Poisson distribution.

Solution:

(i) Using Binomial distribution

$$\text{Probability of twins birth} \quad = \text{p} = \frac{1}{80} = 0.0125$$

$$\therefore \quad \text{q} = 1 - \text{p} = 1 - 0.0125$$

$$= 0.9875$$

$$\text{n} = 30$$

Binomial distribution is given by

$$\text{P}(x) = \text{nC}_\text{x} \ \text{p}^\text{x} \, \text{q}^{\text{n-x}}$$

$$\text{P}(\text{x} \geq 2) = 1 - \text{P}(\text{x} < 2)$$

$$= 1 - \{\text{P}(\text{x} = 0) + \text{P}(\text{x} = 1)\}$$

$$= 1 - \{30\text{C}_0(0.0125)^0 (0.9875)^{30} + 30\text{C}_1 (0.0125)^1 (0.9875)^{29}\}$$

$$= 1 - \{1.1(0.9875)^{30} + 3\ (0.125)\ (0.9875)^{29}\}$$

$$= 1 - \{\ 0.6839 + 0.2597\}$$

$$= 1 - 0.9436$$

$$\text{P}(\ \text{x} \geq 2) = 0.0564$$

(ii) By using Poisson distribution:

The probability mass function of Poisson distribution is given by

$$P(x) = \frac{e^{-m}m^x}{x!}$$

$$\text{Mean} = m = np$$
$$= 30\,(0.0125) = 0.375$$

$$P(x \geq 2) = 1 - P(x < 2)$$
$$= 1 - \{ P(x = 0) + P(x = 1)\}$$
$$= 1 - \{ \frac{e^{-0.375}(0.375)^0}{0!} + \frac{e^{-0.375}(0.375)^1}{1!} \}$$
$$= 1 - e^{-0.375}\ (1 + 0.375)$$
$$= 1 - (0.6873)\ (1.375) = 1 - 0.945 = 0.055$$

3.3 NORMAL DISTRIBUTION:

3.3.0 Introduction:

In the preceding sections we have discussed the discrete distributions, the Binomial and Poisson distribution.

In this section we deal with the most important continuous distribution, known as normal probability distribution or simply normal distribution. It is important for the reason that it plays a vital role in the theoretical and applied statistics.

The normal distribution was first discovered by DeMoivre (English Mathematician) in 1733 as limiting case of binomial distribution. Later it was applied in natural and social science by Laplace (French Mathematician) in 1777. The normal distribution is also known as Gaussian distribution in honour of Karl Friedrich Gauss(1809).

3.3.1 Definition:

A continuous random variable X is said to follow normal distribution with mean μ and standard deviation σ, if its probability density function

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\ e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ;-\infty < x < \infty\ ,\ -\infty < \mu < \infty,\ \sigma > 0.$$

Note:

The mean μ and standard deviation σ are called the parameters of Normal distribution. The normal distribution is expressed by $X \sim N(\mu, \sigma^2)$

3.3.2 Condition of Normal Distribution:

i) Normal distribution is a limiting form of the binomial distribution under the following conditions.

a) n, the number of trials is indefinitely large ie., $n \rightarrow \infty$ and

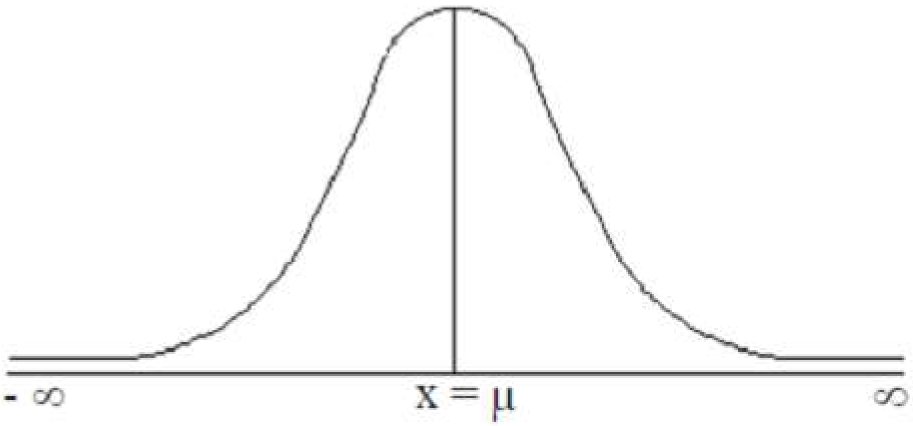
b) Neither p nor q is very small.

ii) Normal distribution can also be obtained as a limiting form of Poisson distribution with parameter $m \rightarrow \infty$

iii) Constants of normal distribution are mean = μ, variation $=\sigma^2$, Standard deviation = σ.

3.3.3 Normal probability curve:

The curve representing the normal distribution is called the normal probability curve. The curve is symmetrical about the mean (μ), bell-shaped and the two tails on the right and left sides of the mean extends to the infinity. The shape of the curve is shown in the following figure.



3.3.4 Properties of normal distribution:

1. The normal curve is bell shaped and is symmetric at x = μ.
2. Mean, median, and mode of the distribution are coincide i.e., Mean = Median = Mode = μ
3. It has only one mode at x = μ (i.e., unimodal)
4. Since the curve is symmetrical, Skewness = $\beta_1 = 0$ and Kurtosis = $\beta_2 = 3$.
5. The points of inflection are at $x = \mu \pm \sigma$
6. The maximum ordinate occurs at x = μ and its value is $= \dfrac{1}{\sigma\sqrt{2\pi}}$
7. The x axis is an asymptote to the curve (i.e. the curve continues to approach but never touches the x axis)
8. The first and third quartiles are equidistant from median.
9. The mean deviation about mean is 0.8 σ
10. Quartile deviation = 0.6745 σ

11. If X and Y are independent normal variates with mean μ_1 and μ_2, and variance σ_1^2 and σ_2^2 respectively then their sum $(X + Y)$ is also a normal variate with mean $(\mu_1 + \mu_2)$ and variance $(\sigma_1^2 + \sigma_2^2)$
12. Area Property $P(\mu - \sigma < x < \mu + \sigma) = 0.6826$
$P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.9544$
$P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.9973$

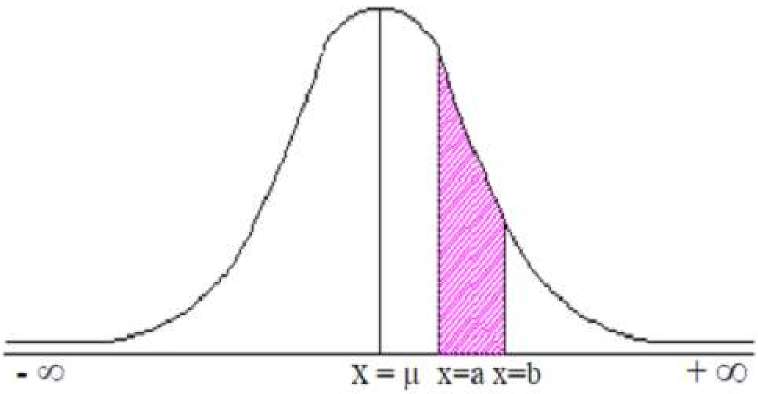
3.3.5 Standard Normal distribution:

Let X be random variable which follows normal distribution with mean μ and variance σ^2 .The standard normal variate is defined as $Z = \dfrac{X-\mu}{\sigma}$ which follows standard normal distribution with mean 0 and standard deviation 1 i.e., $Z \sim N(0,1)$. The standard normal distribution is given by $\phi(z) = \dfrac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}z^2}$; $-\infty < z < \infty$

The advantage of the above function is that it doesn' t contain any parameter. This enable us to compute the area under the normal probability curve.

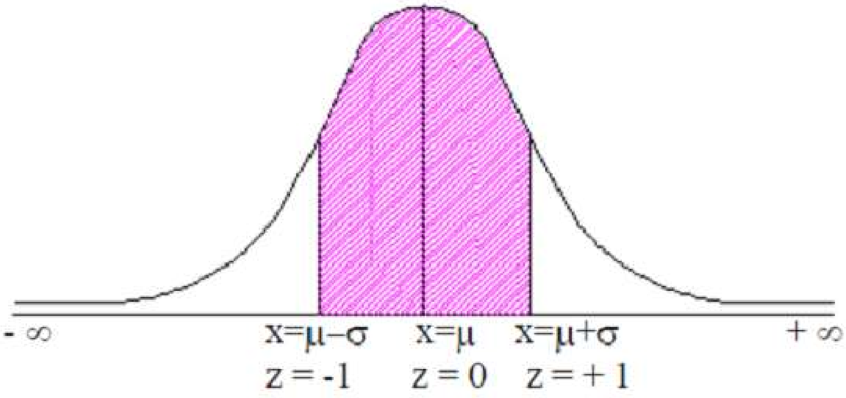
3.3.6 Area properties of Normal curve:

The total area under the normal probability curve is 1. The curve is also called standard probability curve. The area under the curve between the ordinates at x = a and x = b where a < b, represents the probabilities that x lies between x = a and x = b i.e., $P(a \le x \le b)$



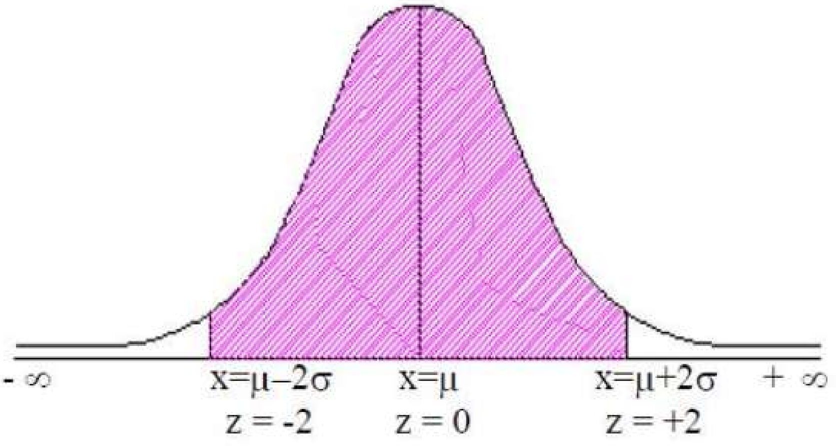
To find any probability value of x, we first standardize it by using $Z = \dfrac{X - \mu}{\sigma}$, and use the area probability normal table. (given in the Appendix).

For Example: The probability that the normal random variable x to lie in the interval (μ−σ , μ+σ) is given by

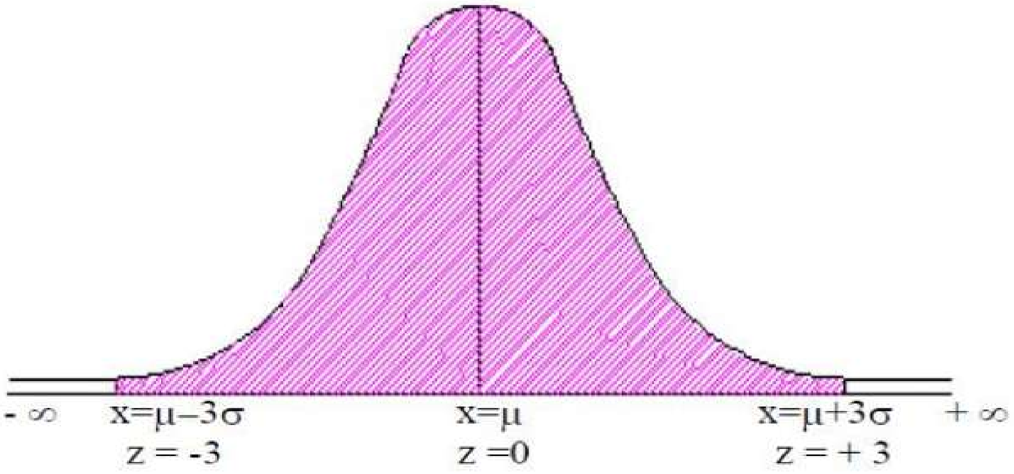


$$\begin{aligned} P(\mu - \sigma < x < \mu+\sigma) &= P(-1 \leq z \leq 1) \\ &= 2P(0 < z < 1) \\ &= 2\,(0.3413) \quad \text{(from the area table)} \\ &= 0.6826 \end{aligned}$$

$$\begin{aligned} P(\mu - 2\sigma < x < \mu+2\sigma) &= P(-2 < z < 2) \\ &= 2P(0 < z < 2) \\ &= 2\,(0.4772) = 0.9544 \end{aligned}$$



$$\begin{aligned} P(\mu - 3\sigma < x < \mu + 3\sigma) &= P(-3 < z < 3) \\ &= 2P(0 < z < 3) \\ &= 2\,(0.49865) = 0.9973 \end{aligned}$$



The probability that a normal variate x lies outside the range μ ± 3σ is given by

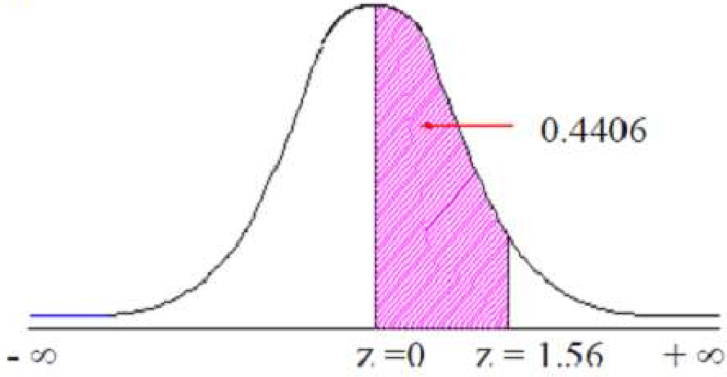
$$P(|x - \mu| > 3\sigma) = P(|z| > 3)$$
$$= 1 - P(-3 \le z \le 3)$$
$$= 1 - 0.9773 = 0.0027$$

Thus we expect that the values in a normal probability curve will lie between the range μ ± 3σ, though theoretically it range from – ∞ to ∞.

Example 15:

Find the probability that the standard normal variate lies between 0 and 1.56

Solution:

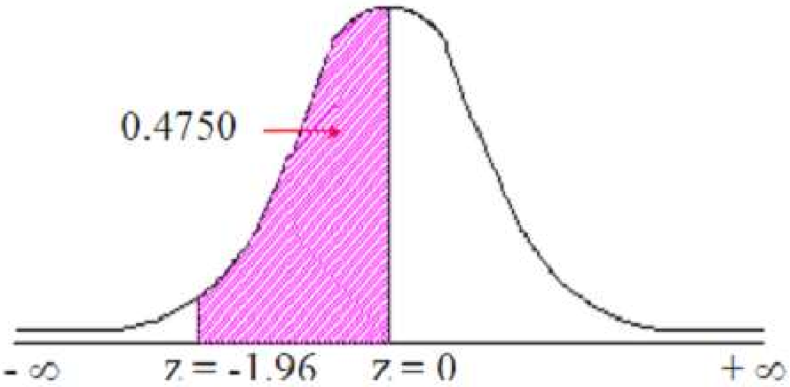


$$P(0<z<1.56) = \text{Area between } z = 0 \text{ and } z = 1.56$$
$$= 0.4406 \quad \text{(from table)}$$

Example 16:

Find the area of the standard normal variate from –1.96 to 0.

Solution:



Area between z = 0 & z =1.96 is same as the area z = –1.96 to z = 0

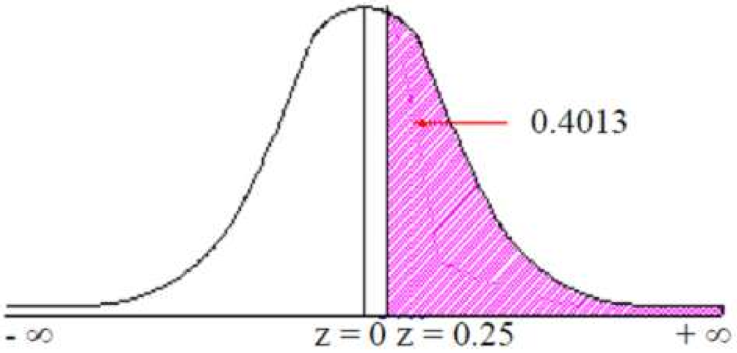
$P(-1.96 < z < 0) = P(0 < z < 1.96)$ (by symmetry)

$= 0.4750$ (from the table)

Example 17:

Find the area to the right of z = 0.25

Solution:

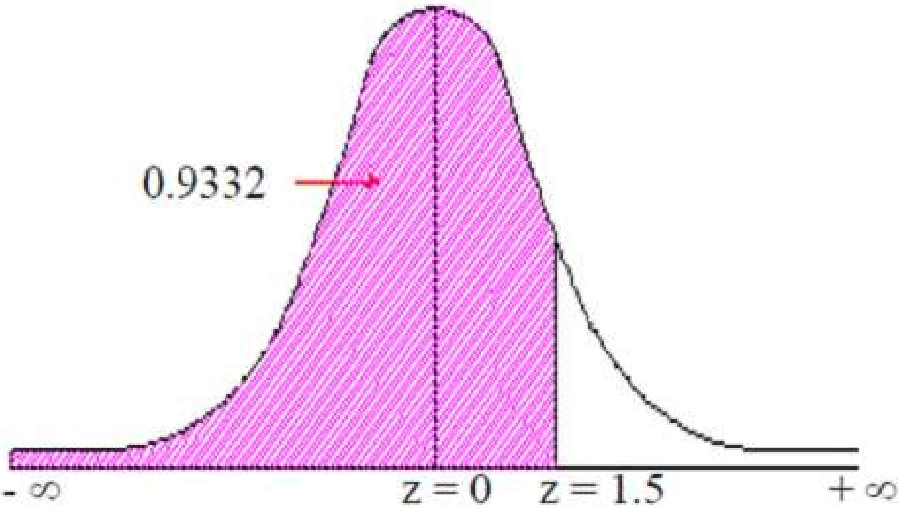


$$P(z > 0.25) = P(0<z<\infty) - P(0<z<0.25)$$
$$= 0.5000 - 0.0987 \quad \text{(from the table)} \quad = 0.4013$$

Example 18:

Find the area to the left of z = 1.5

Solution:



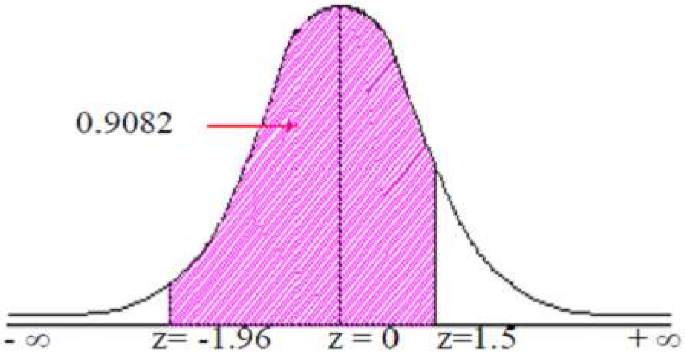
$$P(z < 1.5) = P(-\infty < z < 0) + P(0 < z < 1.5)$$
$$= 0.5 + 0.4332 \quad \text{(from the table)}$$
$$= 0.9332$$

KAHE

Example 19:

Find the area of the standard normal variate between –1.96 and 1.5

Solution:

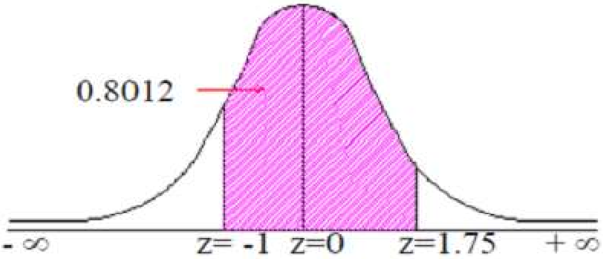


$$\begin{aligned} P(-1.96 < z < 1.5) &= P(-1.96 < z < 0) + P(0 < z < 1.5) \\ &= P(0 < z < 1.96) + P(0 < z < 1.5) \\ &= 0.4750 + 0.4332 \quad \text{(from the table)} \\ &= 0.9082 \end{aligned}$$

Example 20:

Given a normal distribution with μ = 50 and σ = 8, find the probability that x assumes a value between 42 and 64

Solution:



Given that $\mu = 50$ and $\sigma = 8$

The standard normal variate $z = \dfrac{x-\mu}{\sigma}$

$$\text{If } X = 42\text{, } Z_1 = \frac{42-50}{8} = \frac{-8}{8} = -1$$

$$\text{If } X = 64\text{, } Z_2 = \frac{64-50}{8} = \frac{14}{8} = 1.75$$

$$\begin{aligned}\therefore P(42 < x < 64) &= P(-1 < z < 1.75)\\ &= P(-1 < z < 0) + P(0 < z < 1.95)\\ &= P(0<z<1) + P(0 < z < 1.75) \text{ (by symmetry)}\\ &= 0.3413 + 0.4599 \text{ (from the table)}\\ &= 0.8012\end{aligned}$$

Example 21:

Students of a class were given an aptitude test. Their marks were found to be normally distributed with mean 60 and standard deviation 5. What percentage of students scored.

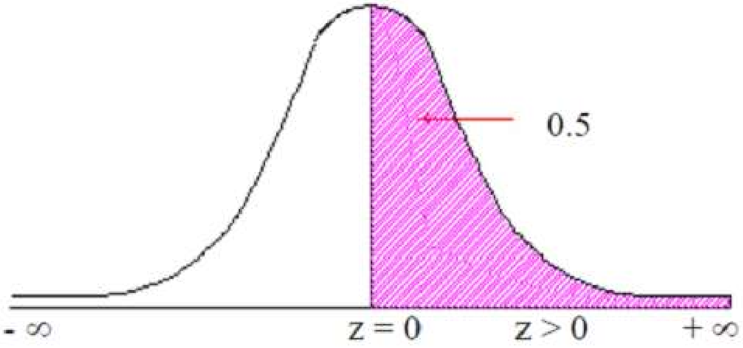
i) More than 60 marks (ii) Less than 56 marks

(iii) Between 45 and 65 marks

Solution:

Given that mean = μ = 60 and standard deviation = σ = 5

i) The standard normal varaiate $Z = \dfrac{x-\mu}{\sigma}$

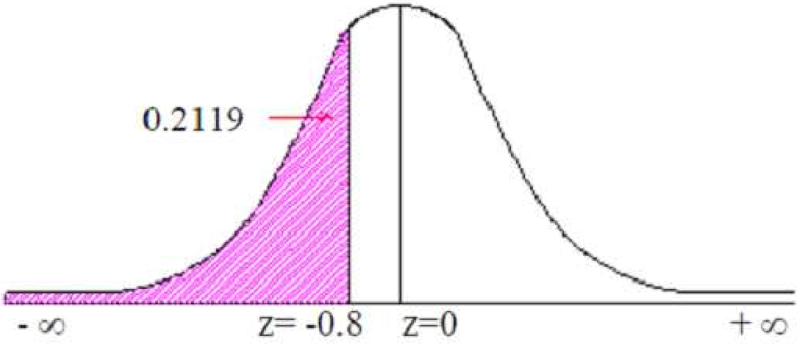


$$\text{If } X = 60,\ Z = \frac{x-\mu}{\sigma} = \frac{60-60}{5} = 0$$

$$\therefore P(x > 60) = P(z > 0)$$
$$= P(0 < z < \infty) = 0.5000$$

Hence the percentage of students scored more than 60 marks is 0.5000(100) = 50 %

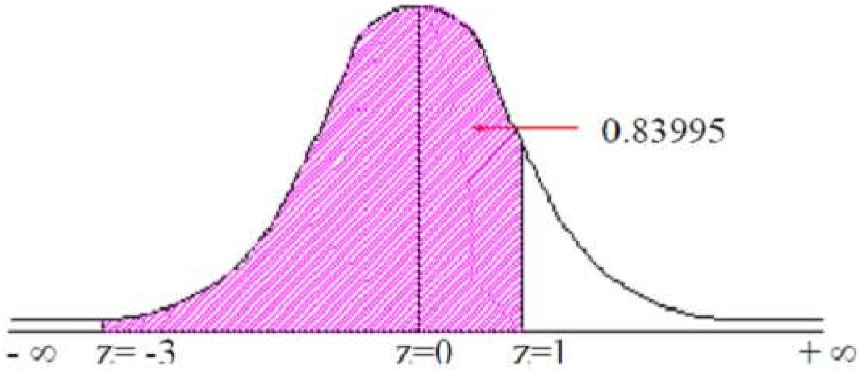
$$\text{ii) If } X = 56,\ Z = \frac{56-60}{5} = \frac{-4}{5} = -0.8$$



$$\begin{aligned} P(x < 56) &= P(z < -0.8) \\ &= P(-\infty < z < 0) - P(-0.8 < z < 0) \quad \text{(by symmetry)} \\ &= P(0 < 2 < \infty) - P(0 < z < 0.8) \\ &= 0.5 - 0.2881 \quad \text{(from the table)} \\ &= 0.2119 \end{aligned}$$

Hence the percentage of students score less than 56 marks is 0.2119(100) = 21.19 %

iii) If X = 45, then $z = \dfrac{45-60}{5} = \dfrac{-15}{5} = -3$



$X = 65$ then $z = \dfrac{65-60}{5} = \dfrac{5}{5} = 1$

$$\begin{aligned} P(45 < x < 65) &= P(-3 < z < 1) \\ &= P(-3 < z < 0) + P(0 < z < 1) \\ &= P(0 < z < 3) + P(0 < z < 1) && \text{(by symmetry)} \\ &= 0.4986 + 0.3413 && \text{(from the table)} \\ &= 0.8399 \end{aligned}$$

Hence the percentage of students scored between 45 and 65 marks is 0.8399(100) = 83.99 %

Example 22:

X is normal distribution with mean 2 and standard deviation 3. Find the value of the variable x such that the probability of the interval from mean to that value is 0.4115

Solution:

Given $\mu = 2, \sigma = 3$

Suppose z_1 is required standard value,

Thus $P(0 < z < z_1) = 0.4115$

From the table the value corresponding to the area 0.4115 is 1.35 that is $z_1 = 1.35$

$$\text{Here } z_1 = \frac{x - \mu}{\sigma}$$

$$1.35 = \frac{x - 2}{3}$$

$$x = 3(1.35) + 2$$

$$= 4.05 + 2 = 6.05$$

Example 23:

In a normal distribution 31 % of the items are under 45 and 8 % are over 64. Find the mean and variance of the distribution.

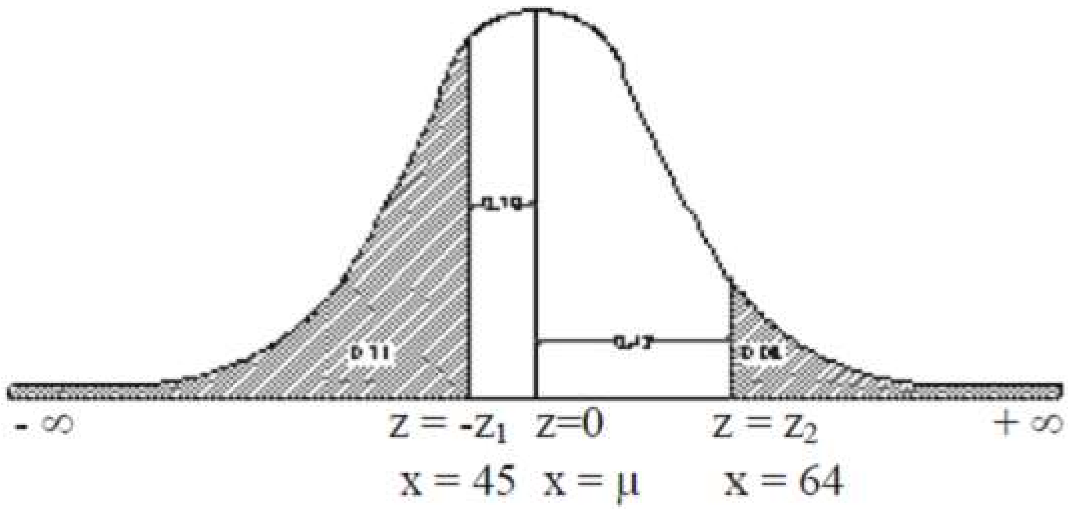
Solution:

Let x denotes the items are given and it follows the normal distribution with mean μ and standard deviation σ

The points $x = 45$ and $x = 64$ are located as shown in the figure.

i) Since 31 % of items are under $x = 45$, position of x into the left of the ordinate $x = \mu$

ii) Since 8 % of items are above $x = 64$, position of this x is to the right of ordinate $x = \mu$



When x = 45, $z = \frac{x-\mu}{\sigma} = \frac{45-\mu}{\sigma} = -z_1$ (say)

Since x is left of $x = \mu$, z_1 is taken as negative

When x = 64, $z = \frac{64-\mu}{\sigma} = z_2$ (say)

From the diagram $P(x < 45) = 0.31$

$P(z < - z_1) = 0.31$

$P(- z_1 < z < 0) = P(- \infty < z < 0) - p(- \infty < z < z_1)$

s $= 0.5 - 0.31 = 0.19$

$P(0 < z < z_1) = 0.19$ (by symmetry)

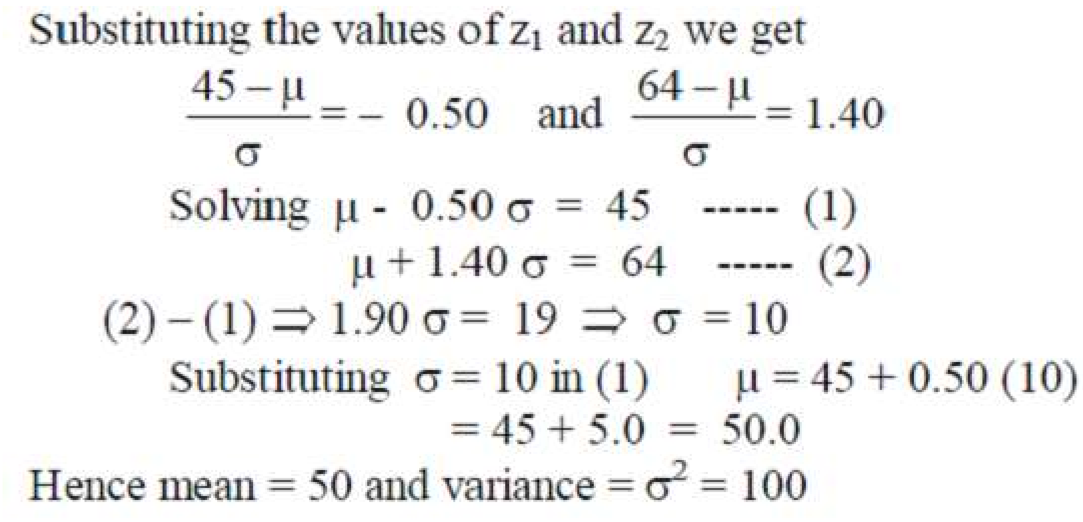
$z_1 = 0.50$ (from the table)

Also from the diagram $p(x > 64) = 0.08$

$P(0 < z < z_2) = P(0 < z < \infty) - P(z_2 < z < \infty)$

$= 0.5 - 0.08 = 0.42$

$z_2 = 1.40$ (from the table)



Substituting the values of z_1 and z_2 we get

$$\frac{45-\mu}{\sigma} = -\ 0.50 \quad \text{and} \quad \frac{64-\mu}{\sigma} = 1.40$$

$$\text{Solving } \mu - \ 0.50\,\sigma = 45 \quad \text{-----}\ (1)$$

$$\mu + 1.40\,\sigma = 64 \quad \text{-----}\ (2)$$

$$(2)-(1) \Rightarrow 1.90\,\sigma = 19 \Rightarrow \sigma = 10$$

$$\text{Substituting } \sigma = 10 \text{ in } (1) \qquad \mu = 45 + 0.50\ (10)$$

$$= 45 + 5.0 = 50.0$$

Hence mean = 50 and variance = $\sigma^2 = 100$

KAHE

POSSIBLE QUESTIONS

PART – B(TWO MARKS)

1.What is the probability of picking a card that was red or black?
2.Define Poisson Distribution.
3.Define Normal Distribution.
4.Define Bernoulli Distribution.
5.Define Bayes' probability.

PART – C(FIVE MARKS)

1.A bag contains 10 white and 6 black balls. 4 balls are successively drawn out ant not replaced. What is the probability that they are alternately of different colours?

2.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, 4 or more will contact the disease?

3.Three horses A, B and C are in a race. A is twice as likely to win as B and B is as likely to win as C.What are the respective probability of winning?

4.In a town 10 accidents took place in a span of 50 days. Assume that the number of accidents per day follows the Poisson distribution; find the probability that there will be three or more accidents in a day.

5.Find the probability that at most 5 defective bolts will be found in a box of 200 bolts, if it is known that 2 % of such bolts are expected to be defective.($e^{-4} = 0. 0183$).

6.12 coins are tossed. What are the probabilities in a single toss for getting,

i) 9 or more heads
ii) less than 3 heads
iii) atleast 8 heads

7.A person is known to hit the target in 3 out of 4 shots, whereas another person is known to hit the target in 2 out of 3 shot. Find the probability of the target being hit at all when they both try.

8.Is there any inconsistency in the statement, the mean of binomial distribution is 20 and it standard deviation 4? If no inconsistency is found what shall be the values of p, q and n.

KAHE

PART – D(TEN MARKS)

1.Find the probability that he value of an item drawn at random from a normal distribution with mean 20 and standard deviation 10 will be between:
(a) 10 and 15 (b) -5 and 10 and (c) 15 and 25
The relevant extract of the area table:

0.5	1.0	1.5	2.0	2.5
0.1915	0.3413	0.4332	0.4772	0.4938

KAHE

UNIT – IV

Testing of Hypothesis in Case of Large and Small Samples*:* Introduction – Large Samples – Assumptions , Testing Hypothesis - Null and alternate hypothesis - Selecting a Significance Level - Preference of type I error - Preference of type II error- Determine appropriate distribution, Two – Tailed Tests and One – Tailed Tests - Two – tailed tests Classification of Test Statistics - Statistics used for testing of hypothesis - Test procedure - How to identify the right statistics for the test , Introduction – small samples, 't' Distribution , Uses of 't' test, Chi-Square - Applications of Chi-Square test - Tests for independence of attributes - Test of goodness of fit - Test for specified variance, F – Distribution and Analysis of Variance (ANOVA): Introduction, Analysis of Variance (ANOVA), Assumptions for F-test - Objectives of ANOVA - ANOVA table - Assumptions for study of ANOVA, Classification of ANOVA - ANOVA table in one-way ANOVA - Two way classifications.

Simple Correlation and Regression*:* Introduction , Correlation - Causation and Correlation - Types of Correlation - Measures of Correlation - Scatter diagram - Karl Pearson's correlation coefficient - Spearman's Rank Correlation Coefficient. Regression - Regression analysis - Regression lines - Regression coefficient , Standard Error of Estimate , Multiple Regression Analysis , Reliability of Estimates , Application of Multiple Regressions

KAHE

Hypothesis:

A statistical hypothesis is an assumption that we make about a population parameter, which may or may not be true concerning one or more variables.

According to Prof. Morris Hamburg "A hypothesis in statistics is simply a quantitative statement about a population".

Hypothesis testing:

Hypothesis testing is to test some hypothesis about parent population from which the sample is drawn.

Example:

A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased.

To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

Null hypothesis:

The hypothesis under verification is known as *null hypothesis* and is denoted by H_0 and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that *"extra coaching has not benefited the students"*. Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that *"the drug is not effective in curing malaria*".

KAHE

Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis H_0 is called alternative hypothesis and is denoted by H_1 or H_a.

For example, if a psychologist who wishes to test whether or not a certain class of people have a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$H_0 : \mu = 100$

Then the alternative hypothesis could be any one of the statements.

$$\begin{aligned} H_1 &: \mu \neq 100 \\ (or)\, H_1 &: \mu > 100 \\ (or)\, H_1 &: \mu < 100 \end{aligned}$$

Errors in testing of hypothesis:

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)
3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e.1) Type-I error

2) Type-II error

1) Type-I error: The type-I error is said to be committed if the null hypothesis (H_0) is true but our test rejects it.

2) Type-II error: The type-II error is said to be committed if the null hypothesis (H_0) is false but our test accepts it.

Level of significance:

The maximum probability of committing type-I error is called level of significance and is denoted by α .

α = P (Committing Type-I error)

= P (H_0 is rejected when it is true)

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc…….

Power of the test:

The probability of rejecting a false hypothesis is called power of the test and is denoted by $1-\beta$ **.**

Power of the test =P (H_0 is rejected when it is false)

= 1- P (H_0 is accepted when it is false)

= 1- P (Committing Type-II error)

= 1- β

- A test for which both α and β are small and kept at minimum level is considered desirable.
- The only way to reduce both α and β simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

Critical region:

A statistic is used to test the hypothesis H_0. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It indicates that if the value of test statistic lies in this region, H_0 will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance α . The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

KAHE

One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called a two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (right tailed alternative) or $H_1 : \theta < \theta_0$ (left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ ------- right tailed test

$H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ ------- left tailed test

Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^{N}c_n$ possible samples. If we calculate some particular statistic from each of the ${}^{N}c_n$ samples, the distribution of sample statistic is called sampling

distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

i.e. S.E (t)=$\sqrt{Var(t)}$

Utility of standard error:

1. It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z| = \left|\frac{t - E(t)}{S.E(t)}\right| > 1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.
2. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.

3. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
4. It is used to determine the size of the sample.

Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

$$\text{i.e. test statistic } Z = \frac{t - E(t)}{S.E(t)}$$

Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0 : \theta = \theta_0$.
2. Set up a alternative hypothesis i.e. $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
3. Choose the level of significance i.e. α .
4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at α % l.o.s i.e. Z_α .
7. Compare the test statistic value with the tabulated value at α % l.o.s. and make a decision whether to accept or to reject the null hypothesis.

Large sample tests:

The sample size which is greater than or equal to 30 is called as large sample and the test depending on large sample is called large sample test.

The assumption made while dealing with the problems relating to large samples are

Assumption-1: The random sampling distribution of the statistic is approximately normal.

Assumption-2: Values given by the sample are sufficiently closed to the population value and can be used on its place for calculating the standard error of the statistic.

Large sample test for single mean (or) test for significance of single mean:

For this test

The null hypothesis is $H_0 : \mu = \mu_0$

against the two sided alternative $H_1 : \mu \neq \mu_0$

where μ is population mean

μ_0 is the value of μ

Let $x_1, x_2, x_3, ..., x_n$ be a random sample from a normal population with mean μ and variance σ^2

i.e. if $X \sim N(\mu, \sigma^2)$ then $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, Where $\bar{x}$ be the sample mean

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H$_0$

If $|Z| < Z_\alpha$, accept the null hypothesis H$_0$

Note: if the population standard deviation is unknown then we can use its estimate s, which will be calculated from the sample. $s = \sqrt{\frac{1}{n-1}\sum(x - \bar{x})^2}$.

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for the first and second populations respectively

Then $\bar{x}_1 \sim N\left(\mu_1, {\sigma_1{}^2}/{n_1}\right)$ and $\bar{x}_2 \sim N\left(\mu_2, {\sigma_2{}^2}/{n_2}\right)$

Therefore $\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}\right)$

For this test

The null hypothesis is $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

against the two sided alternative $H_1 : \mu_1 \neq \mu_2$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} \sim N(0,1) \text{[since } \mu_1 - \mu_2 \text{=0 from H}_0\text{]}$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

KAHE

If $|Z| > Z_\alpha$, reject the null hypothesis H0

If $|Z| < Z_\alpha$, accept the null hypothesis H0

Note: If ${\sigma_1}^2$ and ${\sigma_2}^2$ are unknown then we can consider ${S_1}^2$ and ${S_2}^2$ as the estimate value of ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively..

Large sample test for single standard deviation (or) test for significance of standard deviation:

Let $x_1, x_2, x_3, ..., x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2 ,

for large sample, sample standard deviation s follows a normal distribution with mean σ and variance $\sigma^2/2n$ i.e. $s \sim N\left(\sigma, \sigma^2/2n\right)$

For this test

The null hypothesis is $H_0 : \sigma = \sigma_0$
against the two sided alternative $H_1 : \sigma \neq \sigma_0$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{s - E(s)}{S.E(s)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{s - \sigma}{\sigma/\sqrt{2n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for difference between two standard deviations:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively

Let s_1 and s_2 be the sample standard deviations for the first and second populations respectively

Then $s_1 \sim N\left(\sigma_1, {\sigma_1}^2 / 2n_1\right)$ and $\bar{x}_2 \sim N\left(\sigma_2, {\sigma_2}^2 / 2n_2\right)$

Therefore s_1 - $s_2 \sim N\left(\sigma_1 - \sigma_2, \frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}\right)$

For this test

The null hypothesis is $H_0 : \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$
against the two sided alternative $H_1 : \sigma_1 \neq \sigma_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}} \sim N(0,1) \text{[since } \sigma_1 - \sigma_2 \text{=0 from H}_0\text{]}$$

Now calculate $\left|Z\right|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_{α}

If $\left|Z\right| > Z_{\alpha}$, reject the null hypothesis H_0

If $\left|Z\right| < Z_{\alpha}$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

Let x is number of success in n independent trails with constant probability p, then x follows a binomial distribution with mean np and variance npq.

In a sample of size n let x be the number of persons processing a given attribute then the sample proportion is given by $\hat{p} = \frac{x}{n}$

Then $E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}np = p$

And $V(\hat{p}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}npq = \frac{pq}{n}$

$$S.E(\hat{p}) = \sqrt{\frac{pq}{n}}$$

For this test

The null hypothesis is $H_0 : p = p_0$

against the two sided alternative $H_1 : p \neq p_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p} - E(\hat{p})}{S.E(\hat{p})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H$_0$

If $|Z| < Z_\alpha$, accept the null hypothesis H$_0$

Large sample test for single proportion (or) test for significance of proportion:

let x_1 and x_2 be the number of persons processing a given attribute in a random sample of size n_1 and n_2 then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}$ and $S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

For this test

The null hypothesis is $H_0 : p_1 = p_2$
against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When p is not known p can be calculated by $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H$_0$

If $|Z| < Z_\alpha$, accept the null hypothesis H$_0$

- **As σ is unknown,**

$$\overline{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = \left[\overline{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

Step 2: If μ_0 falls into the above confidence intervals, then

do *not* reject H_0. Otherwise, reject H_0.

Example 1:

The average starting salary of a college graduate is $19000 according to government's report. The average salary of a random sample of 100 graduates is $18800. The standard error is 800. Is the government's report reliable as the level of significance is 0.05. Find the p-value and test the hypothesis in

(a) with the level of significance $\alpha = 0.01$. The other report by some institute indicates that the average salary is $18900. Construct a 95% confidence interval and test if this report is reliable.

[solutions:]

(a)

$$H_0 : \mu = \mu_0 = 19000 \text{ vs. } H_a : \mu \neq \mu_0 = 19000,$$
$$n = 100, \bar{x} = 18800, s = 800, \alpha = 0.05$$

Then,

$$|z| = \left|\frac{\bar{x} - \mu_0}{s/\sqrt{n}}\right| = \left|\frac{18800 - 19000}{800/\sqrt{100}}\right| = |-2.5| = 2.5 > z_{\alpha/2} = z_{0.025} = 1.96.$$

Therefore, reject H_0.

(b)

$$\text{p-value} = P(|Z| > |z|) = P(|Z| > 2.5) = 2 \cdot P(Z > 2.5) = 0.0124 > 0.01$$

Therefore, *not* reject H_0.

KAHE

(c)

$$H_0 : \mu = \mu_0 = 18900 \text{ vs } H_a : \mu \neq \mu_0 = 18900,$$

A 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 18800 \pm z_{0.025} \frac{800}{\sqrt{100}} = 18800 \pm 1.96 \cdot 80 = [18643.2, 18956.8].$$

Since $\mu_0 = 18900 \in [18643.2, 18956.8]$, Therefore, ***not*** reject H_0.

Example 2:

A sample of 49 provides a sample mean of 38 and a sample standard deviation of 7. Let $\alpha = 0.05$. Please test the hypothesis

$$H_0 : u = 40 \text{ } vs. \text{ } H_a : u \neq 40.$$

based on

(a) classical hypothesis test
(b) p-value
(c) confidence interval.
[solution:]

$$\bar{x} = 38, \ s = 7, \ u_0 = 40, \ n = 49, \ z = \frac{\bar{x} - u_0}{s/\sqrt{n}} = \frac{38 - 40}{7/\sqrt{49}} = -2.$$

(a)

$$|z| = 2 > 1.96 = z_{0.025}$$

we reject H_0.

KAHE

(b)

$$p-value = P(|Z| > |z|) = P(|Z| > 2) = 2*(1-0.9772) = 0.0456 < 0.05 = \alpha$$

we reject H_0 .

(c)

$100 \times (1-\alpha)\% = 95\%$ confidence interval is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 38 \pm z_{0.025} \frac{7}{\sqrt{49}} = 38 \pm 1.96 = [36.04, 39.96].$$

Since $40 \notin [36.04, 39.96]$, we reject H_0 .

KAHE

Hypothesis Testing for the Mean (Small Samples)

For samples of size less than 30 and when σ is unknown, if the population has a normal, or nearly normal, distribution, the *t*-distribution is used to test for the mean μ.

Using the t-Test for a Mean μ when the sample is small		
Procedure	**Equations**	**Example 4**
State the claim mathematically and verbally. Identify the null and alternative hypotheses	State H_0 and H_a	$H_0: \mu \geq 16500$ $H_a: \mu < 16500$ $n = 14, \bar{x} = 15700, s = 1250$
Specify the level of significance	Specify α	$\alpha = 0.05$
Identify the degrees of freedom and sketch the sampling distribution	$d.f = n - 1$	$d.f. = 13$
Determine any critical values. If test is left tailed, use One tail, α column with a negative sign. If test is right tailed, use One tail, α column with a positive sign. If test is two tailed, use Two tails, α column with a negative and positive sign.	Table 5 (*t*-distribution) in appendix B	The test is left-tailed. Since test is left tailed and $d.f = 13$, the critical value is $t_0 = -1.771$
Determine the rejection regions.	The rejection region is $t < t_0$	The rejection region is $t < -1.771$
Find the standardized test statistic	$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \approx \frac{\bar{x} - \mu}{s/\sqrt{n}}$	$t = \frac{15700 - 16500}{1250/\sqrt{14}} \approx -2.39$
Make a decision to reject or fail to reject the null hypothesis	If t is in the rejection region, reject H_0, Otherwise do not reject H_0	Since $-2.39 < -1.771$, reject H_0
Interpret the decision in the context of the original claim.		Reject claim that mean is at least 16500.

Chi-square Tests and then F -Distribution

Goodness of Fit

DEFINITION :A **chi-square goodness-of-fit test is** used to test whether a frequency distribution fits an expected distribution.

The test is used in a multinomial experiment to determine whether the number of results in each category fits the null hypothesis:

H_0 : The distribution fits the proposed proportions

H_1 : The distribution differs from the claimed distribution.

To calculate the test statistic for the chi-square goodness-of-fit test, you can use observed frequencies and expected frequencies.

DEFINITION The **observed frequency O** of a category is the frequency for the category observed in the sample data.

The **expected frequency E** of a category is the calculated frequency for the category. Expected frequencies are obtained assuming the specified (or hypothesized) distribution. The expected frequency for the *i*th category is

$E_i = np_i$

where *n* is the number of trials (the sample size) and p_i is the assumed probability of the *i*th category.

The Chi-square Goodness of Fit Test: The sampling distribution for the goodness-of-fit test is a chi-square distribution with $k-1$ degrees of freedom where k is the number of categories. The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequency of each category and *E* represents the expected frequency of each category. To use the chi-square goodness of fit test, *the following must be true* .

KAHE

1. The observed frequencies must be obtained using a random sample.
2. The expected frequencies must be ≥ 5 .

Performing the Chi-Square Goodness-of-Fit Test (p 496)

Procedure	Equations	Example **(p 497)**
Identify the claim. State the null and alternative hypothesis.	State H_0 and H_1	H_0: Classical 4% Country 36% Gospel 11% Oldies 2% Pop 18% Rock 29%
Specify the significance level	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	d.f. = #categories - 1	$d.f. = 6 - 1 = 5$
Find the critical value	χ^2_α: Obtain from Table 6 Appendix B	$\varphi^2_{0.01}(d.f = 5) = 15.086$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 15.086$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Survey results, n = 500 Classical O= 8 E = .04*500 = 20 Country O = 210 E = .36*500 = 180 Gospel O = 7 E = .11*500 = 55 Oldies O = 10 E = .02*500 = 10 Pop O = 75 E = .18*500 = 90 Rock O= 125 E = .29*500 = 145 Substituting $\chi^2 = 22.713$
Make the decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 22.713 > 15.086 we reject the null hypothesis Equivalently $P(X \geq 22.713) < 0.01$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		Music preferences differ from the radio station's claim.

KAHE

Using Minitab to perform the Chi-Square Goodness-of-Fit Test (Manual p 237)

The data from the example above (Example 2 p 497) will be used.

Enter Three columns: Music Type: Classical, etc, Observed: 8 etc, Distribution 0.04, etc. (Note the names of the columns ‘Music Types’, ‘Observed’ and ‘Distribution’ are entered in the gray row at the top.)

Select **Calc->Calculator**, **Store the results in** C4, and calculate the **Expression** C3*500, click **OK**, Name C4 ‘Expected’ since it now contains the expected frequencies

Music Type	Observed	Distribution	Expected
Classical	8	0.04	20
Country	210	0.36	180
Gospel	72	0.11	55
Oldies	10	0.02	10
Pop	75	0.18	90
Rock	125	0.29	145

Next calculate the chi-square statistic, $(O-E)^2/E$ as follows: Click **Calc->Calculator**. **Store the results in** C5 and calculate the **Expression** (C2-C4)**2/C4. Click on **OK** and C5 should contain the calculated values.

7.2000
5.0000
41.8909
0.0000
2.5000
2.7586

KAHE

Next add up the values in C5 and the sum is the test statistic as follows: Click on **Calc->Column Statistics**. Select **Sum** and enter C5 for the **Input Variable**. Click OK. The chi-square statistic is displayed in the session window as follows:

```
Sum of C5
Sum of C5 = 22.7132
```

Next calculate the P-value: Click on **Calc->Probability Distributions->Chi-square**. Select **Cumulative Probability** and enter 5 **Degrees of Freedom** Enter the value of the test statistic 22.7132 for the **Input Constant**. Click **OK**.

The following is displayed on the Session Window.

```
Cumulative Distribution Function
Chi-Square with 5 DF
   x  P( X <= x )
22.7132     0.999617
```

P(X ≤ 22.7132) = 0.999617 So the P-value = 1 – 0.999617 = 0.000383. This is less that $\alpha = 0.01$ so we reject the null hypothesis.

Instead of calculating the P-value, we could have found the critical value from the Chi-Square table (Table 6 Appendix B) for 5 degrees of freedom as we did above. The value is 15.086, and since our test statistic is 22.7132, we reject the null hypothesis.

KAHE

Chi-Square with M&M's

H_0: Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24%
Significance level: α = 0.05
Degrees of freedom: number of categories – 1 = 5
Critical Value: $\chi^2{}_{0.05}(d.f.=5)=11.071$
Rejection Region: $\chi^2 \geq 11.071$
Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where *O* is the actual number of M&M's of each color in the bag and *E* is the proportions specified under H_0 times the total number.
Reject H_0 if the test statistic is greater than the critical value (1.145)

Section 10.2 Independence

This section describes the chi-square test for independence which tests whether two random variables are independent of each other.

DEFINTION An ***r x c*** **contingency table** shows the observed frequencies for the two variables. The observed frequencies are arranged in r rows and c columns. The intersection of a row and a column is called a **cell.**

The following is a contingency table for two variables A and B where f_{ij} is the frequency that A equals A_i and B equals B_j.

	A_1	**A_2**	**A_3**	**A_4**	**A**
B_1	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
B_2	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
B_3	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
B	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	f

KAHE

If A and B are independent, we'd expect

$$f_{ij} = prob(A = A_i) * prob(B = B_j) * f = \left(\frac{f_{i.}}{f}\right)\left(\frac{f_{.j}}{f}\right) f = \frac{(f_{i.})(f_{.j})}{f}$$

$$\frac{(sum\ of\ row\ i) * (sum\ of\ column\ j)}{sample\ size}$$ (

Example 1 Determining the expected frequencies of CEO's ages as a function of company size under the assumption that age is independent of company size.

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	$\frac{300*47}{550} \approx 25.64$	$\frac{300*87}{550} \approx 47.45$	$\frac{300*193}{550} \approx 105.27$	$\frac{300*180}{550} \approx 98.18$	$\frac{300*43}{550} \approx 23.45$	300
Large	$\frac{250*47}{550} \approx 21.36$	$\frac{250*87}{550} \approx 39.55$	$\frac{250*193}{550} \approx 87.73$	$\frac{250*180}{550} \approx 81.82$	$\frac{250*43}{550} \approx 19.55$	250
Total	47	87	193	180	43	550

After finding the expected frequencies under the assumption that the variables are independent, you can test whether they are independent using the chi-square independence test.

DEFINITION A **chi-square independence test** is used to test the independence of two random variables. Using a chi-square test, you can determine whether the occurrence of one variable affects the probability of occurrence of the other variable.

To use the test,

1. The observed frequencies must be obtained from a random sample
2. Each expected frequency must be ≥ 5

The sampling distribution for the test is a chi-square distribution with

$(r-1)(c-1)$

degrees of freedom, where r and c are the number of rows and columns, respectively, of the contingency table. The test statistic for the chi-square independence test is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequencies and *E* represents the expected frequencies.

KAHE

To begin the test we state the null hypothesis that the variables are independent and the alternative hypothesis that they are dependent.

Performing a Chi-Square Test for Independence (p 507)		
Procedure	Equations	Example2 **(p 507)**
Identify the claim. State the null and alternative hypotheses.	State H_0 and H_1	H_0: CEO's ages are independent of company size H_1: CEO's ages are dependent on company size.
Specify the level of significance	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	$d.f. = (r-1)(c-1)$	$d.f. = (2-1)(5-1) = 4$
Find the critical value.	χ^2_α :Obtain from Table 6, Appendix B	$\chi^2_\alpha \geq 13.277$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 13.277$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$\sum \frac{(O-E)^2}{E} \approx 77.9$ Note that O is in the table of actual CEO's ages above, and E is in the table of Expected CEO's ages (if independent of size) above
Make a decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region. Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 77.9 > 13.277 we reject the null hypothesis Equivalently $P(X \geq 77.0) < \alpha$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		CEO's ages and company size are dependent.

The test statistic (77.887) is greater than the critical value obtained from Table 6, Appendix B (13.277) so the null hypothesis is rejected. (Alternatively the P-Value (0.000) is less than the level of significance, α (0.01) so the null hypothesis is rejected.)

3. An urban geographer randomly samples 20 new residents of a neighborhood to determine their ratings of local bus service. The scale used is as follows: 0–very dissatisfied, 1–dissatisfied, 2– neutral, 3–satisfied, 4–very satisfied. The 20 responses are 0,4,3, 2,2,1,1,2,1,0,01,2,1,3,4,2,0,4,1. Use the sign test to see whether the population median is 2.

   Solution:

   There are 5 observations above the hypothesized median. Because the sample size is larger than 10, we test using the sample proportion p = 5/20 = 0.25. Using the PROB-VALUE method the steps in this test are:

   1) H_0: □ = 0.5 and H_A: □ ¹ 0.5
   2) We will use the Z-distribution
   3) We will use the 5%-level, thus □ = 0.05
   4) The test statistic is $z = (0.25 -\ 0.5)/\sqrt{0.25/20} = -\ 2.24$
   5) Table A-4 shows that P(|Z| > 2.24) » 0.025.
   6) Because PROB-VALUE < □, we reject H_0. We conclude □ is different than 0.5, and thus the median is different than 2.

4. A course in statistical methods was team-taught by two instructors, Professor Jovita Fontanez and Professor Clarence Old. Professor Fontanzez used many active learning techniques, whereas Old employed traditional methods. As part of the course evaluation, students were asked to indicate their instructor preference. There was reason to think students would prefer Fontanez, and the sample obtained was consistent with that idea: of the 12 students surveyed, 8 preferred Professor Fontanez and 2 preferred Professor Old. The remaining students were unable to express a preference. Test the hypothesis that the students prefer Fontanez. (*Hint:* Use the sign test.)

   Solution:

KAHE

Although the sample is large enough for a normal approximation, we will use the binomial distribution to illustrate its application. Of the 12 observations, 8 preferred Prof. Fontanez, thus we need the probability of observing 8 or more successes in 12 trials of a Bernoulli process with the probability of success equal to 0.5. From Table A-1, we get

$$P(X\ {}^{3}\ 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$

$$P(X\ {}^{3}\ 8) = 0.1208 \quad + 0.0537 \quad + 0.0161 \quad + 0.0029 \quad + 0.002 \qquad = 0.1937$$

Adopting the 5% uncertainty level, we see that PROB-VALUE > □□□ Thus we fail to reject H_0. We cannot conclude students prefer Fontanez.

5. Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs.

Solution:

(a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

(b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference.

KAHE

We conclude the samples were drawn from different populations.

6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

High Density	Low Density	Sparsely Settled
1.84	2.04	1.07
3.06	2.28	2.31
3.62	4.01	0.91
4.91	1.86	3.28
3.49	1.42	1.31

Solution:

We will use the multi-sample Kruskal-Wallis test with an uncertainly level □ = 0.1. The null hypothesis is that all samples have been drawn from the same population. The rank sums are 55, 39 and 26 for the high density, low density, and sparsely settled samples respectively. The Kruskal-Wallis statistic is

$$H = \frac{12}{15(15+1)}\left(\frac{55^2}{5}+\frac{39^2}{5}+\frac{26^2}{5}\right) - 3(15+1) = 4.22$$

Using the □2 distribution with $3 - 1 = 2$ degrees of freedom, the associated PROB-VALUE is 0.121. We fail to reject the null hypothesis. The sample does not support the hypothesis of differing waste generation rates.

7. The distances travelled to work by a random sample of 12 people to their places of work in 1996 and again in 2006 are shown in the following table.

	Distance (km)			Distance (km)	
Person	1996	2006	Person	1996	2006
1	8.6	8.8	7	7.7	6.5
2	7.7	7.1	8	9.1	9
3	7.7	7.6	9	8	7.1
4	6.8	6.4	10	8.1	8.8
5	9.6	9.1	11	8.7	7.2
6	7.2	7.2	12	7.3	6.4

Has the length of the journey to work changed over the decade?

Solution:

The sample can be considered as twelve paired observations. By taking differences between paired values, we get measures of the change for each individual. If the median change for the population is zero, we expect a sample to have a median difference near zero. Thus we will do a sign test for the median difference with a hypothesized value of zero. In other words, the hypotheses are $\mathrm{H}_0 : \eta = 0$ and $\mathrm{H}_\mathrm{A}\colon \eta \neq 0$ □□We denote samples values whose distance decreased with a minus sign. Sample values with a positive difference get a plus sign. The sample becomes

S = {-,+,+,+,+,0,+,-,+,-,+,+}

Ignoring the tie, this is a sample of size 11 with 8 values above the hypothesized median. We are using Format (C) of Table 10-1, thus the PROB-VALUE is $2P(X \geq 8)$ where X is a binomial variable with □ = 0.5. From the equation for the binomial, the PROB-VALUE is found to be 0.113. At the □ = 10% level, we fail the reject the null hypothesis. We cannot conclude there has been a change in distance.

8. One hundred randomly sampled residents of a city subject to periodic flooding are classified according to whether they are on the floodplain of the major river bisecting the city or off the floodplain. These households are then surveyed to determine whether they currently have flood insurance of any kind. The survey results are as follows:

	On the Floodplain	Off the Floodplain
Insured	50	10
No Insurance	15	25

KAHE

Test a relevant hypothesis.

Solution:

We will do a χ^2 test for a relationship between insurance and house location. The null hypothesis is no relationship (independence). Augmenting the data with expected frequencies, we have:

	On the Floodplain	Off the Floodplain
Insured	50 (39)	10 (21)
No Insurance	15 (26)	25 (14)

The corresponding χ^2 value is 22.16. Table A-8 shows that with 1 degree of freedom, P($\chi^2 > 20$) is zero to 3 decimal places. Thus for any reasonable level of uncertainty (any $\alpha < 0.0005$), we can reject the null hypothesis.

The occurrence of sunshine over a 30-day period was calculated as the percentage of time the sun was visible (i.e., not obscured by clouds). The daily percentages were:

Day	Percentage of sunshine	Day	Percentage of sunshine	Day	Percentage of sunshine
1	75	11	21	21	77
2	95	12	96	22	100
3	89	13	90	23	90
4	80	14	10	24	98
5	7	15	100	25	60
6	84	16	90	26	90
7	90	17	6	27	100
8	18	18	0	28	90
9	90	19	22	29	58
10	100	20	44	30	0

KAHE

If we define a sunny day as one with over 50% sunshine, determine whether the pattern of occurrence of sunny days is random.

Solution:

For this we can use the number-of-runs test. Rather than calculate runs across two samples, here we will simply note if a day has 50% or more sunshine. The sample becomes

S={+,+,+,+,-,+,+,-,+,+,-,+,+,-,+,+,-,-,-,-,+,+,+,+,+,+,+,+,-}

We see that the sample consists of 12 runs. There are $n_x = 21$ sunny days, and $n_y = 9$ cloudy days. Because $n_x < 20$, we cannot use the normal approximation given in Table 10-5. Instead the probability is computed using combinatorial rules, and is approximately 0.4. This is far too large for rejection of the randomness hypothesis. We cannot conclude the pattern is non-random.

9. Test the normality of the DO data (a) using the Kolmogorov–Smirnov test with the ungrouped data of Table 2-4 and (b) using the χ^2 test with $k = 6$ classes of Table 2-6.

Solution:

(a) We will take the mean and standard deviation as known rather than estimated from the sample. Doing so results in calculated PROB-VALUES that are smaller than the true values (i.e., we are more likely to reject the null hypothesis). For the DO data the mean and standard deviation are 5.58 and 0.39 respectively. We sort the data, and then find the differences between the observed and expected cumulative distributions. The table below shows the results for a few of the 50 observations:

x_i	$S(x_i)$	$F(x_i)$	$\lvert S(x_i)-F(x_i)\rvert$
4.2	0.020	0.015	0.005
4.3	0.040	0.023	0.017
4.4	0.060	0.032	0.028
…	…	…	…
5.9	0.780	0.692	0.088
…	…	...	…

6.7	0.960	0.960	0.000
6.8	0.980	0.972	0.008
6.9	1.000	0.981	0.019

The maximum difference is 0.088. Table A-9 shows that with 50 degrees of freedom, the corresponding PROB-VALUE is about 0.6. We obviously cannot reject the hypothesis of normality.

(b) Here we will take the mean and standard deviation as unknown, to be estimated from the sample. In other words, we estimate two parameters from the sample. In building the □^{2} table, we combine the first two and the last two categories in Table 2-6 to ensure at least 2 expected frequencies per cell. This reduces the number of categories 4, as seen in the table below:

Group	Minimum	Maximum	O_j	E_j	$(O_j-E_j)^2/E_j$
1	4.000	4.990	9	3.3	10.13
2	5.000	5.490	10	17.0	2.89
3	5.500	5.990	20	21.7	0.14
4	6.000	6.990	11	7.0	2.24

The observed Chi-square value is 15.4. With $k - p - 1 = 4 - 2 - 1 = 1$ degrees of freedom, Table A-8 shows that the PROB-VALUE is less than 0.0005. We therefore reject the null hypothesis.

Note that with only 4 classes, we can obtain only a rough idea of the distribution of DO. The 4 classes given in Table 2-6 do not yield a distribution that is at all similar to the normal distribution. In practice one would need many classes (and observations) for the □^{2} test to be reliable.

Nonparametric Hypothesis Testing

EARNING OBJECTIVES

By the end of this chapter, you will be able to:

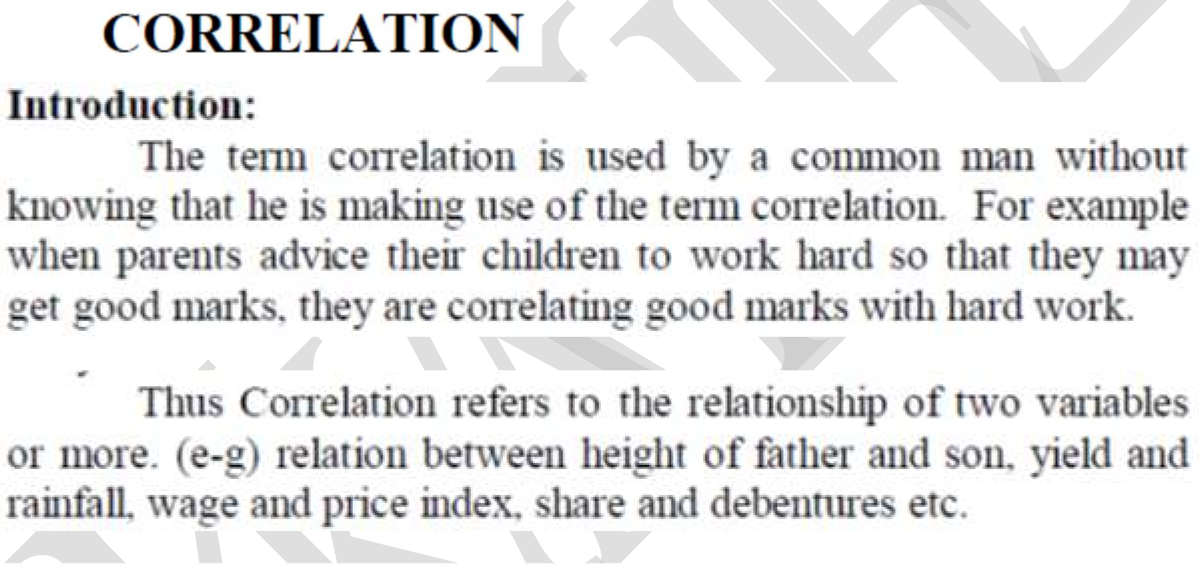
1. Identify and cite examples of situations in which nonparametric tests of hypothesis are appropriate.
2. Explain the logic of nonparametric hypothesis testing for ordinal variables as applied to the Mann-Whitney *U* and runs tests.
3. Perform Mann-Whitney *U* and runs tests using the five-step model as a guide, and correctly interpret the results.
4. Select an appropriate nonparametric test.
5. **FORMULA 1** $$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$
6.
7. where: n_1 = number of cases in sample 1
8. n_2 = number of cases in sample 2
9. $\sum R_1$ = the sum of the ranks for sample 1
10. For our sample problem above
11. $$U = (12)(12) + \frac{12(13)}{2} - 173.5$$
12. $$U = 144 + \frac{156}{2} - 173.5$$
13. $U = 48.5$
14.
15. Note that we could have computed the *U* by using data from sample 2. This alternative solution, which we will label *U*′ (*U* prime), would have resulted in a larger value for *U.* The smaller of the two values, *U* or *U*′, is always taken as the value of *U*. Once *U* has been calculated, *U*′ can be quickly determined by means of Formula 2:
16. The alternative or research hypothesis is usually a statement to the effect that the two populations are different. This form for H_1 would direct the use of a two-tailed test. It is perfectly possible to use one-tailed tests with Mann-Whitney *U* when a direction for the difference can be predicted, but, to conserve space and time, we will consider only the two-tailed case.

17. In step 3, we will take advantage of the fact that, when total sample size (the combined number of cases in the two samples) is greater than or equal to 20, the sampling distribution of *U* approximates normality. This will allow us to use the *Z*-score table (see Appendix A of the textbook) to find the critical region as marked by *Z* (critical).
18. To compute the Mann-Whitney *U* test statistic (step 4), the necessary formulas are
19. **FORMULA 3** $Z \text{ (obtained)} = \dfrac{U - \mu_{\text{U}}}{\sigma_{\text{U}}}$
20. **FORMULA 5** $\sigma_{\text{U}} = \sqrt{\dfrac{n_1\, n_2\, (n_1 + n_2 + 1)}{12}}$
21. We now have all the information we need to conduct a test of significance for *U*.

22. **Step 2. Stating the Null Hypothesis**
23. H_0: The populations from which the samples are drawn are identical on the variable of interest.
24. (H_1: The populations from which the samples are drawn are different on the variable of interest.)
25. **Step 3. Selecting the Sampling Distribution and Establishing the Critical Region**.
26. Sampling distribution = *Z* distribution
27. Alpha = 0.05
28. *Z* (critical) = ± 1.96
29. **Step 4. Calculating the Test Statistic**. With *U* equal to 48.5, □U equal to 72, and σ_{U} of 17.32,
30. *Z* (obtained) = $\dfrac{U - \mu_{\text{U}}}{\sigma_{\text{U}}}$
31. *Z* (obtained) = $\dfrac{48.5 - 72}{17.32}$
32. *Z* (obtained) = -1.36

KAHE

33. **Step 5. Making a Decision**. The test statistic, a *Z* (obtained) of -1.36, does not fall in the critical region as marked by the *Z* (critical) of ± 1.96. Therefore, we fail to reject the null of no difference. Male students are not significantly different from female students in terms of their level of satisfaction with the social life available on campus. Note that if we had used the *U′* value of 95.5 instead of *U* in computing the test statistic, the value of *Z* (obtained) would have been +1.36, and our decision to fail to reject the null would have been exactly the same.
34. **Making Assumptions**.
35. Model: Independent random sampling
36. Level of measurement is ordinal
37. **Step 2. Stating the Null Hypothesis.**
38. H_0: The two populations are identical on level of pain.
39. (H_1: The two populations are different on level of pain.)

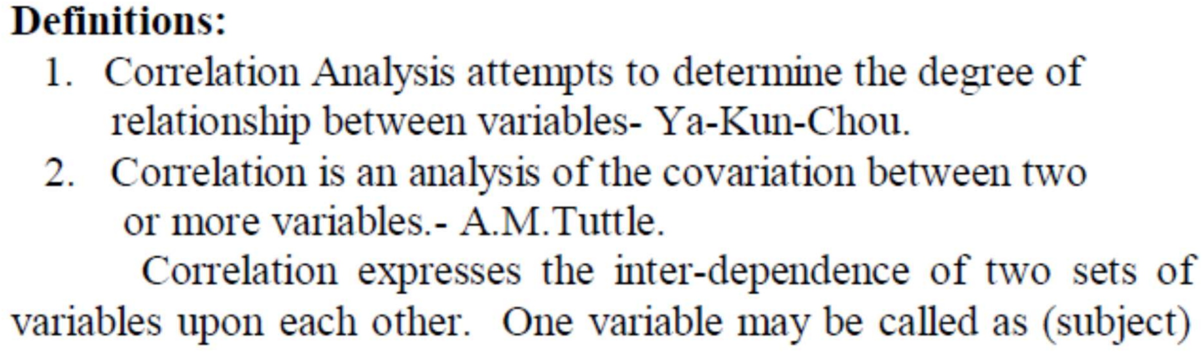


CORRELATION

Introduction:

The term correlation is used by a common man without knowing that he is making use of the term correlation. For example when parents advice their children to work hard so that they may get good marks, they are correlating good marks with hard work.

Thus Correlation refers to the relationship of two variables or more. (e-g) relation between height of father and son, yield and rainfall, wage and price index, share and debentures etc.



Definitions:

1. Correlation Analysis attempts to determine the degree of relationship between variables- Ya-Kun-Chou.
2. Correlation is an analysis of the covariation between two or more variables.- A.M.Tuttle.

Correlation expresses the inter-dependence of two sets of variables upon each other. One variable may be called as (subject) independent and the other relative variable (dependent). Relative variable is measured in terms of subject.

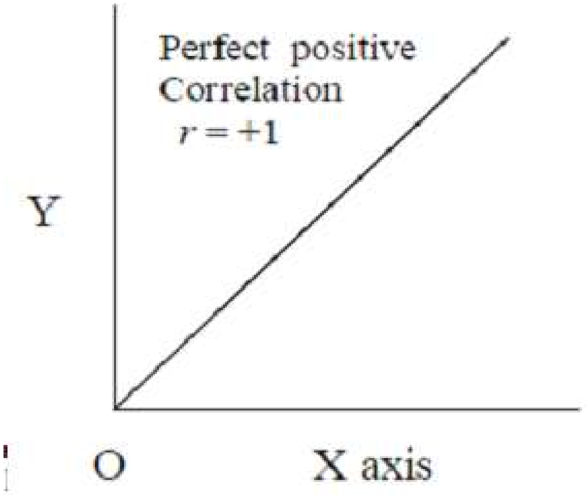
Uses of correlation:

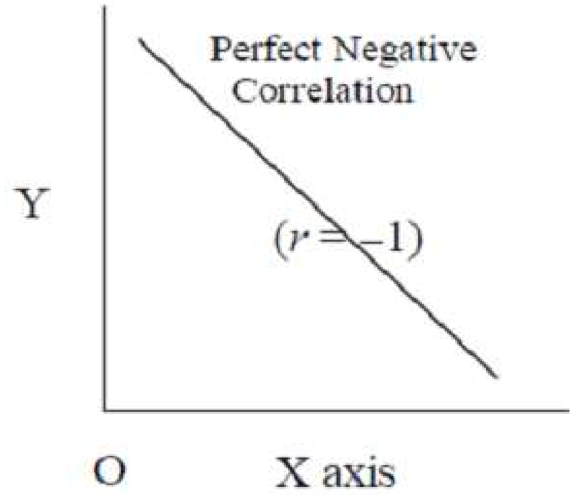
1. It is used in physical and social sciences.
2. It is useful for economists to study the relationship between variables like price, quantity etc. Businessmen estimates costs, sales, price etc. using correlation.
3. It is helpful in measuring the degree of relationship between the variables like income and expenditure, price and supply, supply and demand etc.
4. Sampling error can be calculated.
5. It is the basis for the concept of regression.

Scatter Diagram:

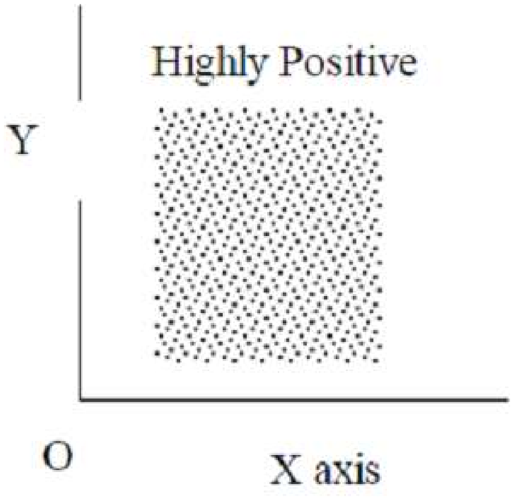
It is the simplest method of studying the relationship between two variables diagrammatically. One variable is represented along the horizontal axis and the second variable along the vertical axis. For each pair of observations of two variables, we put a dot in the plane. There are as many dots in the plane as the number of paired observations of two variables. The direction of dots shows the scatter or concentration of various points. This will show the type of correlation.

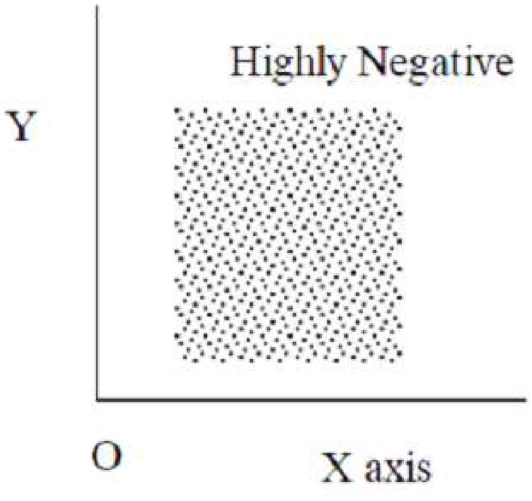
1. If all the plotted points form a straight line from lower left hand corner to the upper right hand corner then there is
   Perfect positive correlation. We denote this as $r = +1$





1. If all the plotted dots lie on a straight line falling from up left hand corner to lower right hand corner, there is a perf negative correlation between the two variables. In this ca the coefficient of correlation takes the value $r = -1$.
2. If the plotted points in the plane form a band and they sh a rising trend from the lower left hand corner to the up right hand corner the two variables are highly positiv correlated.

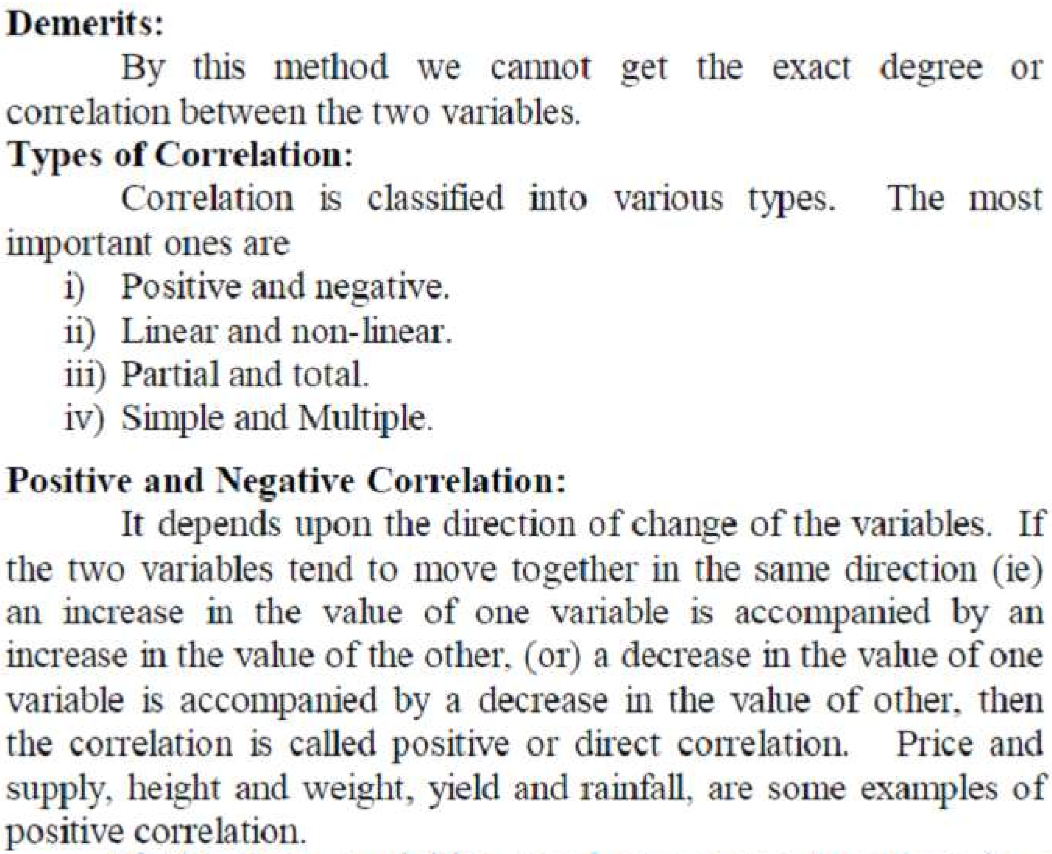




1. If the points fall in a narrow band from the upper left hand corner to the lower right hand corner, there will be a high degree of negative correlation.
2. If the plotted points in the plane are spread all over the diagram there is no correlation between the two variables.

Merits:

1. It is a simplest and attractive method of finding the nature of correlation between the two variables.
2. It is a non-mathematical method of studying correlation. It is easy to understand.
3. It is not affected by extreme items.
4. It is the first step in finding out the relation between the two variables.
5. We can have a rough idea at a glance whether it is a positive correlation or negative correlation.



Demerits:

By this method we cannot get the exact degree or correlation between the two variables.

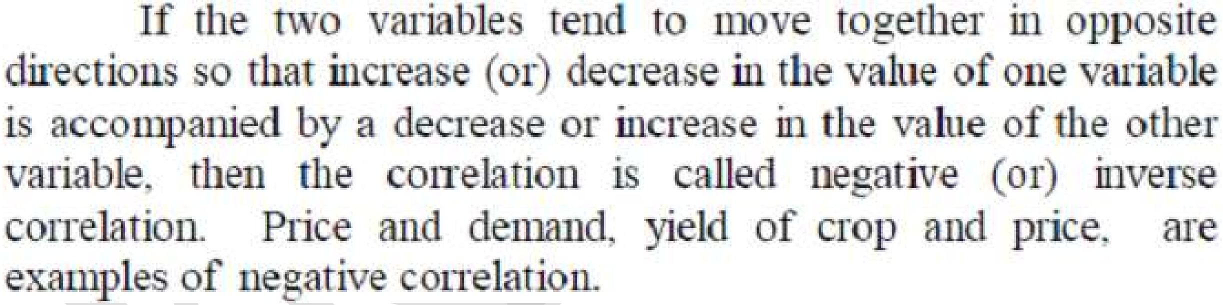
Types of Correlation:

Correlation is classified into various types. The most important ones are

i) Positive and negative.
ii) Linear and non-linear.
iii) Partial and total.
iv) Simple and Multiple.

Positive and Negative Correlation:

It depends upon the direction of change of the variables. If the two variables tend to move together in the same direction (ie) an increase in the value of one variable is accompanied by an increase in the value of the other, (or) a decrease in the value of one variable is accompanied by a decrease in the value of other, then the correlation is called positive or direct correlation. Price and supply, height and weight, yield and rainfall, are some examples of positive correlation.



If the two variables tend to move together in opposite directions so that increase (or) decrease in the value of one variable is accompanied by a decrease or increase in the value of the other variable, then the correlation is called negative (or) inverse correlation. Price and demand, yield of crop and price, are examples of negative correlation.

Linear and Non-linear correlation:

If the ratio of change between the two variables is a constant then there will be linear correlation between them.

Consider the following.

X	2	4	6	8	10	12
Y	3	6	9	12	15	18

Here the ratio of change between the two variables is the same. If we plot these points on a graph we get a straight line.

If the amount of change in one variable does not bear a constant ratio of the amount of change in the other. Then the relation is called Curvi-linear (or) non-linear correlation. The graph will be a curve.

Simple and Multiple correlation:

When we study only two variables, the relationship is simple correlation. For example, quantity of money and price level, demand and price. But in a multiple correlation we study more than two variables simultaneously. The relationship of price, demand and supply of a commodity are an example for multiple correlation.

Partial and total correlation:

The study of two variables excluding some other variable is called **Partial correlation**. For example, we study price and demand eliminating supply side. In total correlation all facts are taken into account.

Computation of correlation:

When there exists some relationship between two variables, we have to measure the degree of relationship. This measure is called the measure of correlation (or) correlation coefficient and it is denoted by ' r' .

Co-variation:

The covariation between the variables x and y is defined as

Cov(x,y) = $\frac{\Sigma(x-\overline{x})(y-\overline{y})}{n}$ where $\overline{x}, \overline{y}$ are respectively means of x and y and ' n' is the number of pairs of observations.

Karl pearson' s coefficient of correlation:

Karl pearson, a great biometrician and statistician, suggested a mathematical method for measuring the magnitude of linear relationship between the two variables. It is most widely used method in practice and it is known as pearsonian coefficient of correlation. It is denoted by 'r'. The formula for calculating 'r' is

(i) $r = \dfrac{Cov(x, y)}{\sigma_x . \sigma_y}$ where σ_x, σ_y are S.D of x and y respectively.

(ii) $r = \dfrac{\Sigma xy}{n \ \sigma_x \ \sigma_y}$

(iii) $r = \dfrac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}}$, $X = x - \overline{x}$, $Y = y - \overline{y}$

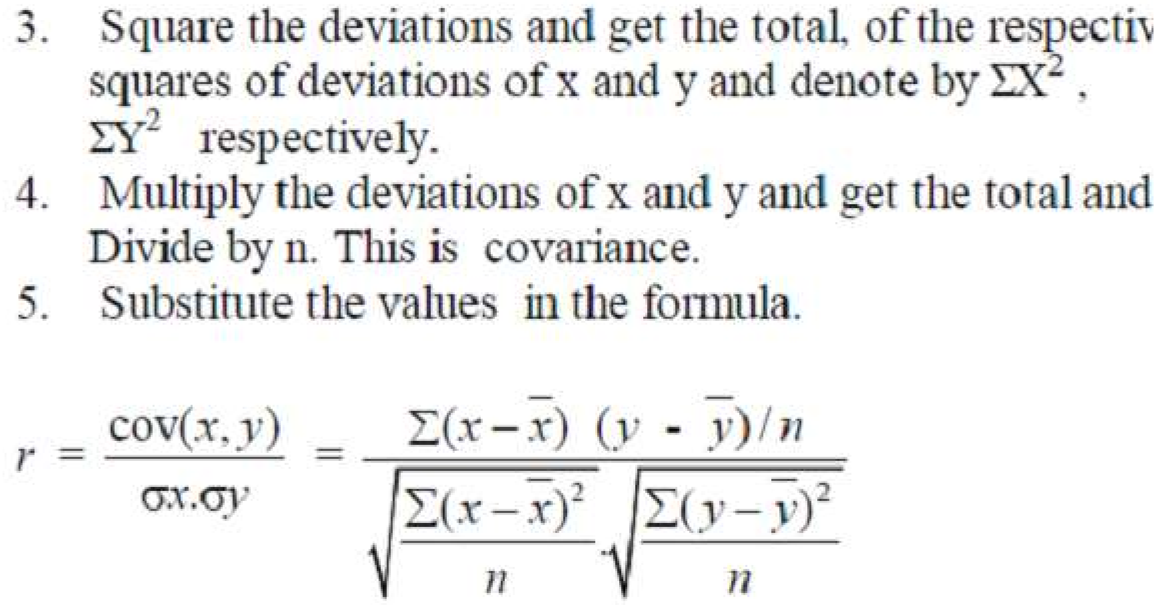
when the deviations are taken from the actual mean we can apply any one of these methods. Simple formula is the third one.

The third formula is easy to calculate, and it is not necessary to calculate the standard deviations of x and y series respectively.

Steps:

1. Find the mean of the two series x and y.
2. Take deviations of the two series from x and y.

   $X = x - \overline{x}$, $Y = y - \overline{y}$



3. Square the deviations and get the total, of the respectiv squares of deviations of x and y and denote by ΣX^2 , ΣY^2 respectively.
4. Multiply the deviations of x and y and get the total and Divide by n. This is covariance.
5. Substitute the values in the formula.

$$r = \frac{\operatorname{cov}(x,y)}{\sigma x.\sigma y} = \frac{\Sigma(x-\bar{x})\ (y - \bar{y})/n}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}.\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}}$$

KAHE

The above formula is simplified as follows

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}}, \quad X = x - \bar{x}\ , Y = y - \bar{y}$$

Example 1:

Find Karl Pearson's coefficient of correlation from the following data between height of father (x) and son (y).

X	64	65	66	67	68	69	70
Y	66	67	65	68	70	68	72

Comment on the result.

Solution:

x	Y	$X = x - \bar{x}$ $X = x - 67$	X^2	$Y = y - \bar{y}$ $Y = y - 68$	Y^2	XY
64	66	-3	9	-2	4	6
65	67	-2	4	-1	1	2
66	65	-1	1	-3	9	3
67	68	0	0	0	0	0
68	70	1	1	2	4	2
69	68	2	4	0	0	0
70	72	3	9	4	16	12
469	476	0	28	0	34	25

$$\bar{x} = \frac{469}{7} = 67 \ ; \ \bar{y} = \frac{476}{7} = 68$$

$$r = \frac{\Sigma XY}{\sqrt{\Sigma X^2 . \Sigma Y^2}} = \frac{25}{\sqrt{28 \times 34}} = \frac{25}{\sqrt{952}} = \frac{25}{30.85} = 0.81$$

Since r = + 0.81, the variables are highly positively correlated. (ie) Tall fathers have tall sons.

Working rule (i)

We can also find r with the following formula

We have $r = \dfrac{Cov(x, y)}{\sigma_x . \sigma_y}$

$$\text{Cov(x,y)} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{n} = \frac{\Sigma(xy + \bar{x}y - \bar{y}x - \bar{x}\bar{y})}{n}$$

$$= \frac{\Sigma xy}{n} - \frac{\bar{y}\Sigma x}{n} - \frac{\bar{x}\Sigma y}{n} + \frac{\Sigma \bar{x}\bar{y}}{n}$$

$$\text{Cov(x,y)} = \frac{\Sigma xy}{n} - \cancel{\bar{y}\bar{x}} - \bar{x}\bar{y} + \cancel{\bar{x}\bar{y}} = \frac{\Sigma xy}{n} - \bar{x}\bar{y}$$

$$\sigma^2 x^2 = \frac{\Sigma x^2}{n} - \bar{x}^2 , \; \sigma^2 y^2 = \frac{\Sigma y^2}{n} - \bar{y}^2$$

$$\text{Now } r = \frac{Cov(x,y)}{\sigma_x . \sigma_y}$$

$$r = \frac{\frac{\Sigma xy}{n} - \overline{xy}}{\sqrt{\left(\frac{\Sigma x^2}{n} - \bar{x}^2\right)} . \sqrt{\left(\frac{\Sigma y^2}{n} - \bar{y}^2\right)}}$$

$$r = \frac{n\Sigma xy - (\Sigma x)\ (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Note: In the above method we need not find mean or standard deviation of variables separately.

Example 2:

Calculate coefficient of correlation from the following data.

X	1	2	3	4	5	6	7	8	9
Y	9	8	10	12	11	13	14	16	15

x	y	x^2	y^2	xy
1	9	1	81	9
2	8	4	64	16
3	10	9	100	30
4	12	16	144	48
5	11	25	121	55
6	13	36	169	78
7	14	49	196	98
8	16	64	256	128
9	15	81	225	135
45	108	285	1356	597

$$r = \frac{n\Sigma xy \ - \ (\Sigma x) \ (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

$$r = \frac{9 \times 597 \ - \ 45 \times 108}{\sqrt{\left(9\times 285 - (45)^2\right).\left(9\times 1356 - (108)^2\right)}}$$

$$r = \frac{5373 \ - \ 4860}{\sqrt{(2565-2025).(12204-11664)}}$$

$$= \frac{513}{\sqrt{540 \times 540}} = \frac{513}{540} = 0.95$$

Working rule (ii) (shortcut method)

We have $r = \dfrac{Cov(x,y)}{\sigma_x.\sigma_y}$

where Cov(x,y) = $\dfrac{\Sigma(x-\bar{x})(y-\bar{y})}{n}$

Take the deviation from x as x – A and the deviation from y as y – B

$$Cov(x,y) = \frac{\Sigma\ [(x-A)-(\bar{x}-A)]\ [(y-B)-(\bar{y}-B)]}{n}$$

$$= \frac{1}{n}\ \Sigma\ [(x-A)\ (y-B)\ -\ (x-A)\ (\bar{y}-B)$$

$$-\ (\bar{x}-A)(y-B)+(\bar{x}-A)(\bar{y}-B)]$$

$$= \frac{1}{n}\ \Sigma\ [(x-A)\ (y-B)\ -\ (\bar{y}-B)\ \frac{\Sigma(x-A)}{n}$$

$$-\ (\bar{x}-A)\frac{\Sigma(y-B)}{n}+\frac{\Sigma(\bar{x}-A)(\bar{y}-B)}{n}$$

$$= \frac{\Sigma(x-A)(y-B)}{n}\ -\ (\bar{y}-B)\ (\bar{x}-\ \frac{nA}{n})$$

$$-\ (\bar{x}-A)\ (\bar{y}-\frac{nB}{n})+(\bar{x}-A)\ (\bar{y}-B)$$

$$= \frac{\Sigma(x - A)(y - B)}{n} - (\bar{y} - B)\ (\bar{x} - A)$$

$$- \cancel{(\bar{x} - A)\ (\bar{y} - B)} + \cancel{(\bar{x} - A)\ (\bar{y} - B)}$$

$$= \frac{\Sigma(x - A)(y - B)}{n} - (\bar{x} - A)\ (\bar{y} - B)$$

Let $x - A = u$; $y - B = v$; $\quad \bar{x} - A = \bar{u}$; $\bar{y} - B = \bar{v}$

$$\therefore \text{Cov}\,(x,y) = \frac{\Sigma uv}{n} - \overline{u}\,\overline{v}$$

$$\sigma\sigma_x^2 = \frac{\Sigma u^2}{n} - \bar{u}^2 = \sigma u^2$$

$$\sigma\sigma_y^2 = \frac{\Sigma \text{v}^2}{n} - \bar{v}^2 = \sigma v^2$$

$$\therefore r = \frac{n\Sigma uv - (\Sigma u)(\Sigma v)}{\sqrt{\left[n\Sigma u^2 - (\Sigma u)^2\right].\left[(n\Sigma v^2) - (\Sigma v)^2\right]}}$$

KAHE

Limitations:

1. Correlation coefficient assumes linear relationship regardless of the assumption is correct or not.
2. Extreme items of variables are being unduly operated on correlation coefficient.
3. Existence of correlation does not necessarily indicate cause-effect relation.

Interpretation:

The following rules helps in interpreting the value of ‘ r’ .

1. When r = 1, there is perfect +ve relationship between the variables.
2. When r = -1, there is perfect –ve relationship between the variables.
3. When r = 0, there is no relationship between the variables.
4. If the correlation is +1 or –1, it signifies that there is a high degree of correlation. (+ve or –ve) between the two variables.

If r is near to zero (ie) 0.1,-0.1, (or) 0.2 there is less correlation.

Rank Correlation:

It is studied when no assumption about the parameters of the population is made. This method is based on ranks. It is useful to study the qualitative measure of attributes like honesty, colour, beauty, intelligence, character, morality etc.The individuals in the group can be arranged in order and there on, obtaining for each individual a number showing his/her rank in the group. This method was developed by Edward Spearman in 1904. It is defined

as $r = 1 - \frac{6\Sigma D^2}{n^3 - n}$ r = rank correlation coefficient.

Note: Some authors use the symbol ρ for rank correlation.
ΣD^2 = sum of squares of differences between the pairs of ranks.
n = number of pairs of observations.

The value of r lies between –1 and +1. If r = +1, there is complete agreement in order of ranks and the direction of ranks is also same. If r = -1, then there is complete disagreement in order of ranks and they are in opposite directions.

Computation for tied observations: There may be two or more items having equal values. In such case the same rank is to be given. The ranking is said to be tied. In such circumstances an average rank is to be given to each individual item. For example if the value so is repeated twice at the 5th rank, the common rank to be assigned to each item is $\frac{5+6}{2} = 5.5$ which is the average of 5 and 6 given as 5.5, appeared twice.

If the ranks are tied, it is required to apply a correction factor which is $\frac{1}{12}$ $(m^3\text{-}m)$. A slightly different formula is used when there is more than one item having the same value.

The formula is

$$r = 1 - \frac{6[\Sigma D^2 + \frac{1}{12}(m^3 - m) + \frac{1}{12}(m^3 - m) + ...]}{n^3 - n}$$

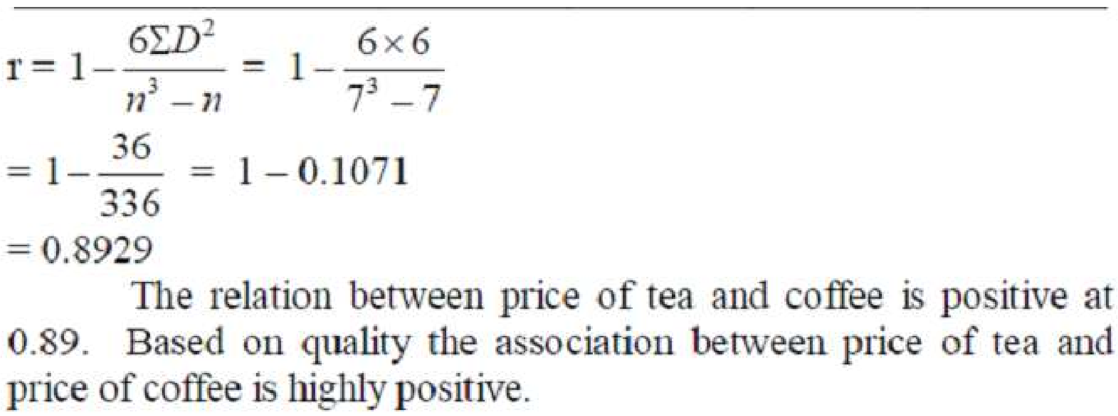
Where m is the number of items whose ranks are common and should be repeated as many times as there are tied observations.

Example 6:

In a marketing survey the price of tea and coffee in a town based on quality was found as shown below. Could you find any relation between and tea and coffee price.

Price of tea	88	90	95	70	60	75	50
Price of coffee	120	134	150	115	110	140	100

Price of tea	Rank	Price of coffee	Rank	D	D^2
88	3	120	4	1	1
90	2	134	3	1	1
95	1	150	1	0	0
70	5	115	5	0	0
60	6	110	6	0	0
75	4	140	2	2	4
50	7	100	7	0	0
					$\Sigma D^2 = 6$



$$r = 1-\frac{6\Sigma D^2}{n^3-n} = 1-\frac{6\times 6}{7^3-7}$$

$$= 1-\frac{36}{336} = 1-0.1071$$

$$= 0.8929$$

The relation between price of tea and coffee is positive at 0.89. Based on quality the association between price of tea and price of coffee is highly positive.

REGRESSION

In mathematics, regression is one of the important topics in statistics. The process of determining the relationship between two variables is called as regression. It is also one of the statistical analysis methods that can be used to assessing the association between the two different variables

9.1 Introduction:

After knowing the relationship between two variables we may be interested in estimating (predicting) the value of one variable given the value of another. The variable predicted on the basis of other variables is called the "dependent" or the ' explained' variable and the other the ' independent' or the ' predicting' variable. The prediction is based on average relationship derived statistically by regression analysis. The equation, linear or otherwise, is called the regression equation or the explaining equation.

For example, if we know that advertising and sales are correlated we may find out expected amount of sales for a given advertising expenditure or the required amount of expenditure for attaining a given amount of sales.

The relationship between two variables can be considered between, say, rainfall and agricultural production, price of an input and the overall cost of product, consumer expenditure and disposable income. Thus, regression analysis reveals average relationship between two variables and this makes possible estimation or prediction.

9.1.1 Definition:

Regression is the measure of the average relationship between two or more variables in terms of the original units of the data.

9.2 Types Of Regression:

The regression analysis can be classified into:

a) Simple and Multiple
b) Linear and Non –Linear
c) Total and Partial

a) Simple and Multiple:

In case of simple relationship only two variables are considered, for example, the influence of advertising expenditure on sales turnover. In the case of multiple relationship, more than two variables are involved. On this while one variable is a dependent variable the remaining variables are independent ones.

For example, the turnover (y) may depend on advertising expenditure (x) and the income of the people (z). Then the functional relationship can be expressed as $y = f(x,z)$.

b) Linear and Non-linear:

The linear relationships are based on straight-line trend, the equation of which has no-power higher than one. But, remember a linear relationship can be both simple and multiple. Normally a linear relationship is taken into account because besides its simplicity, it has a better predective value, a linear trend can be easily projected into the future. In the case of non-linear relationship curved trend lines are derived. The equations of these are parabolic.

c) Total and Partial:

In the case of total relationships all the important variables are considered. Normally, they take the form of a multiple relationships because most economic and business phenomena are affected by multiplicity of cases. In the case of partial relationship one or more variables are considered, but not all, thus excluding the influence of those not found relevant for a given purpose.

9.3 Linear Regression Equation:

If two variables have linear relationship then as the independent variable (X) changes, the dependent variable (Y) also changes. If the different values of X and Y are plotted, then the two straight lines of best fit can be made to pass through the plotted points. These two lines are known as regression lines. Again, these regression lines are based on two equations known as regression equations. These equations show best estimate of one variable for the known value of the other. The equations are linear.

Linear regression equation of Y on X is

$$Y = a + bX \;......(1)$$

And X on Y is

$$X = a + bY \;......(2)$$

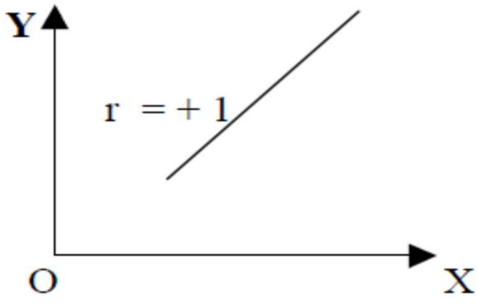
a, b are constants.

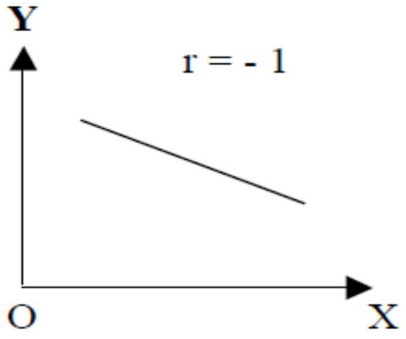
From (1) We can estimate Y for known value of X.
(2) We can estimate X for known value of Y.

9.3.1 Regression Lines:

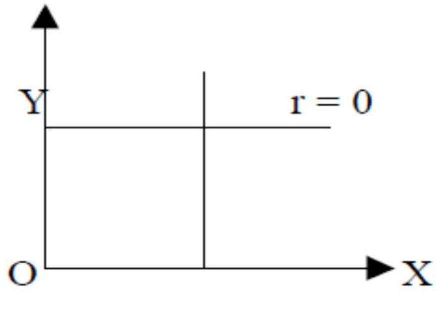
For regression analysis of two variables there are two regression lines, namely Y on X and X on Y. The two regression lines show the average relationship between the two variables.

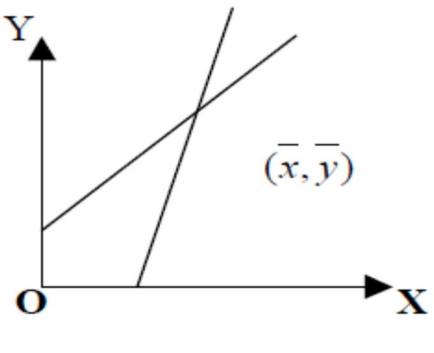
For perfect correlation, positive or negative i.e., r = $\underline{+}$ 1, the two lines coincide i.e., we will find only one straight line. If r = 0, i.e., both the variables are independent then the two lines will cut each other at right angle. In this case the two lines will be parallel to X and Y-axes.





Lastly the two lines intersect at the point of means of X and Y. From this point of intersection, if a straight line is drawn on X-axis, it will touch at the mean value of x. Similarly, a perpendicular drawn from the point of intersection of two regression lines on Y-axis will touch the mean value of Y.





9.3.2 Principle of ' Least Squares' :

Regression shows an average relationship between two variables, which is expressed by a line of regression drawn by the method of "least squares". This line of regression can be derived graphically or algebraically. Before we discuss the various methods let us understand the meaning of least squares.

A line fitted by the method of least squares is known as the line of best fit. The line adapts to the following rules:

(i) The algebraic sum of deviation in the individual observations with reference to the regression line may be equal to zero. i.e.,

$\Sigma(X - Xc) = 0$ or $\Sigma (Y- Yc) = 0$

Where Xc and Yc are the values obtained by regression analysis.

(ii) The sum of the squares of these deviations is less than the sum of squares of deviations from any other line. i.e.,

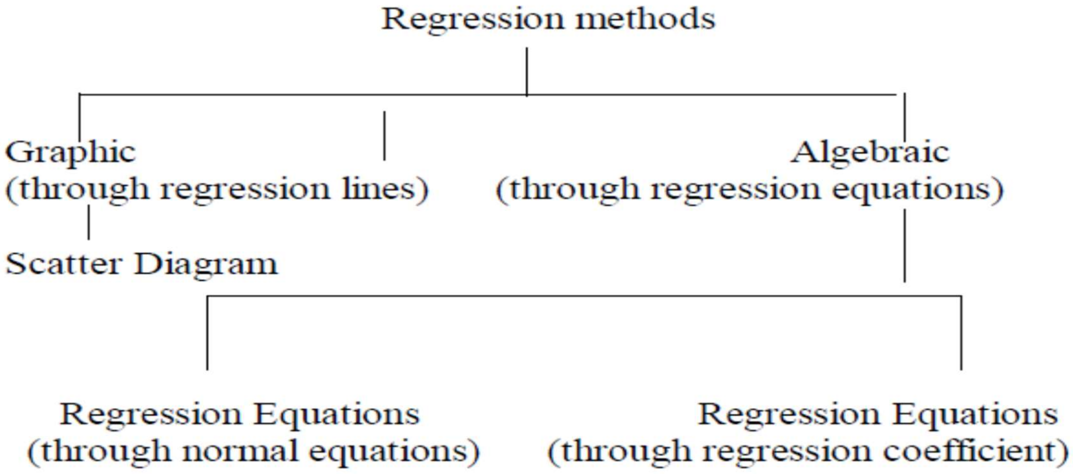
$\Sigma(Y - Yc)^2 < \Sigma (Y - Ai)^2$

Where Ai = corresponding values of any other straight line.

(iii) The lines of regression (best fit) intersect at the mean values of the variables X and Y, i.e., intersecting point is $\overline{x}, \overline{y}$.

9.4 Methods of Regression Analysis:

The various methods can be represented in the form of chart given below:


Regression methods
Graphic
(through regression lines)
Algebraic
(through regression equations)
Scatter Diagram
Regression Equations
(through normal equations)
Regression Equations
(through regression coefficient)

9.4.1 Graphic Method:

Scatter Diagram:

Under this method the points are plotted on a graph paper representing various parts of values of the concerned variables. These points give a picture of a scatter diagram with several points spread over. A regression line may be drawn in between these points either by free hand or by a scale rule in such a way that the squares of the vertical or the horizontal distances (as the case may be) between the points and the line of regression so drawn is the least. In other words, it should be drawn faithfully as the line of best fit leaving equal number of points on both sides in such a manner that the sum of the squares of the distances is the best.

9.4.2 Algebraic Methods:

(i) Regression Equation.

The two regression equations

for X on Y; $X = a + bY$

And for Y on X; $Y = a + bX$

Where X, Y are variables, and a,b are constants whose values are to be determined

For the equation, $X = a + bY$

The normal equations are

$$\sum X = na + b\sum Y \text{ and}$$

$$\sum XY = a\sum Y + b\sum Y^2$$

For the equation, $Y = a + bX$, the normal equations are

$$\sum Y = na + b\sum X \text{ and}$$

$$\sum XY = a\sum X + b\sum X^2$$

From these normal equations the values of a and b can be determined.

Example 1:

Find the two regression equations from the following data:

X:	6	2	10	4	8
Y:	9	11	5	8	7

Solution:

X	Y	X^2	Y^2	XY
6	9	36	81	54
2	11	4	121	22
10	5	100	25	50
4	8	16	64	32
8	7	64	49	56
30	40	220	340	214

Regression equation of Y on X is $Y = a + bX$ and the normal equations are

$$\sum Y = na + b\sum X$$
$$\sum XY = a\sum X + b\sum X^2$$

Substituting the values, we get

$$40 \;=\; 5a + 30b \;\ldots\ldots\; (1)$$
$$214 = 30a + 220b \;\ldots\ldots(2)$$

Multiplying (1) by 6

$$240 = 30a + 180b\ldots\ldots(3)$$

(2) – (3) $\quad - 26 = 40b$

$$\text{or } b = -\frac{26}{40} = -0.65$$

Now, substituting the value of ‘ b’ in equation (1)

$$40 = 5a - 19.5$$
$$5a = 59.5$$
$$a = \frac{59.5}{5} = 11.9$$

Hence, required regression line Y on X is $Y = 11.9 - 0.65X$.
Again, regression equation of X on Y is

$X = a + bY$ and

The normal equations are

$$\sum X = na + b\sum Y \text{ and}$$
$$\sum XY = a\sum Y + b\sum Y^2$$

Now, substituting the corresponding values from the above table, we get

$$30 = 5a + 40b \ldots(3)$$
$$214 = 40a + 340b \ldots(4)$$

Multiplying (3) by 8, we get

$$240 = 40a + 320\, b \ldots(5)$$

(4) – (5) gives

$$-26 = 20b$$
$$b = -\frac{26}{20} = -1.3$$

Substituting $b = -1.3$ in equation (3) gives

$$30 = 5a - 52$$
$$5a = 82$$
$$a = \frac{82}{5} = 16.4$$

Hence, Required regression line of X on Y is

X = 16.4 – 1.3Y

(ii) Regression Co-efficents:

The regression equation of Y on X is $y_e = \overline{y} + r\frac{\sigma_y}{\sigma_x}(x - \overline{x})$

Here, the regression Co.efficient of Y on X is

$$b_1 = b_{yx} = r\frac{\sigma_y}{\sigma_x}$$
$$y_e = \overline{y} + b_1(x - \overline{x})$$

The regression equation of X on Y is

$$X_e = \overline{x} + r\frac{\sigma_x}{\sigma_y}(y - \overline{y})$$

Here, the regression Co-efficient of X on Y

$$b_2 = b_{xy} = r\frac{\sigma_x}{\sigma_y}$$
$$X_e = \overline{X} + b_2(y - \overline{y})$$

If the deviation are taken from respective means of x and y

$$b_1 = b_{yx} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (X-\overline{X})^2} = \frac{\sum xy}{\sum x^2} \quad \text{and}$$

$$b_2 = b_{xy} = \frac{\sum (X-\overline{X})(Y-\overline{Y})}{\sum (Y-\overline{Y})^2} = \frac{\sum xy}{\sum y^2}$$

where $x = X - \overline{X}, y = Y - \overline{Y}$

If the deviations are taken from any arbitrary values of x and y (short – cut method)

$$b_1 = b_{yx} = \frac{n\sum uv - \sum u \sum v}{n\sum u^2 - \left(\sum u\right)^2}$$

$$b_2 = b_{xy} = \frac{n\sum uv - \sum u \sum v}{n\sum v^2 - \left(\sum v\right)^2}$$

where u = x – A : v = Y-B
A = any value in X
B = any value in Y

9.5 Properties of Regression Co-efficient:

1. Both regression coefficients must have the same sign, ie either they will be positive or negative.
2. correlation coefficient is the geometric mean of the regression coefficients ie, $r = \pm\sqrt{b_1 b_2}$
3. The correlation coefficient will have the same sign as that of the regression coefficients.
4. If one regression coefficient is greater than unity, then other regression coefficient must be less than unity.
5. Regression coefficients are independent of origin but not of scale.
6. Arithmetic mean of b_1 and b_2 is equal to or greater than the coefficient of correlation. Symbolically $\frac{b_1 + b_2}{2} \geq r$

7. If r=0, the variables are uncorrelated , the lines of regression become perpendicular to each other.
8. If r= $\pm$1, the two lines of regression either coincide or parallel to each other
9. Angle between the two regression lines is $\theta = \tan^{-1}\left[\frac{m_1 - m_2}{1 + m_1 m_2}\right]$

   where m_1 and,m_2 are the slopes of the regression lines X on Y and Y on X respectively.
10. The angle between the regression lines indicates the degree of dependence between the variables.

Example 2:

If 2 regression coefficients are $b_1=\frac{4}{5}$ and $b_2 = \frac{9}{20}$ .What would be the value of r?

Solution:

The correlation coefficient , $r = \pm\sqrt{b_1 b_2}$

$$= \sqrt{\frac{4}{5} \text{x} \frac{9}{20}}$$

$$= \sqrt{\frac{36}{100}} = \frac{6}{10} = 0.6$$

Example 3:

Given $b_1 = \frac{15}{8}$ and $b_2 = \frac{3}{5}$, Find r

Solution:

$r = \pm\sqrt{b_1 b_2}$

$= \sqrt{\frac{15}{8} \text{x} \frac{3}{5}}$

$= \sqrt{\frac{9}{8}} \quad =1.06$

It is not possible since *r*, cannot be greater than one. So the given values are wrong

Example 4:

Compute the two regression equations from the following data.

X	1	2	3	4	5
Y	2	3	5	4	6

If x =2.5, what will be the value of y?

Solution:

X	Y	$x = X - \overline{X}$	$y = Y - \overline{Y}$	x^2	y^2	xy
1	2	-2	-2	4	4	4
2	3	-1	-1	1	1	-1
3	5	0	1	0	1	0
4	4	1	0	1	0	0
5	6	2	2	4	4	4
15	20	20		10	10	9

$$\overline{X} = \frac{\sum X}{n} = \frac{15}{5} = 3$$

$$\overline{Y} = \frac{\sum Y}{n} = \frac{20}{5} = 4$$

Regression Co efficient of Y on X

$$b_{yx} = \frac{\sum xy}{\sum x^2} = \frac{9}{10} = 0.9$$

Hence regression equation of Y on X is

$$Y = \overline{Y} + b_{yx}(X - \overline{X})$$

$$= 4 + 0.9 \ (X - 3)$$

$$= 4 + 0.9X - 2.7$$

$$=1.3 + 0.9X$$

when X = 2.5

$$Y = 1.3 + 0.9 \times 2.5$$

$$= 3.55$$

Regression co efficient of X on Y

$$b_{xy} = \frac{\sum xy}{\sum y^2} = \frac{9}{10} = 0.9$$

So, regression equation of X on Y is

$$X = \overline{X} + b_{xy}(Y - \overline{Y})$$
$$= 3 + 0.9\,(Y - 4)$$
$$= 3 + 0.9Y - 3.6$$
$$= 0.9Y - 0.6$$

Example 6:

In a correlation study, the following values are obtained

	X	Y
Mean	65	67
S.D	2.5	3.5

Co-efficient of correlation = 0.8
Find the two regression equations that are associated with the above values.

Solution:

Given,

$\overline{X} = 65$, $\overline{Y} = 67$, $\sigma_x = 2.5$, $\sigma_y = 3.5$, $r = 0.8$

The regression co-efficient of Y on X is

$$b_{yx} = b_1 = r\frac{\sigma_y}{\sigma_x}$$
$$= 0.8 \times \frac{3.5}{2.5} = 1.12$$

The regression coefficient of X on Y is

$$b_{xy} = b_2 = r\,\frac{\sigma_x}{\sigma_y}$$

$$= 0.8 \times \frac{2.5}{3.5} = 0.57$$

Hence, the regression equation of Y on X is

$$Y_e = \overline{Y} + b_1(X - \overline{X})$$

$$= 67 + 1.12\ (X-65)$$

$$= 67 + 1.12\ X - 72.8$$

$$= 1.12X - 5.8$$

The regression equation of X on Y is

$$X_e = \overline{X} + b_2(Y - \overline{Y})$$

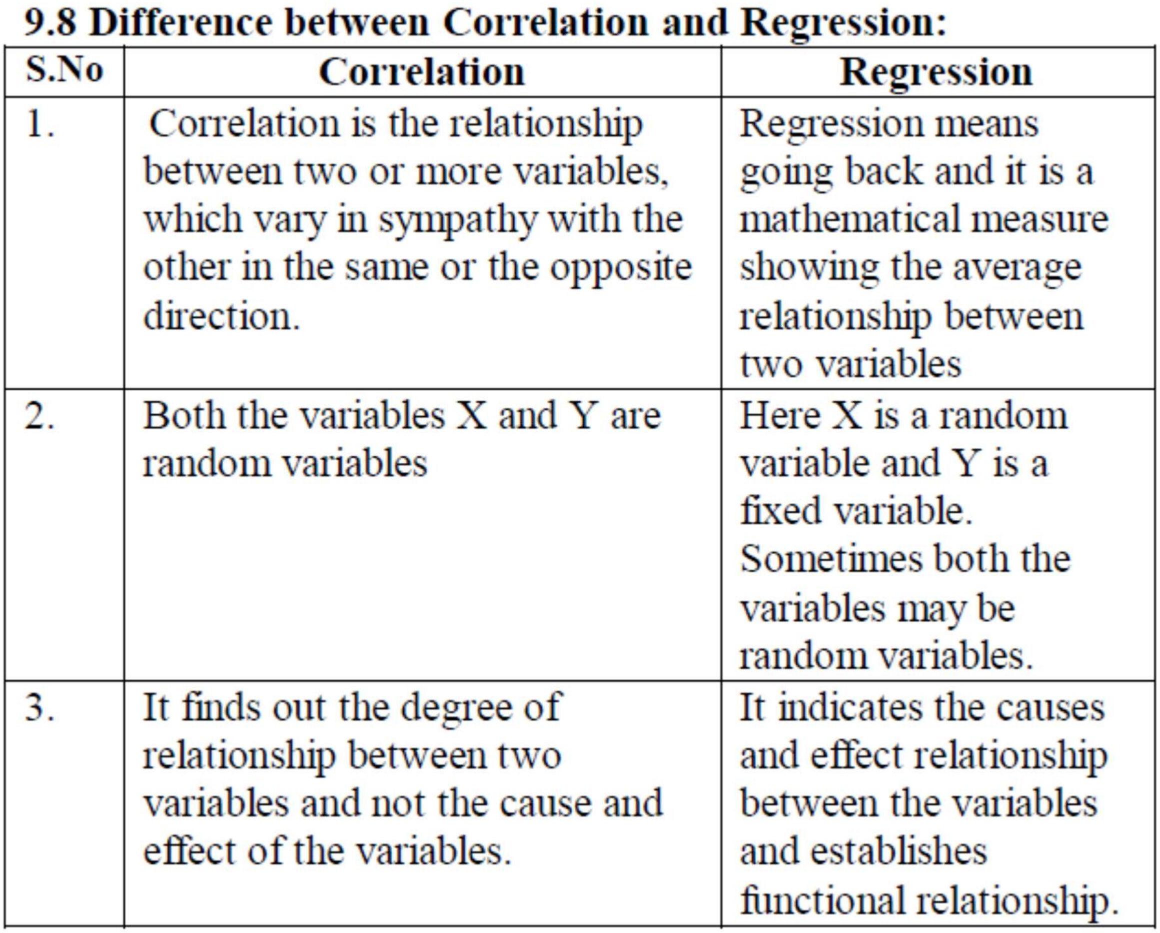
$$= 65 + 0.57\ (Y-67)$$

$$= 65 + 0.57Y - 38.19$$

$$= 26.81 + 0.57Y$$

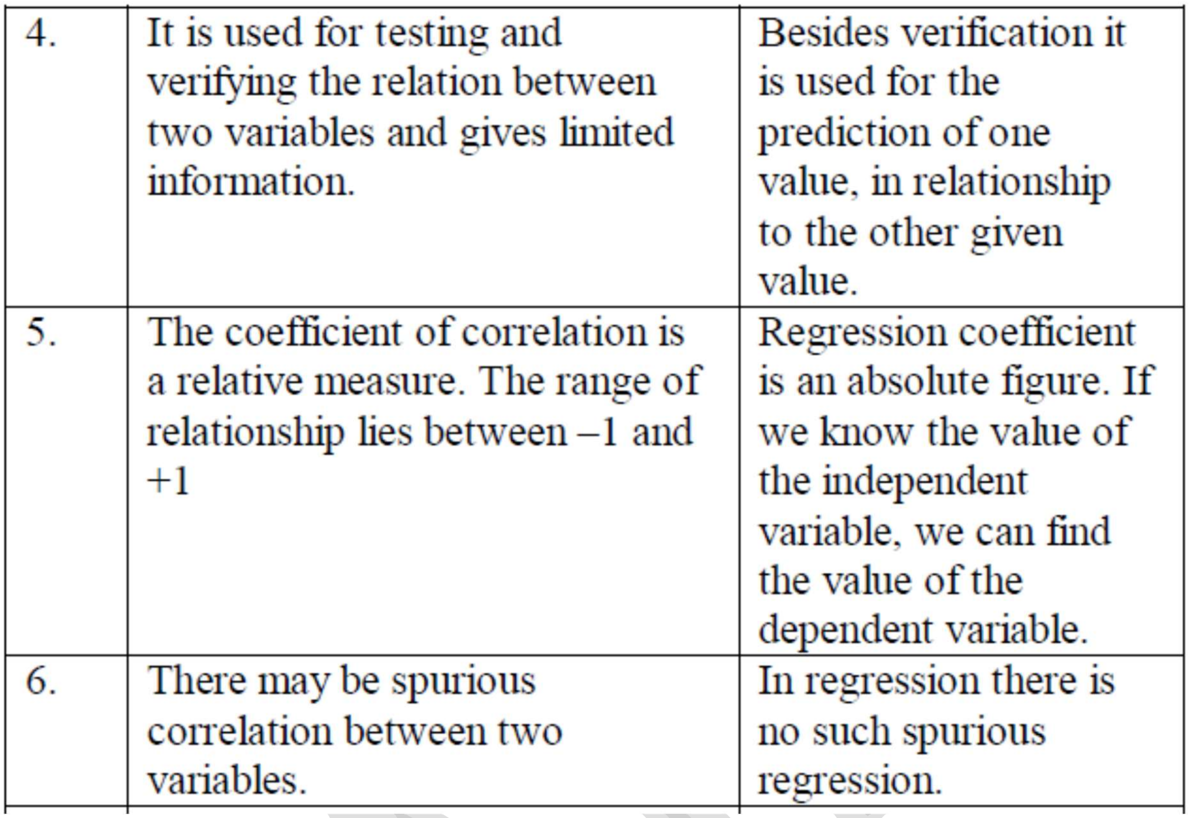
9.7 Uses of Regression Analysis:

1. Regression analysis helps in establishing a functional relationship between two or more variables.
2. Since most of the problems of economic analysis are based on cause and effect relationships, the regression analysis is a highly valuable tool in economic and business research.
3. Regression analysis predicts the values of dependent variables from the values of independent variables.
4. We can calculate coefficient of correlation (r) and coefficient of determination (r^2) with the help of regression coefficients.
5. In statistical analysis of demand curves, supply curves, production function, cost function, consumption function etc., regression analysis is widely used.



9.8 Difference between Correlation and Regression:

S.No	Correlation	Regression
1.	Correlation is the relationship between two or more variables, which vary in sympathy with the other in the same or the opposite direction.	Regression means going back and it is a mathematical measure showing the average relationship between two variables
2.	Both the variables X and Y are random variables	Here X is a random variable and Y is a fixed variable. Sometimes both the variables may be random variables.
3.	It finds out the degree of relationship between two variables and not the cause and effect of the variables.	It indicates the causes and effect relationship between the variables and establishes functional relationship.



4.	It is used for testing and verifying the relation between two variables and gives limited information.	Besides verification it is used for the prediction of one value, in relationship to the other given value.
5.	The coefficient of correlation is a relative measure. The range of relationship lies between –1 and +1	Regression coefficient is an absolute figure. If we know the value of the independent variable, we can find the value of the dependent variable.
6.	There may be spurious correlation between two variables.	In regression there is no such spurious regression.

POSSIBLE QUESTIONS

PART – B (TWO MARKS)

1.What are the features of Spearman's correlation coefficient?
2.What do you mean by regression equations?
3.State Rank Correlation.
4.Define standard error.
5.Define Chi-Square test.

PART – C (FIVE MARKS)

1. In 600 throws of a six faced dice, odd points appeared 360 times. would you say that the dice is fair at 5% level of significance?

2.Sample of two different types of bulbs were tested for length of life,and the following data were obtained:

	Type I	Type II
Sample size	8	7
Sample mean	1234 hours	1136 hours
Sample of SD	36 hours	40 hours

Is the difference in the means significant?
(Given that the significant value of t at 5 % level of significance for 13 d.f. is 2.16)

3.In a hospital 480 female and 520 male babies were born in a week. Do these figures confirm the hypothesis that males and females are born in equal number?

4.You are given the following data:

	X	Y
Arithmetic mean	36	85
Standard deviation	11	8

Correlation coefficient between X and Y = 0.66
i) Find the two regression equations. ii) Find r.

5.Marks obtained by 7 students in Accountancy (X) and Statistics (Y) are given below. Compute Rank Correlation Coefficient.

X	15	20	28	12	40	60	20
Y	40	30	50	30	20	10	30

6.For the following data calculate the rank correlation coefficient between X and Y.

X :	1	2	3	4	5	6	7	8	9	10	11	12
Y :	12	9	6	10	3	5	4	7	8	2	11	1

KAHE

7.Calculate the two regression equation from the following data:

X :	10	12	13	12	16	15
Y :	40	38	43	45	37	43

8.From the following data calculate the regression equations taking deviation of items from the mean of X and Y series:

X:	6	2	10	4	8
Y:	9	11	5	8	7

PART- D (TEN Marks)

1. Calculate Karl Pearson's correlation coefficient between the marks in English and Hindi obtained by 10 students:

Marks in English :	10	25	13	25	22	11	12	25	21	20
Marks in Hindi:	12	22	16	15	18	18	17	23	24	17

2.In a village 'A' out of random sample of 1000 persons 100 were found to be vegetarians while in another village 'B' out of 1500 persons 180 were to be vegetarians.Do you find a significant difference in the food habits of the people of the two villages?

KAHE

UNIT – V

Hypothesis testing – Errors in Hypothesis testing - large sample test (Z – test) single and two mean test, Small sample test (t – test)-Single mean-Two mean- Chi–square test –Goodness of fit, ANOVA – one way-f-test.

KAHE

Hypothesis:

A statistical hypothesis is an assumption that we make about a population parameter, which may or may not be true concerning one or more variables.

According to Prof. Morris Hamburg "A hypothesis in statistics is simply a quantitative statement about a population".

Hypothesis testing:

Hypothesis testing is to test some hypothesis about parent population from which the sample is drawn.

Example:

A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased.

To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

Null hypothesis:

The hypothesis under verification is known as *null hypothesis* and is denoted by H_0 and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that *"extra coaching has not benefited the students"*. Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that *"the drug is not effective in curing malaria*".

Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis H_0 is called alternative hypothesis and is denoted by H_1 or H_a.

For example, if a psychologist who wishes to test whether or not a certain class of people have a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$H_0 : \mu = 100$

Then the alternative hypothesis could be any one of the statements.

$$\begin{aligned} H_1 &: \mu \neq 100 \\ (or) H_1 &: \mu > 100 \\ (or) H_1 &: \mu < 100 \end{aligned}$$

Errors in testing of hypothesis:

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)
3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e.1) Type-I error

2) Type-II error

1) Type-I error: The type-I error is said to be committed if the null hypothesis (H_0) is true but our test rejects it.

2) Type-II error: The type-II error is said to be committed if the null hypothesis (H_0) is false but our test accepts it.

Level of significance:

The maximum probability of committing type-I error is called level of significance and is denoted by α .

$$\alpha = \text{P (Committing Type-I error)}$$

$$= \text{P (H}_0 \text{ is rejected when it is true)}$$

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc……

Power of the test:

The probability of rejecting a false hypothesis is called power of the test and is denoted by $1-\beta$ **.**

Power of the test =P (H_0 is rejected when it is false)

= 1- P (H_0 is accepted when it is false)

= 1- P (Committing Type-II error)

= 1- β

- A test for which both α and β are small and kept at minimum level is considered desirable.
- The only way to reduce both α and β simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

Critical region:

A statistic is used to test the hypothesis H_0. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It indicates that if the value of test statistic lies in this region, H_0 will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance α . The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

KAHE

One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called a two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (right tailed alternative) or $H_1 : \theta < \theta_0$ (left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ ------- right tailed test

$H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ ------- left tailed test

Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^{N}c_n$ possible samples. If we calculate some particular statistic from each of the ${}^{N}c_n$ samples, the distribution of sample statistic is called sampling

distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

i.e. S.E (t)= $\sqrt{Var(t)}$

Utility of standard error:

1. It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z| = \left| \frac{t - E(t)}{S.E(t)} \right| > 1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.
2. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.

3. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
4. It is used to determine the size of the sample.

Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

$$\text{i.e. test statistic } Z = \frac{t - E(t)}{S.E(t)}$$

Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0 : \theta = \theta_0$.
2. Set up a alternative hypothesis i.e. $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
3. Choose the level of significance i.e. α .
4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at α % l.o.s i.e. Z_α .
7. Compare the test statistic value with the tabulated value at α % l.o.s. and make a decision whether to accept or to reject the null hypothesis.

Large sample tests:

The sample size which is greater than or equal to 30 is called as large sample and the test depending on large sample is called large sample test.

The assumption made while dealing with the problems relating to large samples are

Assumption-1: The random sampling distribution of the statistic is approximately normal.

Assumption-2: Values given by the sample are sufficiently closed to the population value and can be used on its place for calculating the standard error of the statistic.

KAHE

Large sample test for single mean (or) test for significance of single mean:

For this test

The null hypothesis is $H_0 : \mu = \mu_0$

against the two sided alternative $H_1 : \mu \neq \mu_0$

where μ is population mean

μ_0 is the value of μ

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots, x_n$ be a random sample from a normal population with mean μ and variance σ^2

i.e. if $X \sim N(\mu, \sigma^2)$ then $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, Where $\bar{x}$ be the sample mean

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H$_0$

If $|Z| < Z_\alpha$, accept the null hypothesis H$_0$

Note: if the population standard deviation is unknown then we can use its estimate s, which will be calculated from the sample. $s = \sqrt{\frac{1}{n-1}\sum (x - \bar{x})^2}$.

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for the first and second populations respectively

Then $\bar{x}_1 \sim N\left(\mu_1, {\sigma_1{}^2}/{n_1}\right)$ and $\bar{x}_2 \sim N\left(\mu_2, {\sigma_2{}^2}/{n_2}\right)$

Therefore $\bar{x}_1$ - $\bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}\right)$

For this test

The null hypothesis is $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$
against the two sided alternative $H_1 : \mu_1 \neq \mu_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{\sigma_1{}^2}{n_1} + \frac{\sigma_2{}^2}{n_2}}} \sim N(0,1) \text{[since } \mu_1 - \mu_2 \text{=0 from H}_0\text{]}$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H0

If $|Z| < Z_\alpha$, accept the null hypothesis H0

Note: If $\sigma_1{}^2$ and $\sigma_2{}^2$ are unknown then we can consider $S_1{}^2$ and $S_2{}^2$ as the estimate value of $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively..

Large sample test for single standard deviation (or) test for significance of standard deviation:

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots\ldots\ldots, x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2 ,

for large sample, sample standard deviation s follows a normal distribution with mean σ and variance $\sigma^2/2n$ i.e. $s \sim N\left(\sigma, \sigma^2/2n\right)$

For this test

The null hypothesis is $H_0 : \sigma = \sigma_0$
against the two sided alternative $H_1 : \sigma \neq \sigma_0$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{s - E(s)}{S.E(s)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{s - \sigma}{\sigma/\sqrt{2n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H0

If $|Z| < Z_\alpha$, accept the null hypothesis H0

Large sample test for difference between two standard deviations:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances $\sigma_1{}^2$ and $\sigma_2{}^2$ respectively

Let s_1 and s_2 be the sample standard deviations for the first and second populations respectively

Then $s_1 \sim N\left(\sigma_1, {\sigma_1}^2 \Big/ 2n_1\right)$ and $\bar{x}_2 \sim N\left(\sigma_2, {\sigma_2}^2 \Big/ 2n_2\right)$

Therefore $s_1 - s_2 \sim N\left(\sigma_1 - \sigma_2, \frac{\sigma_1{}^2}{2n_1} + \frac{\sigma_2{}^2}{2n_2}\right)$

For this test

The null hypothesis is $H_0 : \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$
against the two sided alternative $H_1 : \sigma_1 \neq \sigma_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{\sigma_1{}^2}{2n_1} + \frac{\sigma_2{}^2}{2n_2}}} \sim N(0,1) \text{[since } \sigma_1 - \sigma_2 \text{=0 from H}_0\text{]}$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

Let x is number of success in n independent trails with constant probability p, then x follows a binomial distribution with mean np and variance npq.

In a sample of size n let x be the number of persons processing a given attribute then the sample proportion is given by $\hat{p} = \frac{x}{n}$

Then $E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}np = p$

And $V(\hat{p}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}npq = \frac{pq}{n}$

$$S.E(\hat{p}) = \sqrt{\frac{pq}{n}}$$

For this test

The null hypothesis is $H_0 : p = p_0$

against the two sided alternative $H_1 : p \neq p_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p} - E(\hat{p})}{S.E(\hat{p})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

let x_1 and x_2 be the number of persons processing a given attribute in a random sample of size n_1 and n_2 then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}$ and $S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$

For this test

The null hypothesis is $H_0 : p_1 = p_2$

against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When p is not known p can be calculated by $p = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H$_0$

If $|Z| < Z_\alpha$, accept the null hypothesis H$_0$

- **As σ is unknown,**

$$\overline{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = \left[\overline{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \overline{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

KAHE

Step 2: If μ_0 falls into the above confidence intervals, then

do *not* reject H_0. Otherwise, reject H_0.

Example 1:

The average starting salary of a college graduate is $19000 according to government's report. The average salary of a random sample of 100 graduates is $18800. The standard error is 800. Is the government's report reliable as the level of significance is 0.05. Find the p-value and test the hypothesis in

(a) with the level of significance $\alpha = 0.01$. The other report by some institute indicates that the average salary is $18900. Construct a 95% confidence interval and test if this report is reliable.

[solutions:]

(a)

$$H_0 : \mu = \mu_0 = 19000 \quad \text{vs.} \quad H_a : \mu \neq \mu_0 = 19000\,,$$
$$n = 100,\ \bar{x} = 18800\,,\ s = 800,\ \alpha = 0.05$$

Then,

$$|z| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| = \left| \frac{18800 - 19000}{800/\sqrt{100}} \right| = |-2.5| = 2.5 > z_{\alpha/2} = z_{0.025} = 1.96\,.$$

Therefore, reject H_0.

(b)

$$\text{p-value} = P(|Z| > |z|) = P(|Z| > 2.5) = 2 \cdot P(Z > 2.5) = 0.0124 > 0.01$$

Therefore, ***not*** reject H_0.

(c)

$$H_0 : \mu = \mu_0 = 18900 \;\text{ vs }\; H_a : \mu \neq \mu_0 = 18900,$$

A 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 18800 \pm z_{0.025}\frac{800}{\sqrt{100}} = 18800 \pm 1.96 \cdot 80 = [18643.2, 18956.8].$$

Since $\mu_0 = 18900 \in [18643.2, 18956.8]$, Therefore, ***not*** reject H_0.

Example 2:

A sample of 49 provides a sample mean of 38 and a sample standard deviation of 7. Let $\alpha = 0.05$. Please test the hypothesis

$$H_0 : u = 40 \;\; vs. \;\; H_a : u \neq 40.$$

based on

(a) classical hypothesis test
(b) p-value
(c) confidence interval.
[solution:]

$$\bar{x} = 38,\; s = 7,\; u_0 = 40,\; n = 49,\; z = \frac{\bar{x} - u_0}{s/\sqrt{n}} = \frac{38 - 40}{7/\sqrt{49}} = -2.$$

(a)

$$|z| = 2 > 1.96 = z_{0.025}$$

we reject H_0.

KAHE

(b)

$$p-value = P(|Z| > |z|) = P(|Z| > 2) = 2*(1-0.9772) = 0.0456 < 0.05 = \alpha$$

we reject H_0.

(c)

$100\times(1-\alpha)\% = 95\%$ confidence interval is

$$\bar{x} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} = 38 \pm z_{0.025}\frac{7}{\sqrt{49}} = 38 \pm 1.96 = [36.04, 39.96].$$

Since $40 \notin [36.04, 39.96]$, we reject H_0.

KAHE

Hypothesis Testing for the Mean (Small Samples)

For samples of size less than 30 and when σ is unknown, if the population has a normal, or nearly normal, distribution, the *t*-distribution is used to test for the mean μ.

Using the t-Test for a Mean μ when the sample is small

Procedure	Equations	Example 4
State the claim mathematically and verbally. Identify the null and alternative hypotheses	State H_0 and H_a	$H_0 : \mu \geq 16500$ $H_a : \mu < 16500$ $n = 14, \bar{x} = 15700, s = 1250$
Specify the level of significance	Specify α	$\alpha = 0.05$
Identify the degrees of freedom and sketch the sampling distribution	$d.f = n - 1$	$d.f. = 13$
Determine any critical values. If test is left tailed, use One tail, α column with a negative sign. If test is right tailed, use One tail, α column with a positive sign. If test is two tailed, use Two tails, α column with a negative and positive sign.	Table 5 (*t*-distribution) in appendix B	The test is left-tailed. Since test is left tailed and $d.f = 13$, the critical value is $t_0 = -1.771$
Determine the rejection regions.	The rejection region is $t < t_0$	The rejection region is $t < -1.771$
Find the standardized test statistic	$t = \dfrac{\bar{x} - \mu}{\sigma_{\bar{x}}} \approx \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$	$t = \dfrac{15700 - 16500}{1250/\sqrt{14}} \approx -2.39$
Make a decision to reject or fail to reject the null hypothesis	If t is in the rejection region, reject H_0, Otherwise do not reject H_0	Since $-2.39 < -1.771$, reject H_0
Interpret the decision in the context of the original claim.		Reject claim that mean is at least 16500.

KAHE

Chi-square Tests and then F -Distribution

Goodness of Fit

DEFINITION :A **chi-square goodness-of-fit test is** used to test whether a frequency distribution fits an expected distribution.

The test is used in a multinomial experiment to determine whether the number of results in each category fits the null hypothesis:

H_0 : The distribution fits the proposed proportions

H_1 : The distribution differs from the claimed distribution.

To calculate the test statistic for the chi-square goodness-of-fit test, you can use observed frequencies and expected frequencies.

DEFINITION The **observed frequency O** of a category is the frequency for the category observed in the sample data.

The **expected frequency E** of a category is the calculated frequency for the category. Expected frequencies are obtained assuming the specified (or hypothesized) distribution. The expected frequency for the *i*th category is

$E_i = np_i$

where *n* is the number of trials (the sample size) and p_i is the assumed probability of the *i*th category.

The Chi-square Goodness of Fit Test: The sampling distribution for the goodness-of-fit test is a chi-square distribution with $k-1$ degrees of freedom where k is the number of categories. The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequency of each category and *E* represents the expected frequency of each category. To use the chi-square goodness of fit test, *the following must be true* .

1. The observed frequencies must be obtained using a random sample.
2. The expected frequencies must be ≥ 5 .

Performing the Chi-Square Goodness-of-Fit Test (p 496)		
Procedure	Equations	Example **(p 497)**
Identify the claim. State the null and alternative hypothesis.	State H_0 and H_1	H_0 : Classical 4% Country 36% Gospel 11% Oldies 2% Pop 18% Rock 29%
Specify the significance level	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	d.f. = #categories - 1	$d.f. = 6 - 1 = 5$
Find the critical value	χ^2_α : Obtain from Table 6 Appendix B	$\varphi^2_{0.01}(d.f = 5) = 15.086$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 15.086$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Survey results, n = 500 Classical O= 8 E = .04*500 = 20 Country O = 210 E = .36*500 = 180 Gospel O = 7 E = .11*500 = 55 Oldies O = 10 E = .02*500 = 10 Pop O = 75 E = .18*500 = 90 Rock O= 125 E = .29*500 = 145 Substituting $\chi^2 = 22.713$
Make the decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 22.713 > 15.086 we reject the null hypothesis Equivalently $P(X \geq 22.713) < 0.01$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		Music preferences differ from the radio station's claim.

Using Minitab to perform the Chi-Square Goodness-of-Fit Test (Manual p 237)

The data from the example above (Example 2 p 497) will be used.

Enter Three columns: Music Type: Classical, etc, Observed: 8 etc, Distribution 0.04, etc. (Note the names of the columns ‘Music Types’, ‘Observed’ and ‘Distribution’ are entered in the gray row at the top.)

Select **Calc->Calculator**, **Store the results in** C4, and calculate the **Expression** C3*500, click **OK**, Name C4 ‘Expected’ since it now contains the expected frequencies

Music Type	Observed	Distribution	Expected
Classical	8	0.04	20
Country	210	0.36	180
Gospel	72	0.11	55
Oldies	10	0.02	10
Pop	75	0.18	90
Rock	125	0.29	145

Next calculate the chi-square statistic, $(O-E)^2/E$ as follows: Click **Calc->Calculator**. **Store the results in** C5 and calculate the **Expression** (C2-C4)**2/C4. Click on **OK** and C5 should contain the calculated values.

7.2000
5.0000
41.8909
0.0000
2.5000
2.7586

KAHE

Next add up the values in C5 and the sum is the test statistic as follows: Click on **Calc->Column Statistics**. Select **Sum** and enter C5 for the **Input Variable**. Click OK. The chi-square statistic is displayed in the session window as follows:

```
Sum of C5
Sum of C5 = 22.7132
```

Next calculate the P-value: Click on **Calc->Probability Distributions->Chi-square**. Select **Cumulative Probability** and enter 5 **Degrees of Freedom** Enter the value of the test statistic 22.7132 for the **Input Constant**. Click **OK**.

The following is displayed on the Session Window.

```
Cumulative Distribution Function
Chi-Square with 5 DF
   x  P( X <= x )
22.7132     0.999617
```

$P(X \leq 22.7132) = 0.999617$ So the P-value = 1 – 0.999617 = 0.000383. This is less that $\alpha = 0.01$ so we reject the null hypothesis.

Instead of calculating the P-value, we could have found the critical value from the Chi-Square table (Table 6 Appendix B) for 5 degrees of freedom as we did above. The value is 15.086, and since our test statistic is 22.7132, we reject the null hypothesis.

KAHE

Chi-Square with M&M's

H_0: Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24%
Significance level: α = 0.05
Degrees of freedom: number of categories – 1 = 5
Critical Value: $\chi^2{}_{0.05}(d.f.=5)=11.071$
Rejection Region: $\chi^2 \geq 11.071$
Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where *O* is the actual number of M&M's of each color in the bag and *E* is the proportions specified under H_0 times the total number.
Reject *H_0* if the test statistic is greater than the critical value (1.145)

Section 10.2 Independence

This section describes the chi-square test for independence which tests whether two random variables are independent of each other.

DEFINTION An ***r x c*** **contingency table** shows the observed frequencies for the two variables. The observed frequencies are arranged in r rows and c columns. The intersection of a row and a column is called a **cell.**

The following is a contingency table for two variables A and B where f_{ij} is the frequency that A equals A_i and B equals B_j.

	A₁	**A₂**	**A₃**	**A₄**	**A**
B₁	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
B₂	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
B₃	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
B	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	f

KAHE

If A and B are independent, we'd expect

$$f_{ij} = prob(A = A_i) * prob(B = B_j) * f = \left(\frac{f_{i.}}{f}\right)\left(\frac{f_{.j}}{f}\right) f = \frac{(f_{i.})(f_{.j})}{f}$$

$$\frac{(sum\ of\ row\ i) * (sum\ of\ column\ j)}{sample\ size}$$ (

Example 1 Determining the expected frequencies of CEO's ages as a function of company size under the assumption that age is independent of company size.

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	$\frac{300*47}{550}$ ≈ 25.64	$\frac{300*87}{550}$ ≈ 47.45	$\frac{300*193}{550}$ ≈ 105.27	$\frac{300*180}{550}$ ≈ 98.18	$\frac{300*43}{550}$ ≈ 23.45	300
Large	$\frac{250*47}{550}$ ≈ 21.36	$\frac{250*87}{550}$ ≈ 39.55	$\frac{250*193}{550}$ ≈ 87.73	$\frac{250*180}{550}$ ≈ 81.82	$\frac{250*43}{550}$ ≈ 19.55	250
Total	47	87	193	180	43	550

After finding the expected frequencies under the assumption that the variables are independent, you can test whether they are independent using the chi-square independence test.

KAHE

DEFINITION A **chi-square independence test** is used to test the independence of two random variables. Using a chi-square test, you can determine whether the occurrence of one variable affects the probability of occurrence of the other variable.

To use the test,

1. The observed frequencies must be obtained from a random sample
2. Each expected frequency must be ≥ 5

The sampling distribution for the test is a chi-square distribution with

$(r-1)(c-1)$

degrees of freedom, where r and c are the number of rows and columns, respectively, of the contingency table. The test statistic for the chi-square independence test is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequencies and *E* represents the expected frequencies.

KAHE

To begin the test we state the null hypothesis that the variables are independent and the alternative hypothesis that they are dependent.

Performing a Chi-Square Test for Independence (p 507)		
Procedure	Equations	Example2 **(p 507)**
Identify the claim. State the null and alternative hypotheses.	State H_0 and H_1	H_0: CEO's ages are independent of company size H_1: CEO's ages are dependent on company size.
Specify the level of significance	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	$d.f. = (r-1)(c-1)$	$d.f. = (2-1)(5-1) = 4$
Find the critical value.	χ^2_α :Obtain from Table 6, Appendix B	$\chi^2_\alpha \geq 13.277$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 13.277$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$\sum \frac{(O-E)^2}{E} \approx 77.9$ Note that O is in the table of actual CEO's ages above, and E is in the table of Expected CEO's ages (if independent of size) above
Make a decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region. Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 77.9 > 13.277 we reject the null hypothesis Equivalently $P(X \geq 77.0) < \alpha$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		CEO's ages and company size are dependent.

KAHE

The test statistic (77.887) is greater than the critical value obtained from Table 6, Appendix B (13.277) so the null hypothesis is rejected. (Alternatively the P-Value (0.000) is less than the level of significance, α (0.01) so the null hypothesis is rejected.)

3. An urban geographer randomly samples 20 new residents of a neighborhood to determine their ratings of local bus service. The scale used is as follows: 0–very dissatisfied, 1–dissatisfied, 2– neutral, 3–satisfied, 4–very satisfied. The 20 responses are 0,4,3, 2,2,1,1,2,1,0,01,2,1,3,4,2,0,4,1. Use the sign test to see whether the population median is 2.

   Solution:

   There are 5 observations above the hypothesized median. Because the sample size is larger than 10, we test using the sample proportion *p* = 5/20 = 0.25. Using the PROB-VALUE method the steps in this test are:

   1) H_0: □ = 0.5 and H_A: □ ¹ 0.5
   2) We will use the *Z*-distribution
   3) We will use the 5%-level, thus □ = 0.05
   4) The test statistic is $z = (0.25 - 0.5)/\sqrt{0.25/20} = -2.24$
   5) Table A-4 shows that P(|*Z*| > 2.24) » 0.025.
   6) Because PROB-VALUE < □, we reject H_0. We conclude □ is different than 0.5, and thus the median is different than 2.

4. A course in statistical methods was team-taught by two instructors, Professor Jovita Fontanez and Professor Clarence Old. Professor Fontanzez used many active learning techniques, whereas Old employed traditional methods. As part of the course evaluation, students were asked to indicate their instructor preference. There was reason to think students would prefer Fontanez, and the sample obtained was consistent with that idea: of the 12 students surveyed, 8 preferred Professor Fontanez and 2 preferred Professor Old. The remaining students were unable to express a preference. Test the hypothesis that the students prefer Fontanez. (*Hint:* Use the sign test.)

   Solution:

KAHE

Although the sample is large enough for a normal approximation, we will use the binomial distribution to illustrate its application. Of the 12 observations, 8 preferred Prof. Fontanez, thus we need the probability of observing 8 or more successes in 12 trials of a Bernoulli process with the probability of success equal to 0.5. From Table A-1, we get

$$P(X^{3}\ 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$

$$P(X^{3}\ 8) = 0.1208 \quad + 0.0537 \quad + 0.0161 \quad + 0.0029 \quad + 0.002 \quad = 0.1937$$

Adopting the 5% uncertainty level, we see that PROB-VALUE > □□□ Thus we fail to reject H_0. We cannot conclude students prefer Fontanez.

5. Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs.

Solution:

(a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

(b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference.

We conclude the samples were drawn from different populations.

6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

High Density	Low Density	Sparsely Settled
1.84	2.04	1.07
3.06	2.28	2.31
3.62	4.01	0.91
4.91	1.86	3.28
3.49	1.42	1.31

Solution:

We will use the multi-sample Kruskal-Wallis test with an uncertainly level □ = 0.1. The null hypothesis is that all samples have been drawn from the same population. The rank sums are 55, 39 and 26 for the high density, low density, and sparsely settled samples respectively. The Kruskal-Wallis statistic is

$$H = \frac{12}{15(15+1)}\left(\frac{55^2}{5}+\frac{39^2}{5}+\frac{26^2}{5}\right) - 3(15+1) = 4.22$$

Using the □2 distribution with $3 - 1 = 2$ degrees of freedom, the associated PROB-VALUE is 0.121. We fail to reject the null hypothesis. The sample does not support the hypothesis of differing waste generation rates.

7. The distances travelled to work by a random sample of 12 people to their places of work in 1996 and again in 2006 are shown in the following table.

	Distance (km)			Distance (km)	
Person	1996	2006	Person	1996	2006
1	8.6	8.8	7	7.7	6.5
2	7.7	7.1	8	9.1	9
3	7.7	7.6	9	8	7.1
4	6.8	6.4	10	8.1	8.8
5	9.6	9.1	11	8.7	7.2
6	7.2	7.2	12	7.3	6.4

KAHE

Has the length of the journey to work changed over the decade?

Solution:

The sample can be considered as twelve paired observations. By taking differences between paired values, we get measures of the change for each individual. If the median change for the population is zero, we expect a sample to have a median difference near zero. Thus we will do a sign test for the median difference with a hypothesized value of zero. In other words, the hypotheses are $\mathrm{H}_0 : \eta = 0$ and $\mathrm{H}_A : \eta \neq 0$ □□We denote samples values whose distance decreased with a minus sign. Sample values with a positive difference get a plus sign. The sample becomes

$$S = \{-,+,+,+,+,0,+,-,+,-,+,+\}$$

Ignoring the tie, this is a sample of size 11 with 8 values above the hypothesized median. We are using Format (C) of Table 10-1, thus the PROB-VALUE is $2P(X \geq 8)$ where X is a binomial variable with □ = 0.5. From the equation for the binomial, the PROB-VALUE is found to be 0.113. At the □ = 10% level, we fail the reject the null hypothesis. We cannot conclude there has been a change in distance.

8. One hundred randomly sampled residents of a city subject to periodic flooding are classified according to whether they are on the floodplain of the major river bisecting the city or off the floodplain. These households are then surveyed to determine whether they currently have flood insurance of any kind. The survey results are as follows:

	On the Floodplain	Off the Floodplain
Insured	50	10
No Insurance	15	25

Test a relevant hypothesis.

Solution:

We will do a □^{2} test for a relationship between insurance and house location. The null hypothesis is no relationship (independence). Augmenting the data with expected frequencies, we have:

	On the Floodplain	Off the Floodplain
Insured	50 (39)	10 (21)
No Insurance	15 (26)	25 (14)

The corresponding □^{2} value is 22.16. Table A-8 shows that with 1 degree of freedom, P(□^{2} > 20) is zero to 3 decimal places. Thus for any reasonable level of uncertainty (any □ < 0.0005), we can reject the null hypothesis.

The occurrence of sunshine over a 30-day period was calculated as the percentage of time the sun was visible (i.e., not obscured by clouds). The daily percentages were:

	Percentage		Percentage		Percentage
Day	of sunshine	Day	of sunshine	Day	of sunshine
1	75	11	21	21	77
2	95	12	96	22	100
3	89	13	90	23	90
4	80	14	10	24	98
5	7	15	100	25	60
6	84	16	90	26	90
7	90	17	6	27	100
8	18	18	0	28	90
9	90	19	22	29	58
10	100	20	44	30	0

KAHE

If we define a sunny day as one with over 50% sunshine, determine whether the pattern of occurrence of sunny days is random.

Solution:

For this we can use the number-of-runs test. Rather than calculate runs across two samples, here we will simply note if a day has 50% or more sunshine. The sample becomes

S={+,+,+,+,-,+,+,-,+,+,-,+,+,-,+,+,-,-,-,-,+,+,+,+,+,+,+,+,-}

We see that the sample consists of 12 runs. There are $n_x = 21$ sunny days, and $n_y = 9$ cloudy days. Because $n_x < 20$, we cannot use the normal approximation given in Table 10-5. Instead the probability is computed using combinatorial rules, and is approximately 0.4. This is far too large for rejection of the randomness hypothesis. We cannot conclude the pattern is non-random.

9. Test the normality of the DO data (a) using the Kolmogorov–Smirnov test with the ungrouped data of Table 2-4 and (b) using the χ^2 test with $k = 6$ classes of Table 2-6.

Solution:

(a) We will take the mean and standard deviation as known rather than estimated from the sample. Doing so results in calculated PROB-VALUES that are smaller than the true values (i.e., we are more likely to reject the null hypothesis). For the DO data the mean and standard deviation are 5.58 and 0.39 respectively. We sort the data, and then find the differences between the observed and expected cumulative distributions. The table below shows the results for a few of the 50 observations:

x_i	$S(x_i)$	$F(x_i)$	$\lvert S(x_i)-F(x_i)\rvert$
4.2	0.020	0.015	0.005
4.3	0.040	0.023	0.017
4.4	0.060	0.032	0.028
…	…	…	…
5.9	0.780	0.692	0.088
…	…	...	…

6.7	0.960	0.960	0.000
6.8	0.980	0.972	0.008
6.9	1.000	0.981	0.019

The maximum difference is 0.088. Table A-9 shows that with 50 degrees of freedom, the corresponding PROB-VALUE is about 0.6. We obviously cannot reject the hypothesis of normality.

(b) Here we will take the mean and standard deviation as unknown, to be estimated from the sample. In other words, we estimate two parameters from the sample. In building the □^{2} table, we combine the first two and the last two categories in Table 2-6 to ensure at least 2 expected frequencies per cell. This reduces the number of categories 4, as seen in the table below:

Group	Minimum	Maximum	O_j	E_j	$(O_j-E_j)^2/E_j$
1	4.000	4.990	9	3.3	10.13
2	5.000	5.490	10	17.0	2.89
3	5.500	5.990	20	21.7	0.14
4	6.000	6.990	11	7.0	2.24

The observed Chi-square value is 15.4. With $k - p - 1 = 4 - 2 - 1 = 1$ degrees of freedom, Table A-8 shows that the PROB-VALUE is less than 0.0005. We therefore reject the null hypothesis.

Note that with only 4 classes, we can obtain only a rough idea of the distribution of DO. The 4 classes given in Table 2-6 do not yield a distribution that is at all similar to the normal distribution. In practice one would need many classes (and observations) for the □^{2} test to be reliable.

Nonparametric Hypothesis Testing

EARNING OBJECTIVES

By the end of this chapter, you will be able to:

1. Identify and cite examples of situations in which nonparametric tests of hypothesis are appropriate.
2. Explain the logic of nonparametric hypothesis testing for ordinal variables as applied to the Mann-Whitney *U* and runs tests.
3. Perform Mann-Whitney *U* and runs tests using the five-step model as a guide, and correctly interpret the results.
4. Select an appropriate nonparametric test.
5. **FORMULA 1** $U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$
6.
7. where: n_1 = number of cases in sample 1
8. n_2 = number of cases in sample 2
9. $\sum R_1$ = the sum of the ranks for sample 1
10. For our sample problem above
11. $U = (12)(12) + \frac{12(13)}{2} - 173.5$
12. $U = 144 + \frac{156}{2} - 173.5$
13. $U = 48.5$
14.
15. Note that we could have computed the *U* by using data from sample 2. This alternative solution, which we will label *U*′ (*U* prime), would have resulted in a larger value for *U*. The smaller of the two values, *U* or *U*′, is always taken as the value of *U*. Once *U* has been calculated, *U*′ can be quickly determined by means of Formula 2:
16. The alternative or research hypothesis is usually a statement to the effect that the two populations are different. This form for H_1 would direct the use of a two-tailed test. It is perfectly possible to use one-tailed tests with Mann-Whitney *U* when a direction for the difference can be predicted, but, to conserve space and time, we will consider only the two-tailed case.

17. In step 3, we will take advantage of the fact that, when total sample size (the combined number of cases in the two samples) is greater than or equal to 20, the sampling distribution of *U* approximates normality. This will allow us to use the *Z*-score table (see Appendix A of the textbook) to find the critical region as marked by *Z* (critical).
18. To compute the Mann-Whitney *U* test statistic (step 4), the necessary formulas are
19. **FORMULA 3** $Z \text{ (obtained)} = \dfrac{U - \mu_U}{\sigma_U}$
20. **FORMULA 5** $\sigma_U = \sqrt{\dfrac{n_1 \, n_2 \, (n_1 + n_2 + 1)}{12}}$
21. We now have all the information we need to conduct a test of significance for *U*.

22. **Step 2. Stating the Null Hypothesis**
23. H_0: The populations from which the samples are drawn are identical on the variable of interest.
24. (H_1: The populations from which the samples are drawn are different on the variable of interest.)
25. **Step 3. Selecting the Sampling Distribution and Establishing the Critical Region**.
26. Sampling distribution = *Z* distribution
27. Alpha = 0.05
28. *Z* (critical) = ± 1.96
29. **Step 4. Calculating the Test Statistic**. With *U* equal to 48.5, □$_U$ equal to 72, and σ_U of 17.32,
30. *Z* (obtained) = $\dfrac{U - \mu_U}{\sigma_U}$
31. *Z* (obtained) = $\dfrac{48.5 - 72}{17.32}$
32. *Z* (obtained) = -1.36

KAHE

33. **Step 5. Making a Decision**. The test statistic, a *Z* (obtained) of -1.36, does not fall in the critical region as marked by the *Z* (critical) of ± 1.96. Therefore, we fail to reject the null of no difference. Male students are not significantly different from female students in terms of their level of satisfaction with the social life available on campus. Note that if we had used the *U*′ value of 95.5 instead of *U* in computing the test statistic, the value of *Z* (obtained) would have been +1.36, and our decision to fail to reject the null would have been exactly the same.
34. **Making Assumptions**.
35. Model: Independent random sampling
36. Level of measurement is ordinal
37. **Step 2. Stating the Null Hypothesis.**
38. H_0: The two populations are identical on level of pain.
39. (H_1: The two populations are different on level of pain.)

IV. The ANOVA Table: Sums of Squares and Degrees of Freedom

A. Introduction

At the heart of any analysis of variance is the ANOVA Table. The formulas for the sums of squares in ANOVA are simplified if the *k* samples are all of the same size n_S. In the interests of simplicity, therefore, the following discussion assumes that all *k* samples contain the same number of observations n_S.

B. Notation

- The index i represents the ith population or treatment, where i ranges from 1 to *k*
- The index j represents the jth obsevation within a sample, where j ranges from 1 to n_S
- *n* is the total number of observations from all samples
- y_{ij} is the value of the jth observation in the ith sample
- $\overline{y}_i$ is the mean of the ith sample
- $\overline{\overline{y}}$ (read “y double-bar”) is the mean of all *n* observations, $\overline{\overline{y}} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_S} y_{ij}$, or the mean of the sample means (hence the “double-bar” in the name), $\overline{\overline{y}} = \frac{\overline{y}_1 + \overline{y}_2 + \cdots + \overline{y}_k}{k}$

KAHE

C. Sums of Squares

$$\boldsymbol{SSR} = n_{\mathrm{S}} \sum_{i=1}^{k} \left(\overline{\mathrm{y}}_{\mathrm{i}} - \overline{\overline{\mathrm{y}}} \right)^2$$

<u>S</u>um of <u>S</u>quares for <u>T</u>reatments, is the “Between Group” variation, where the *k* “groups” or populations are represented by their sample means. If the sample means differ substantially then SST will be large.

$$\boldsymbol{SSE} = \sum_{\mathrm{i=1}}^{\mathrm{k}} \sum_{\mathrm{j=1}}^{n_S} \left(\mathrm{y}_{\mathrm{ij}} - \overline{\mathrm{y}}_{\mathrm{i}} \right)^2$$

<u>S</u>um of <u>S</u>quares for <u>E</u>rror, is the “Within Group” variation and represents the random or sample-to-sample variation

$$\boldsymbol{SST} = \sum_{\mathrm{i=1}}^{\mathrm{k}} \sum_{\mathrm{j=1}}^{n_S} \left(\mathrm{y}_{\mathrm{ij}} - \overline{\overline{\mathrm{y}}} \right)^2$$

<u>Total</u> <u>S</u>um of <u>S</u>quares, is the total variation in the values of the response variables over all *k* samples. (Note: SST is the same as in regression)

The ANOVA Table below summarizes some of the information in this section

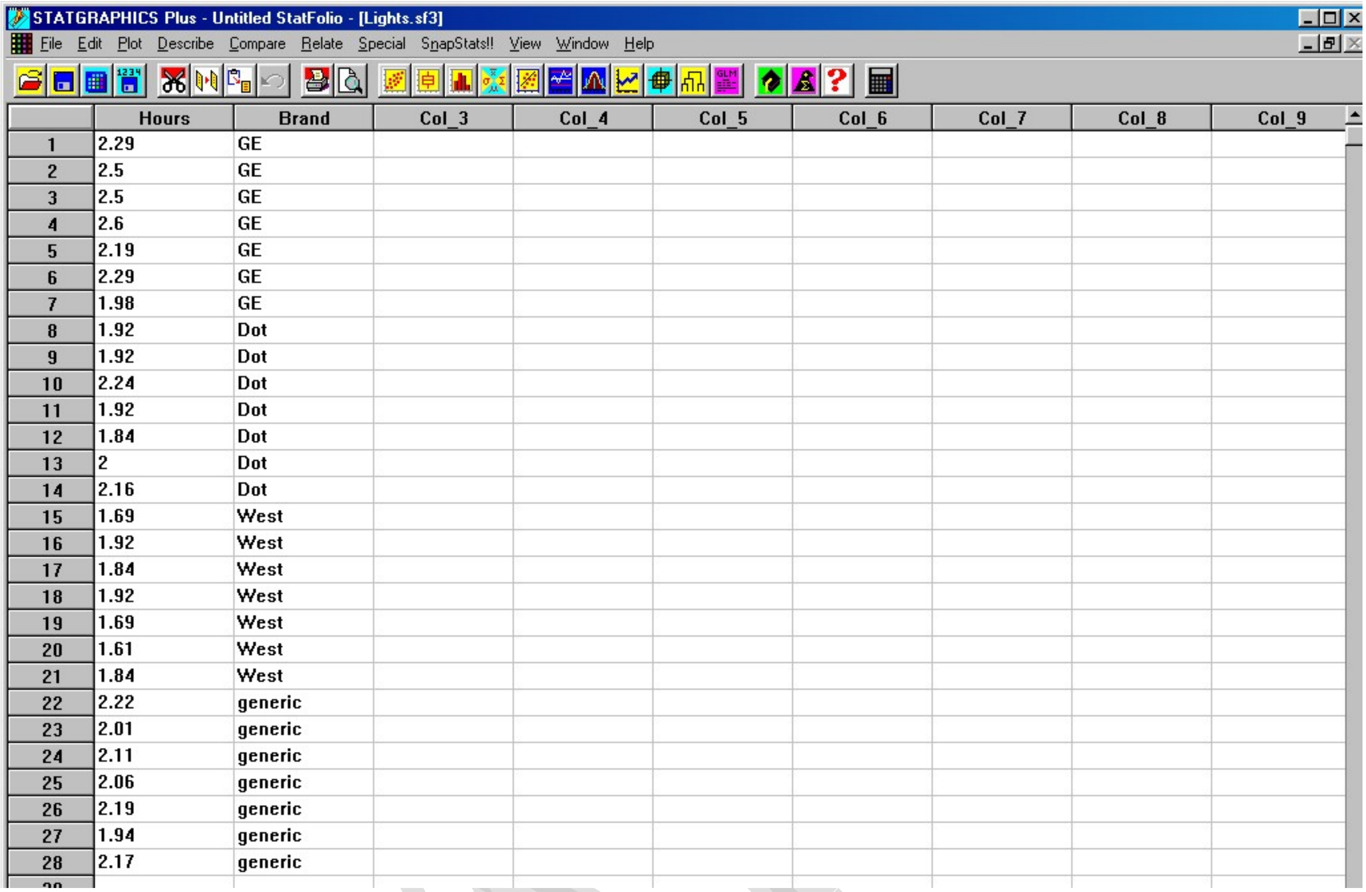
ANOVA Table for One-Way Analysis of Variance

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-Ratio*	*P-Value*
Between groups	***SSR***	*k* - 1	***MSR* = *SSR*/(*k*-1)**	***F* = *MSR/MSE***	
Within groups	***SSE***	*n* - *k*	***MSE* = *SSE*/(n-k)**		
Total (Corr.)	***SST***	*n* - 1			

V. Using Statgraphics

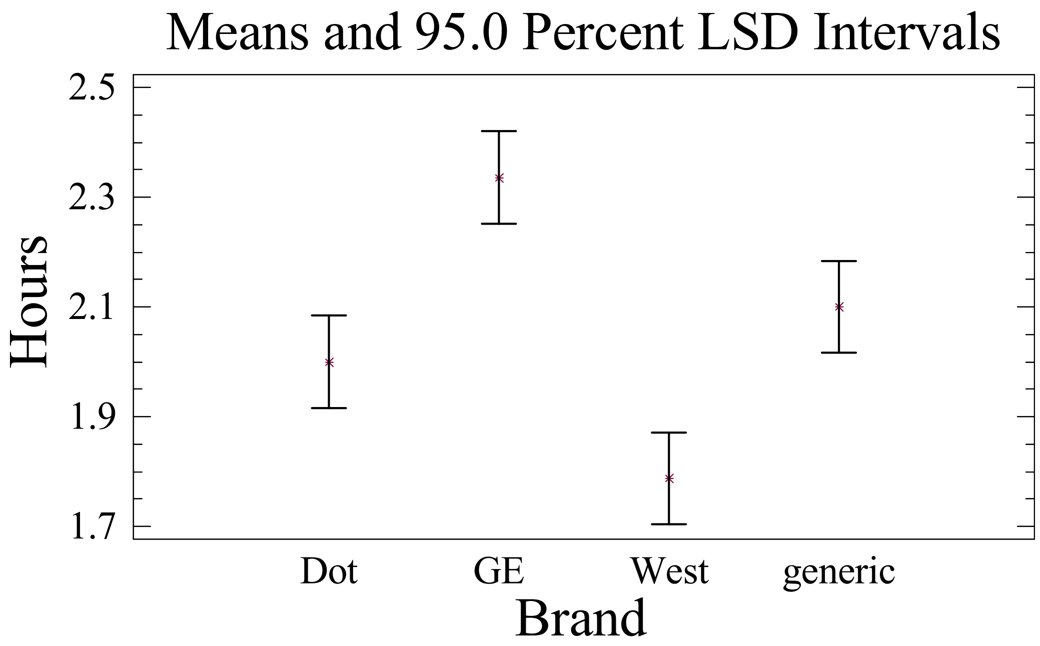
To perform a one-way analysis of variance in Statgraphics, follow *<u>C</u>ompare > <u>A</u>nalysis of Variance > <u>O</u>ne-Way ANOVA* and enter the response and factor into the dependent variable and factor fields, respectively.

Example 1 (continued): For the lightbulb problem, the spreadsheet might look like the one below. Notice that the qualitative factor Brand doesn’t need to be numeric. Statgraphics will treat the factor in ANOVA as qualitative, so there is no need to recode it as a numeric variable. For the same reason there is no need to create dummy variables as in regression.


STATGRAPHICS Plus - Untitled StatFolio - [Lights.sf3]
File Edit Plot Describe Compare Relate Special SnapStats!! View Window Help

	Hours	Brand	Col_3	Col_4	Col_5	Col_6	Col_7	Col_8	Col_9
1	2.29	GE							
2	2.5	GE							
3	2.5	GE							
4	2.6	GE							
5	2.19	GE							
6	2.29	GE							
7	1.98	GE							
8	1.92	Dot							
9	1.92	Dot							
10	2.24	Dot							
11	1.92	Dot							
12	1.84	Dot							
13	2	Dot							
14	2.16	Dot							
15	1.69	West							
16	1.92	West							
17	1.84	West							
18	1.92	West							
19	1.69	West							
20	1.61	West							
21	1.84	West							
22	2.22	generic							
23	2.01	generic							
24	2.11	generic							
25	2.06	generic							
26	2.19	generic							
27	1.94	generic							
28	2.17	generic							

This leads to the ANOVA Table below. Looking at the *P*-value for the *F*-test, we conclude that there is strong evidence that at least two of the mean lifetimes differ.



VI. Two-Way ANOVA

When the effects of two qualitative factors upon a quantitative response variable are investigated, the procedure is called two-way ANOVA. Although a model exists for two-way analysis of variance, similar to the multiple regression model, it will not be covered in this class. Neither will we cover the details of the ANOVA Table. Nevertheless, there are some new considerations in two-way ANOVA stemming from the presence of the second factor in the model.

Example 2: The EPA (Environmental Protection Agency) tests public bodies of water for the presence of *coliform* bacteria. Aside from being potentially harmful to people in its own right, this bacteria tend to proliferate in polluted water, making the presence of *coliform* bacteria a surrogate for polution. Water samples are collected off public beaches, and the number of *coliform* bacterial per cc is determined. (See the file "**Bacteria**."

The EPA is interested in determining the factors that affect *coliform* bacterial formation in a particular county. The county has beaches adjacent to the ocean, a bay, and a sound. The EPA beleives that the amount of "flushing" a beach gets may affect the ability of polution to accumulate in the waters off the beach. The EPA also believes that the geographical location of the beach may be significant. (There could be several reasons for this: the climate may be different in different parts of the county, or the land-use may vary across the county, etc.)

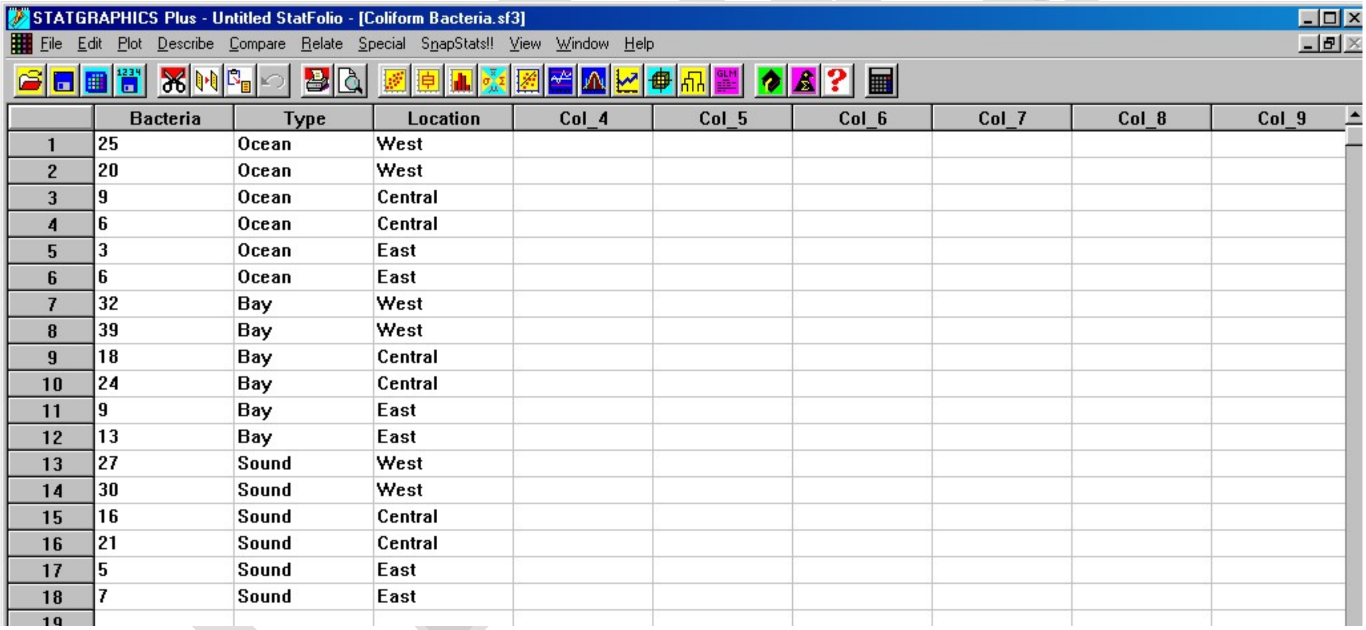
KAHE

As luck would have it, there is at least one beach for each combination of type (ocean, bay, sound) and location (west, central, east) within the county. Because of this, the EPA decides to sample a beach at each of the 9 possible combinations of type and location and conduct a two-way analysis of variance for *coliform* bacterial count. Two independent samples are taken at each beach to allow for an estimation of the natural variation in *coliform* bacterial count (this “repetition” is needed for the computation of MSE, which estimates the sample-to-sample variance in bacterial counts).

VII. Two-Way ANOVA Using Statgraphics

To perform a two-way analysis of variance in Statgraphics, follow *Compare > Analysis of Variance > Multifactor ANOVA* and enter the response and factors into the dependent variable and factor fields, respectively.

Example 2 (continued): Since data from such a study often appears in the form of a two-way table, with one factor as the row variable, the second as the column variable, and the observations as values in the row-by-column cells, it is important to remember that each variable must have its own column in the spreadsheet as in the example below. (This may require that you re-format the original spreadsheet prior to beginning the analysis.)



	Bacteria	Type	Location	Col_4	Col_5	Col_6	Col_7	Col_8	Col_9
1	25	Ocean	West						
2	20	Ocean	West						
3	9	Ocean	Central						
4	6	Ocean	Central						
5	3	Ocean	East						
6	6	Ocean	East						
7	32	Bay	West						
8	39	Bay	West						
9	18	Bay	Central						
10	24	Bay	Central						
11	9	Bay	East						
12	13	Bay	East						
13	27	Sound	West						
14	30	Sound	West						
15	16	Sound	Central						
16	21	Sound	Central						
17	5	Sound	East						
18	7	Sound	East						
19									

The default ANOVA Table below has separate rows for the factors Type (called factor A) and Location (called factor B). A test of the significance of each factor is performed and the corresponding p-value displayed. It appears that both the type of beach and its location affect *coliform* bacterial count.

But does the effect of the beach type on bacteria count depend upon its location within the county? If the particular pairings of factor levels are important, the factors are said to “interact.”

```
Source                 Sum of Squares     Df    Mean Square    F-Ratio    P-Value
--------------------------------------------------------------------------------
MAIN EFFECTS
 A:Type                       364.778      2        182.389      16.00     0.0003
 B:Location                   1430.11      2        715.056      62.71     0.0000

RESIDUAL                      148.222     13        11.4017
--------------------------------------------------------------------------------
TOTAL (CORRECTED)             1943.11     17
```

Before interpreting the results in the ANOVA table above, we should consider the role that interaction plays. If the effect of beach type on bacteria formation depends on the location of the beach then it is better to investigate the *combinations* of the levels of the factors type and location for their affect on bacteria. It will come as no surprise to you that there is a hypothesis test for interactions.

KAHE

POSSIBLE QUESTIONS

PART – B (SIX MARKS)

1.In 600 throws of a six faced dice, odd points appeared 360 times. would you say that the dice is fair at 5% level of significance?

2.A wholesaler in apples claims that only 4% of the apples supplied by him are defective. A random 0f 600 apples contained 36 defective apples. Test the claim of the wholesaler.

3.In a sample of 400 population from a village 230 are found to be eaters of vegetarian ietems and the rest non vegetarian items.can we assume that both vegetarian and non vegetarian food are equally popular?

4.A random sample of 500 pineapples were taken from a large consignment and 65 were found to be bed. Show that the standard error of the population of bad once in a sample of the size is 0.015,and deduce that the percentage of bad pineapples in the consignment almost certainly lies between 8.5 and 17.5

5.A machine puts out 16 imperfect articles of 500.After the machine is overhauled,it puts out 3 imperfect articles in a batch of 100.Has the machine improved?

6.In a simple random of 600 men taken from a big city,400 are found to be smokers. In another simple random sample 900 men taken from another city 450 are smokers. Do the data indicate that there is a significant difference in the habit of smoking in the two cities.

7.Sample of two different types of bulbs were tested for length of life,and the following data were obtained:

	Type I	Type II
Sample size	8	7
Sample mean	1234 hours	1136 hours
Sample of SD	36 hours	40 hours

Is the difference in the means significant?
(Given that the significant value of t at 5 % level of significance for 13 d.f. is 2.16)

8.A dice is tossed 120 times with the following results:

No.turned up	1	2	3	4	5	6	Total
Frequency	30	25	18	10	22	25	120

Test the hypothesis that the dice is unbiased.

9.The following results are obtained from a sample of 10 boxes of biscuits:
Mean weight of contents = 490gms
Standard deviation of the weight = 9 gms.
Could the sample come from a population having a mean of 500 gms.

10.4 coins were tossed 160 times and the following results were obtained :

No of heads	0	1	2	3	4
Observed frequencies	17	52	54	31	6

Under the assumption that coins are balanced,find the expected frequencies of getting 0,1,2,3,or 4 heads and test the goodness of fit.

PART – C (TEN MARKS)

11. Sample of two different types of bulbs were tested for length of life, and the following data were obtained:

	Type I	Type II
Sample size	9	7
Sample mean	1752 hours	1137 hours
Sample of SD	38 hours	42 hours

Is the difference in the means significant?
(Given that the significant value of t at 5 % level of significance for 13 d.f. is 2.16)

KAHE