9.	**Random Variables and Mathematical Expectation**	**159**
	Random variable	159
	Probability mass function	160
	Probability density function	160
	Mathematical Expectation	164
10.	**Joint Probability Distributions**	**171**
	Marginal Probability mass function	172
	marginal probability density function	173
	Correlation by using bivariate probability	197
11.	**Theoretical Probability Distributions**	**195**
	Binomial distribution	197
	Poisson distribution	204
	Negative Binomial distribution	211
	Hyper geometric distribution	216
12.	**Continuous Probability Distribution**	**225**
	Normal distribution	227
	Standard Normal distribution	229
	Gamma distribution	238
	Chi-square distribution	242
13.	**Theory of Estimation**	**249**
	Point estimation	256
	Interval Estimation	257
	Method of Estimation	261
14.	**Statistical Inference**	**269**
	Error in sampling	270
	Level of Significance	272
	One tailed and two tailed test	272
	Large sample test (Z-Test)	274
15.	**Exact Sampling Distribution**	**297**
	t-test	300
	f-test	318
16.	**Categorical Data Analysis (Chi-square test)**	**329**
	Goodness of fit	332
	Test of Independence	337
	Yate's Correction	340
17.	**Bibliography**	**351**
18	**Statistical Table**	**353**
	(Normal table, T-table , chi-square table and f-table)	353

CHAPTER 1

Statistics

ORIGIN AND DEVELOPMENT OF STATISTICS

The word "Statistics is very popular and widely used word. It has been derived from the Latin word "Status", German word "Statistik" and the Italian word "statista" all of which means the political state. In the ancient time Statistics was regarded as the 'Science of statecraft' and was the by product of the administrative activity of the state and the scope of the statistics was primarily limited to the collection of the data regarding 'age and sex wise' and 'property and wealth' of the country by the government for framing military and fiscal policies. But at present statistics is used by Economists, Scientists, Engineers, Doctors, and Researchers etc. So these days there is hardly a place where statistics is not used.

A comprehensive analysis and interpretation of data goes back to 1662 AD, when John Graunt published a book based on the birth and death records of England. Inspired by his work, Sir William Petty developed the idea of Political Arithmetic, the art of reasoning by figures upon thing relating to government. In around 1700 AD, the work of Edmund Halley on mortality tables and life expectancy (very useful in life insurance companies) had established statistics as a science both viable and worthwhile. All these works and efforts gave birth to the idea of statistics.

Over the next 150 years, many mathematicians, including the famous German mathematician Carl Friedrich Gauss, philosophers and scholars of different European countries attempted to develop the theory of statistics. By the end of the nineteenth century the discipline of statistics was well recognized; and in the first half of the twentieth century the field of statistics made remarkable progress both in theory and practice under the guidance of famous statisticians, such as Karl Pearson, Sir R. A. Fisher, William Sealey Gosset, Andrey Kolmogorov, Jerzy Neyman, Abraham Wald, John Tukey and so on.

The continuous efforts made by statisticians, mathematicians and research scholars over the past have succeeded to develop a wide variety of scientifically proven statistical methods that can be used in various disciplines for various data related activities. Statistical methods coupled with micro-computer and appropriate software packages are increasingly becoming powerful fact finding tools in a variety of disciplines and in a variety of situations.

1.2 DIVISION OF STATISTICS

The word Statistics conveys two different meanings when it is used in singular and plural senses. In plural sense, it refers to the numerical results obtained by applying statistical methods to a set of data systematically collected for specific pre-determined purpose. In singular sense, it refers to a subject of study that deals with various scientific methods, which are essential from the initial stage of data collection to the final stage of data presentation. The numerical data may relate to quantity of cement used, strength of steel used, assemble time of computer, wage of workers in an industry, demand and supply of cement etc.

The subject statistics is divided into two parts: mathematical statistics and applied statistics. Mathematical statistics deals with the development of statistical theory and methods based on certain principles and mathematics, while the applied statistics deals with the applications of statistical methods to data. In this context, Business Statistics is considered as applied statistics. The statistics is also divided into descriptive statistics and inferential statistics.

Descriptive statistics is used to summarize or present the data, either numerically or graphically. Numerical descriptors of data include the followings.

i. One-way or two-way frequency tables

ii. Various kinds of summary measures, such as mean, variance, correlation coefficient and so on

iii. Statistical models

While graphical summarizations include various kinds of charts and graphs, such as pie chart, line graph, scatter plot and so on. The main objective of this book is to teach descriptive statistics.

Inferential statistics is used to draw inferences about the population from a sample data drawn from the population. These inferences may take the form of answers to yes/no questions (hypothesis testing), estimates of numerical characteristics (estimation). Inferential statistical methods involve quite advanced methods and will not be considered in this book.

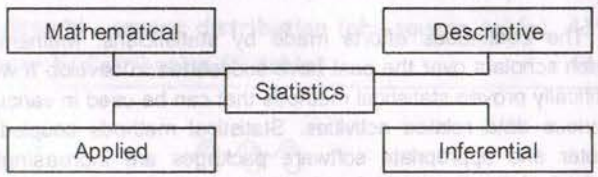


Figure of Division of Statistics

Learning of statistics will require the knowledge of basic mathematics. Like in other disciplines, statistics has its own vocabularies. An important and most frequently used term in statistics is variable. Whenever such terminologies are felt necessary to use we shall explain them.

If we observe or measure a characteristic we find that it takes on different values in different persons, places, or things, and it is customary to label the characteristics a variable. Some examples of variables include salary, age, sex, post and educational degree of employees of a company, amount of daily sales of a store, and the marks obtained by students in an exam. Data are considered as the values of a variable (in case of uni-variate data) or values of several variables (in case of multivariate data). The values of a variable are generated by measuring, asking or observing the characteristics of similar subjects under study.

Data generated by quantitative variables (such as salary, age, amount of sales and so on) are called quantitative data, while the data generated by qualitative variables (such as sex, post, educational level and so on) are called qualitative or categorical data.

1.3 DEFINITION OF STATISTICS

1.3.1 Singular Sense

In Singular Sense, there are many definitions given by number of Statistician that cover single aspect. A.L. Bowley has defined Statistics as the 'Science of Counting'. He has also defined Statistics as the 'Science of averages'.

Both definition of A.L. Bowley's are not completed by the comprehensive view of Statistics.

Boddington has defined Statistics as the 'Science of estimates and probabilities'. This definition is also not completed and satisfactory because estimates and probabilities are only the part of statistical methods. It does not cover the basic form of Statistics.

According to Ya–Lun–Chou, ' Statistics is a method of decision making in the face of uncertainty on the basis of numerical data and calculated risks.'

Croxton and Cowden give the most accepted and comprehensive definition of statistics in singular sense as "Statistics as the science which deals with the collection, presentation (Classification, tabulation, diagram and graphs), analysis and interpretation of the numerical data". According to this definition there are four stages.

i. **Collection of data**

The first step of an investigation is the collection of data. Careful collection is needed because further analysis is based on this. There are

different methods of collection of data and they must be reliable. If the collected data are faulty, results will also be faulty. Therefore, the investigator must take special care in collection of data.

ii. Presentation of data

The collected data are generally in an unintelligible form and need to be classified and tabulated before they can be analyzed. For example the researcher is interested to know the average expenditure of 900 families of a town. The mass data collected is difficult to understand and analyze. Therefore the collected data are to be presented in tabular or diagrammatic or graphic form. The data presented in systematic order will facilitate further analysis.

iii. Analysis of data

After the presentation of data, the next step is to analyze the presented data. Analysis includes condensation, summarization, conclusion etc. through the means of various descriptive statistical measures.

iv Interpretation of data

Figures don't speak for themselves. The duty of statistician is not complete with mere collection and analysis of data. But, valid conclusion must be drawn on the basis of analysis. A high degree of skill and experience is necessary for the interpretation. Correct interpretation leads to valid conclusion.

1.3.2 Plural Sense

Definition of Statistics In Plural sense covers many aspects. Prof. Horace Secrist gave the most comprehensive definition. According to him," By statistics, we mean aggregate of facts affected to a marked extent by multiplicity of causes, enumerated or estimated according to a reasonable standard of accuracy, collected in a systematic manner, for a predetermined purpose and placed in relation to each other."

This plural sense definition possesses the following characteristics of Statistics:

i. Aggregate of facts

Statistics deals with the aggregate of facts. It does not deal with the single figure. For instance, income of an employee of a company does not form statistics. But, the income of employees of the company constitutes statistics.

ii Affected by multiplicity of causes

Statistical figures should be affected by multiplicity of factors. Statistical techniques have been devised to study the joint effect of a number of factors on a single variable. For example, sale of a commodity depends upon the its demand, advertisement etc.

iii. Numerically expressed

Only numerical information constitutes statistics. Statistics deals with the quantitative data. The variables like beauty, honesty etc., unless expressed numerically, cannot be regarded as statistics.

iv. Enumerated according to reasonable standards of accuracy

Statistical data are collected either by enumeration or by estimation. The method of enumeration may be used when the population size is small. In such a case, data will be exact and accurate. If the population size is large, enumeration is not possible, so in such case the figures are estimated. However, the estimated values will not be as precise and accurate as the actually values collected by enumeration, the degree of accuracy largely depends on the nature and propose of the study. Any way, 100% accuracy in statistical work is rare. However, certain standards of accuracy must be maintained for drawing meaningful conclusions.

v. Collected in a systematic manner

The data must be collected after a well plan in a systematic manner. Otherwise, they will give misleading results.

vi. Collected for a pre-determined purpose

It is essential to state clearly the objective or the purpose of the enquiry before collecting the data. Collection of too much information and too little information should be avoided.

vii. Placed in relation to each other

The data should be comparable. The data can't be regarded as statistics if they are not comparable. In order to make valid comparisons the data should be homogenous i.e. they should relate to the same phenomenon or subject.

Thus, from the definition of Prof. Horace Secrist, we may conclude that

"All statistics are numerical statements of facts but all numerical statements of facts are not statistics."

1.4 USE OF STATISTICS

When we are dealing with uni-variate data scope of engineering statistics primarily concerned with statistical summaries such as average, variability shape and sizes of different engineering variables. And for bi-variate data there are two statistical techniques as correlation analysis and regression analysis. Correlation analysis summarizes the strength of relationship between two factors, while regression analysis shows how to predict the value of dependent variable using the independent one.

Furthermore, in engineering probability and probability distribution are used to calculate the numerical value of chances of happening and non happening of certain variables, expected value of random variables and overall fitting of probability distributions.

According to Wallis and Roberts, "Statistics may be regarded as a body of methods for making wise decisions in the face of uncertainty ". This definition reflects the application of statistics in engineering regarding formulation of policies, forecasting future trends, making statistical hypothesis and testing of these hypothesis and so on. Accurate estimation of population parameter of engineering variables and decision making of an engineer depends upon his experience and proper use of statistical tools and methods.

Engineers apply physical and chemical laws and mathematics to design, develop, test, and supervise various products and services.



Engineers perform tests to learn how things behave under stress, and at what point they might fail

As engineers perform experiments, they collect data that can be used to explain relationships better and to reveal information about the quality of products and services they provide and the information can be better analyzed

An engineer is someone who solves problems of interest to society by the efficient application of scientific principles. Engineers accomplish this by either refining an existing product or process or by designing a new product or process that meets customers' needs. The engineering, or scientific, method is the approach to formulating and solving these problems. The steps in the engineering method are as follows:

i. Develop a clear and concise description of the problem.
ii. Identify, at least tentatively, the important factors that affect this problem or that may play a role in its solution.
iii. Propose a model for the problem, using scientific or engineering knowledge of the phenomenon being studied, State any limitations or assumptions of the model.
iv. Conduct appropriate experiments and collect data to test or validate the tentative model or conclusions made in steps 2 and 3.
v. Refine model on the basis of the observed data.
vi. Manipulate the model to assist in developing a solution to the problem.
vii. Conduct an appropriate experiment to conform that the proposed solution to the problem is both effective and efficient.
viii. Draw conclusions or make recommendations based on the problem solution.

1.5 FUNCTION OF STATISTICS

The five major functions of statistics are illustrated in Fig 1.2; and they will be more elaborated in subsequent sections. Here we shall make brief elaboration in the context of function of statistics.

i. Statistics avoids collecting data by avoiding our biases

Data are the fundamental elements of statistics and they need to be collected. Collecting data sounds simple but may be the most difficult,

particularly when we want to avoid biases in the process of collecting data. There are numerous scientific sampling methods in statistics which avoids biases while collecting data. Findings based on data, which are free of biases, can be used to estimate parameters and such estimate get public credibility.

ii. Statistics provides techniques for organizing data scientifically

Collected data are in the raw form. They need to organize scientifically. Statistics provides techniques that help us to organize data scientifically. Nowadays it is customary to organize data in computer. Organization or management of data is very essentials in modern research, since well organized data eases the work of data analysis as well as helps for drawing desired information very quickly.

iii. Statistics enables to draw findings from masses of data

One of the main objectives of data analysis is to draw desired findings from masses of data. There are statistical methods or techniques that help us to draw desired findings from a large data set. The desired findings can be used in variety of situations, such as comparison of the performance between two or more business companies, formulating policies, and to take decisions.

iv. Statistics helps to take decision under uncertainty

One major topic of inferential statistics is hypothesis testing, which enables us to take decision under uncertainty with desired amount of risk.

v. Statistics enables to interpret data

The derived findings from a data set may need to interpret so that the people of various walks of life can understand the meaning of the findings. Statistics provides knowledge for interpreting the findings.

vi. Statistics enhances presentation of findings effectively

Presentation of numerical facts to public is as important as finding of facts from the data. Statistics provides various methods, such as tabular, graphical, summary measures, statistical models etc. for presenting the numerical findings.

1.6 LIMITATION OF STATISTICS

Like in many other disciplines, statistics has certain limitations. Some of them are listed below.

i. Statistics works on aggregate level only

Statistical findings are usually interpreted in terms of average, which may not be true to an individual. For example, statistical finding, such as, average monthly salary of employees of a company is Rs 40,000, but no employee in the company may be receiving this amount of salary.

ii. Statistics is relatively weak in the analysis of categorical data

Categorical data are frequently occurred in practice. However, there are some statistical methods that do not work for categorical data, but they work for quantitative data.

iii. Statistics is likely to be misused

Statistical methods are often intentionally misused, in the sense finding ways to interpret data that are favorable to the presenter. Inappropriate use of the available tools can produce misleading conclusions. In this context, it is worth mentioning the famous quote believed to be due to Benjamin Disraeli: There are three types of lies - lies, damn lies, and statistics.

iv. Statistics cannot prove anything

The logic employed in inferential statistics is inductive in nature (drawing inference from small part - sample - to a larger part - population), which is opposite to the deductive logic used in mathematics. The deductive logic or argument does not prove anything.

v. Statistical results are sometimes distrusting

Sometimes conflicting nature of statistical statements are available in the literature, particularly in the medical sciences. For example, statements like "doing X reduces high blood pressure" and also statements like "doing X actually worsens high blood pressure". This kind of conflicting statement is mainly when studies were conducted on different groups with different protocols. However, many readers may fail to notice these distinctions, or the media may oversimplify this vital contextual information, and the public's distrust of statistics is thereby increased.

1.7 DATA COLLECTION

We are at the age of information. Data is one of the main sources of information. Statistical methods are applied to data in order to derive information with the hope that the information will fulfill the desired objectives. For example, data is one of the main sources of information in order to know the potential buyers of a new product or to know the consumers' satisfaction out of a marketed product. Similarly, data is the main source of information to know whether or not to set up a new industry in a certain location. In such cases, marketing manager or statistician can use secondary data, primary data or both to fulfill the desired objectives. Both types of data may require and need to be collected from various sources or by various means.

It is important to note that the data is collected with the hope that information obtained from the data will able to answer our subject of enquiry; consequently, utmost care must be taken to collect as reliable and relevant data as possible. For this purpose, there are several crucial steps that need to be followed during the process of data collection in order to ensure that the data collection process and measurement systems are reliable. Incorporating these steps into a data collection plan will improve the likelihood that the data and measurements can be used to support the resulting analysis. The pre-data collection steps include the followings.

i. Clearly define the objectives of the data collection
ii. Reach understanding and agreement on operational definitions and methodology for the data collection plan
iii. Ensure data accuracy and measurement reliability.

The numerical statements of facts collected systematically for pre-determined purpose is called statistical data. The collection of data means the methods that are to be used for gathering the necessary information from the unit under consideration. The method of collection of data depends upon the nature, objective and the scope of the enquiry. The person who conducts the statistical enquiry is known as 'investigator'. The person from whom the information are collected is known as 'Respondent' or 'Informant'

The data can be collected in several ways. An agent or investigator himself may collect data or supervise its collection. In general, there are two techniques of data collection

i. Census technique and
ii. The Sampling technique

In the census method of data collection data is obtained from each and every unit of the population unit under consideration. In the sample technique, data are obtained only from a part of the population and the conclusion is drawn for the whole population. The second method, though easier, requires a lot of care and effort.

The persons or organizations that have collected the data and the reports and publication in which the data are published are known as the source of data. Data can be obtained either from "Primary sources" or from "secondary sources".

Statistical data depending upon the sources are of two types.

i. Primary data and
ii. Secondary data

1.8 PRIMARY AND SECONDARY DATA

The data required for a statistical enquiry are some times collected by the investigator or his agents. These data collected for the first time are original in nature and they are called primary data. These types of data are obtained in the survey and enquiries conducted by government, some individual, institution and research bodies. For example if an investigator wants to collect the data regarding the wage of workers of an institution he/she must go to that institution and taking interview to the workers and collect the necessary information.

Dates which are originally collected but obtained from some published or unpublished sources are "secondary data". This type of data is not original in character. For example the reports & publications trade by Nepal Engineering council are primary for that organization but secondary for those who use it.

The main difference between primary and secondary data is only of degree one. Data which are primary in the hands of one becomes secondary in the hands of other i.e. primary data once collected and published becomes secondary data for others investigators. For example the data relating to the income and expenditure of people of Nepal published by "Nepal Rastra Bank" are primary for that organization but secondary for those who use it.

1.9 METHODS OF COLLECTING PRIMARY DATA

Primary data is data that has to be collected by the researcher for the specific purpose of addressing the problem at hand. Collection of primary data is necessary when available secondary data cannot fulfill the research objective. For example, while lunching of a project an engineering researcher might be interested to know the following characteristics of the customers.

i. Environmental and socioeconomic effect
ii. Attitudes, opinions or interests
iii. Awareness and knowledge
iv. Intentions, motivation and behavior

The above nature of characteristics may change over time and location, in which case the available secondary data may fail to provide the current situations. The primary data is usually collected by conducting sample survey or experiment, and the data generated by these methods are correspondingly called survey data and experimental data. Statisticians have made both of these methods as scientific as possible; and both of these methods are now regarded as the fact-finding instruments.

The following methods are commonly used for the collection of primary data.

i. Direct personal interview
ii Indirect oral interview
iii. Information from local correspondents
iv. Mailed questionnaire method
v. Schedule sent through enumerators

1.9.1 Direct personal interview

In this method, the investigator himself goes to the spot and makes direct personal contacts with the people, called informants, from whom the data are to be collected. He asks them various questions related to the problem of study and collects first hand data from them. For example, if an investigator wants to make an enquiry on the wages of the workers of a factory, he will visit the factory, meet the workers personally and get the information from them.

This method is suitable when the scope of the enquiry is limited and confined to a small area, original or highly accurate data are required and the data are to be kept secret and confidential.

This method of collection of primary data is adopted in the following cases.

i. Where greater accuracy is needed.
ii. Where the field of enquiry is not large.
iii. Where confidential data are to be collected.
iv. Where the field is a complex one.
v. Where intensive study is needed.
vi. Where sufficient time is available.

Merits

i. More accurate and uniform information can be obtained as the interviewer can clear up the doubt of informants about certain questions.
ii. This method is most suitable for intensive study where the field of enquiry is small.
iii. Most of the informants response when the interviewer approaches them personally.
iv. Sensitive questions can be asked at such time when the informants feel more comfortable at home with the interviewer.
v. The interviewer can get proper information through respondents by taking them at their educational level and if need by asking questions in their language of communication.

Demerits

i. This method takes a lot of time and it is costly too.
ii. It is not suitable for extensive studies. If the area of enquiry is wide, it is not possible for a man to meet every informant and collect data.
iii. There may be a greater chance of personal prejudice and biasness.

1.9.2 Indirect Oral Interview

In this method, the investigator gets the information indirectly. So, in this method the interviewer collects the data from a third person who may be directly or indirectly in touch with the information sought. The third person is said to be the 'witness'. This method is used in case when the informants hesitate to give information directly. In this case, it is necessary to get information from the people who may know them.

This method is suitable when the field of enquiry is very wide and the information can't be collected by direct contact or the informants hesitate to give the information.

Merits

i. This method saves time, money and labour.
ii. It can cover a much wider area easily.
iii. In this method, the opinions and suggestions of experts can also be obtained.

Demerits

i. The information is obtained from the informants not directly connected, so it is doubtful that the results will be fully true.
ii. If the investigator is a biased person, he may twist the fact.

1.9.3 Information through Local Correspondents

In this method, the local agents called correspondents, employed in various places, collect data. They collect information in their own ways and send them to the central office where the data are processed and analyzed. This method is generally used by newspaper agencies to get information about important events regarding politics, sports, accidents, strikes, etc. Various government departments also use this method.

This method is suitable when information of different types is to be obtained on a regular basis at a low cost and when the area is vast.

Merits

i. Information can be obtained from a wide area by this method.

ii. This method is very cheap and yields results easily and promptly.

iii. Information can be obtained on a regular basis.

Demerits

i. Information sent by local correspondents may not be accurate and reliable due to personal prejudice and biasness.

ii. It is usually difficult to check or verify the accuracy of the collected data by this method.

1.9.4 Mailed Questionnaire Method

In this method, a questionnaire consisting of a list of questions pertaining to the enquiry is prepared. There are blank spaces for answers. This questionnaire is sent to the respondents, who are expected to write the answers in the blank spaces. A covering letter is also sent along with the questionnaire, requesting the respondents to extend their full cooperation by giving the correct replies and returning the questionnaire duly filled in time. To get a quick and better response, the return postage expense is borne by the investigator, by sending a self addressed and stamped envelope. Research workers, individuals, non–official agencies and the government usually adopt this method.

This method is suitable when the area of enquiry is very vast and the informants are literate and they know the importance of such enquiries. The success of this method depends upon the intelligence of the respondents concerned to understand the questions.

Merits

i. This method is most economical and it needs the least amount of money, time and manpower.

ii. Information in this method is given by informants themselves. Hence, information is free from biasness of the investigator.

iii. Information can be easily obtained however vast the area may be.

Demerits

i. This method can't be used where the informants are not literate.

ii. The greatest drawback of this method is considerable non-responses. Most of the informants don't return the questionnaire.

iii. The returned questionnaire may not be carefully filled and answers are haphazardly given. This may be due to that the informants might have misunderstood the question.

1.9.5 Schedule Sent Through Enumerators

It is the most widely used method of collection of primary data. A number of enumerators are selected and trained. They are provided with standardized questionnaires. Specific training and instructions are given to them for filling up schedules. Such enumerators will be in charge of a certain area of study. They go to the informants along with the schedule and get replies to the questions. Then they record their answers in the schedule. Enumerators explain clearly the object and purpose of the enquiry.

This method is suitable when the field of investigation is not very vast and the informants are even illiterate.

The difference between the questionnaire and this method is that in the former method, the questionnaire is sent to the informants whereas in this method, the enumerator carries the schedule to the informants. Public organizations, research institutions and others use this method.

Merits of Questionnaire method

i. Information can also be obtained from the person who can't read or write.

ii. In this method, the enumerators can check the accuracy of information by asking supplementary questions.

iii. There is very little possibility of non-responses as the enumerators go personally to obtain the data.

Demerits of Questionnaire method

i. Data collected by inefficient enumerators may lead to wrong conclusion.

ii. It is an expensive method since the enumerators have to be paid for their service.

iii. It is more time consuming in comparison to the 'Mailed questionnaire method'.

1.10 SOURCES OF SECONDARY DATA

Secondary data are borrowed from the collections which have already been collected by others. Use of primary data in statistical investigation is not always practicable due to the availability of time, money and so on .In this method investigator collects the secondary data. There is a lot of published and unpublished data from which the further studies can be made. The sources of secondary data may be classified as: - published sources and unpublished sources

The major sources of secondary data are published reports, website and CD-Rom launched by government or non-government organizations. Statistical Year Book published by the Central Bureau of Statistics (CBS), Economic Survey published by the Ministry of Finance, Quarterly Economic Bulletin published by Nepal Rastra Bank are some the examples of published secondary data sources. Unpublished official or survey data of various organizations also constitute the sources of secondary data. Apart from these international agencies, such as United Nations (UN), International Monetary Fund (IMF), World Bank (WB) and Asian Development Bank (ADB) etc., also regularly publish statistical reports.

Statistical data are not always published and they are kept in un-published form. There are various sources of un-published statistical materials such as the records maintained by private form, business enterprises, various departments and office of the government, hospital records, research carried out by the individual research scholars in the universities or research institutions.

1.10.1 Advantages of Secondary Data

The following are the advantages of secondary data.

i. It saves time and cost of survey.

ii. If specially trained persons collect it, the quality of secondary data is better.

iii. The scope of inquiry can be increased in terms of area and time period to be covered.

1.10.2 Disadvantages of Secondary Data

Following are the main disadvantages of secondary data:

i. Many times the exact definition of terms and units used in a secondary data are not known.

ii. When the secondary data are obtained from two sources, it may not be comparable in terms of definitions, units and time period covered.

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

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

1.10.2 Precautions Used in Secondary Data

Some others might have collected secondary data for their own purpose. They may not be reliable or suitable for the enquiry to be conducted. They can be used only when they are carefully scrutinized and found satisfactory after the following points are dually considered.

i. Data should be reliable.

ii. Data should be suitable.

iii. Data should be adequate

1.11 PROBLEM IN DATA COLLECTION

While collecting data, many problems arise. Some of them are as follows.

1.11.1 Problems in Primary Data Collection

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

i. The informants may show dishonesty in giving the answers of questions put by the interviewer.

ii. If the informants are illiterate and the informants do not understand the questions put by the interviewer, the real information can't be obtained.

iii. A problem may arise due to lack of transportation facility.

iv. While using 'Mailed questionnaire method' in collecting the data, only a few questionnaires are returned. Even if they are returned, they may not be sincerely answered. So, accurate data can't be expected.

v. If the scope of enquiry is wide and the high degree of accuracy is required, then money and time should sufficiently be available. If not, inadequate and inaccurate data may be obtained.

vi. The data should be collected by well-trained and skilled investigators/enumerators. But mostly, agents of investigators without any training conduct the investigations. So, the accuracy of data collected by them may be questioned. Unless the information collected are relevant, the conclusion obtained will not be fruitful.

i. The informants may show dishonesty in giving the answers of questions put by the interviewer.

ii. If the informants are illiterate and the informants do not understand the questions put by the interviewer, the real information can't be obtained.

iii. A problem may arise due to lack of transportation facility.

iv. While using 'Mailed questionnaire method' in collecting the data, only a few questionnaires are returned. Even if they are returned, they may not be sincerely answered. So, accurate data can't be expected.

v. If the scope of enquiry is wide and the high degree of accuracy is required, then money and time should sufficiently be available. If not, inadequate and inaccurate data may be obtained.

vi. The data should be collected by well-trained and skilled investigators/enumerators. But mostly, agents of investigators without any training conduct the investigations. So, the accuracy of data collected by them may be questioned. Unless the information collected are relevant, the conclusion obtained will not be fruitful.

Problems in Secondary Data Collection

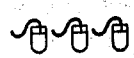
While collecting secondary data, we face different problems. Some of them are as follows.

i. Suitable, reliable and adequate data to the enquiry under study can be rarely obtained.

ii. Sometimes, the unpublished data can't easily be obtained as the persons hesitate to give it.

Exercise – 2

1. Distinguish between 'primary data' and 'secondary data; State the sources of secondary data. What precautions are to be observed when such data are to be used for any investigation?
2. What method would you employ in collection of data regarding statistical enquiry about "Graduate unemployment" in Kathmandu considering accuracy, time and cost of enquiry?
3. What is primary data? Enumerate the various methods of collecting primary data.
4. What is secondary data? Point out the sources of secondary data.
5. What are the problems in data collection ?



Classification and Tabulation

INTRODUCTION

Whether the raw data is primary or secondary it is necessary that they will need to organize or manage scientifically before they can be analyzed. A pile of completed questionnaires, or a long list of sales data, convey nothing to the statisticians or marketing officers. It is important to note that each questionnaire contains various types of data that may be broken down or classified under a number of variables; consequently, the answers of the questionnaires must be arranged of each variable and of each questionnaire scientifically in a spreadsheet like table or tables. Likewise, the sales data may have to be broken down by commodity, store, location and time, and need to be organized in tables accordingly. All these data related activities fall under the topic of data organization or management and organized data constitute data file or files. Nowadays, organization of data is usually carried out in computer.

Scientifically organized data can be analyzed. The preliminary stage of data analysis is the tabulation of data in terms of frequencies, and such table is called frequency table. Note that each variable in a data file can generate one frequency table. The process of generating frequency tables from a data file is dealt under the topic frequency distribution.

DATA ORGANIZATION

A spreadsheet like table contains a number of rows and a number of columns. The organization of data in a table must follow a rule; and the rule is that each column corresponds to a variable and each row corresponds to entity (unit about which data is collected). For example, the data on socio-demographic variables of 5 employees is organized as shown in the following table, where the variable MS stands for marital status and EA stands for educational attainment.

Table 3.1: Part of an organized data file

Employee	Sex	Age	MS	EA	Salary
1	M	32	Married	Master	18000
2	F	40	Married	Bachelor	15000
3	M	28	Married	Bachelor	16000
4	M	23	Unmarried	Master	14000
5	F	20	Unmarried	Inter	12000

In the above table, sex, marital status and educational attainment are categorical variables, while age and salary are quantitative variables.

Nowadays, it is customary to use coding system for categorical variables. For example, we may use the following coding system.

For sex: Male = 1 and Female = 2

For MS: Unmarried =1 and Married = 2

For EA: Inter = 1, Bachelor = 2 and Master = 3

When this coding system is followed, then all entries, or cell values, of the above table will convert to numbers. During data organization coding system is essential for expediting the data entry work as well as for facilitating the process of data analysis.

The statements - the second employee is twice as old as the fifth employee and the first employee's salary is one and half times higher than that of the fifth employee - are perfectly correct. However, the statements - female (=2) is twice as high as male (=1) or female is greater than male, since 2 is greater than one - are incorrect. The lesson that we learn from these examples is as follows.

One should not only take into consideration of the face values of a variable while comparing them, but also take into consideration of the scale of measurement associated with the variable.

2.3 CLASSIFICATION OF VARIABLES AND DATA

Data generated by four scales of measurement are correspondingly called nominal, ordinal, interval scale and ratio scale data. It is important to note that nominal and ordinal scale of measurement produce categorical data, while the interval and ratio scale produce quantitative data.

Data collected one short of time is called cross-sectional, while data collected or generated over a regular time interval is called time series. Usually, survey data are cross-sectional, while the hourly, daily, weekly, monthly, or annually recorded data are time series.

Quantitative variables are classified into two categories: discrete and continuous. A variable is discrete if there are natural gaps among all possible values of the variable. For example, the number of items sold in a day, the number of employees in a company, and the number of finance companies in a locality are some examples of discrete variables. A variable is continuous if there is no natural gap among all possible values of the variable. Age, height, and weight are some examples of continuous variables.

2.4 SIGMA NOTATION

Sigma notation is used for getting the aggregate of the same type of variable. The Greek letter Σ (capital sigma) is used to denote the sum of a series.

For example $\sum_{k=1}^{4} X_k = X_1 + X_2 + X_3 + X_4$

2.5 FREQUENCY

The number of times, which a variate value occurs, is known as frequency. It is denoted by f.

Total frequency is, $\Sigma f = N$

2.6 FREQUENCY DISTRIBUTION

Frequency distribution is simply a table in which the collected and classified data are presented into different classes. It shows the frequency of occurrence of different values of a single phenomenon or characteristic. The different values of a variable are presented in suitable way i.e. individual, ungrouped and grouped.

Following are the types of frequency distribution.

i. Raw data

ii. Discrete frequency distribution

iii. Continuous frequency distribution

Array

Arrangement of statistical raw data called individual data or series in an ascending or descending order of magnitude is called array of the data. However, this presentation though better than the raw data, does not reduce the volume of the data.

2.6.1 Discrete Frequency Distribution

The series formed from a discrete variable is known as discrete series. It is also called discrete frequency distribution or ungrouped frequency distribution.

For example,

Variable	0	1	2	3	4
Frequency	5	3	10	9	11

If the variables are not largely distributed, discrete frequency distribution is suitable.

2.6.2 Construction of a Discrete Frequency Distribution

A discrete frequency distribution changes the given raw data into condensed and simpler forms by using the following steps.

a. Write all the possible values of the variable from the lowest to the highest. (X- Variable column)

b. Put a vertical bar, known as tally bar against the value of a variable to which it occurs.

c. Count the number of bars with respect to each value of variable as frequency of variable.

Example 1

The marks obtained by 30 students of a class are given below.

4	5	6	7	8	9	4	5	6	5
3	4	5	7	8	2	3	5	6	7
8	7	4	5	3	2	4	2	3	9

Construct the discrete frequency distribution of the given raw data.

Solution

Construction of discrete frequency distribution

Marks	Tally Bar	Frequency
2	|||	3
3	||||	4
4	𝍸	5
5	𝍸 |	6
6	|||	3
7	||||	4
8	|||	3
9	||	2
Total (N = f)		30

2.6.3 Continuous Frequency Distribution

If the variate values lie between two numbers which is known as an interval. If the frequency is counted in the interval, then the distribution is called continuous frequency distribution.

For example,

Mark Obtained	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
No. of students	10	20	30	15	5

2.6.4 The Terms Used in Continuous Frequency Distribution

The following technical terms are important when a continuous frequency distribution is formed or data are classified according to class intervals.

a. Class b. Class limits c. Class size

d. Class frequency e. Mid value of a class

Class

To manage and analyze such that, the variate values of a variable are partitioned or classified into groups of equal length or interval termed as a class. The classes should be clearly defined and should not lead to any ambiguity. Each of the given value should be included in one of the classes.

Class limits

The class limits are the smallest and the largest values in the class. For example, let us take the class 20–30. The lowest value is 20 and the highest value is 30. The two boundaries of the class are known as lower limit and upper limit of the class. Class limit are also known as class boundaries.

Class size

The difference between the upper and the lower limit of a class is known as class size or class interval. For example, in the class 20–30, class size is 10. It is denoted by i or h.

Class frequency

The number of observations falling within a class interval is known as its class frequency. For example, the frequency of the class 20–30 is 13. This means that there are 13 numbers of observations lying between 20 and 30. If we add the frequency of all individual classes, we obtain the total frequency.

Mid value of class

The central value of a class interval is called mid-point or mid-value. It lies halfway between the lower limit and the upper limit of a class interval.

$$\text{Mid point of a class} = \frac{\text{U L of the class} - \text{LL of the class}}{2}$$

Where, UL= Upper Limit and LL = Lower Limit

2.6.5 Formation of Continuous Frequency Distribution

For the construction of continuous frequency distribution the following term must be identified.

Number of Classes

The number of classes largely depends upon the number of variable values in the data; the nature of data. It should preferably lie between 5 and 15, but may be more than 15 depending upon the total frequency. It should not be less than 5 because this classification may not reveal the required characteristics of the population.

The number of classes is determined by formula of Professor Sturges called Sturges' rule.

$$K = 1 + 3.322 \log_{10} N$$

N = Number of observations

According to rule if there are 200 numbers of observations,

$$K = 1 + 3.322 \log_{10} 200$$

$$= 1 + 3.322 \times 2.3010 = 8.64 \text{ (nearly 9 classes)}$$

If N = 40, i.e. the number of observation is 40

Then, $K = 1 + 3.322 \log_{10} 40 = 1 + 3.322 \times 1.6021$

$= 1 + 5.3381 = 6.322$

Thus, suitable numbers of classes are 6 or 7

Class Size

In general, class size should be taken as a multiple of five or ten. But from the Sturges' rule,

$h = \frac{L - S}{K}$, h = Class size , K = Number of classes

L = The Largest value, S = The smallest value.

2.7 Formation of Class Intervals

Continuous frequency distribution is formed by the following three methods.

i. Exclusive type classes ii. Inclusive type classes

iii. Open end classes

2.7.1 Exclusive Method

In this method, the upper limit of one class is equal to the lower limit of the next class to make it continuous without any gap. This is mostly useful in all statistical study of continuous frequency distribution. In this method, upper limit of the value of variable is excluded in that class while counting, which is counted in the next class.

Class interval	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50
Frequency	5	10	7	8	8

2.7.2 Inclusive Method

There is a gap between the upper limit of one class and lower limit of next class. But both lower limit and upper limit are included in the same class.

For example,

Class interval	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49
Frequency	5	10	7	8	8

For some statistical measures, we should change inclusive classes into exclusive by using the following correction factor (C_f)

$$C_f = \frac{\text{Lower limit of succeeding class} - \text{Upper limit of preceding class}}{2}$$

Then, exclusive classes will be

Lower limit of each class = Lower limit - C_f

Upper limit of each class= Upper limit + C_f

Example 2

Convert the following inclusive classes into exclusive.

Class	0 - 9	10 - 19	20 - 29	30 - 39	40 - 49
Frequency	5	10	7	8	6

Solution

Here, the correction factor is given by

$$C_f = \frac{\text{Lower limit of } 2^{nd} \text{ class} - \text{Upper limit of } 1^{st} \text{ class}}{2} = \frac{10 - 9}{2} = 0.5$$

So, Lower limit of first exclusive class = 0 – 0.5 = – 0.5

Upper limit of first exclusive class = 9 + 0.5 = 9.5

Hence, the first exclusive class = (- 0.5 to 9.5) and so on

Therefore, the exclusive classes are given below:

Class interval	Frequency
– 0.5 - 9.5	5
9.5 - 19.5	10
19.5 - 29.5	7
29.5 - 39.5	8
39.5 - 49.5	6

2.7.3 Open End Classes

In continuous frequency distribution, if the lower limit of the first class or the upper limits of the last class or both are not bound, they are called open end classes.

For example,

Class	Below 10	10 - 20	20 - 40	Above 40
Frequency	10	8	7	5

General Rules for Forming Continuous Frequency Distribution

While forming continuous frequency distribution generally the following rules are considered.

i. Try to make class interval equal.

ii. The number of class lies between 5 and 15.

iii. Class size is generally taken as 5 or multiple of 5.

iv. Lower limit are generally 5 or 0 or multiple of 5.

v. Classes must be mutually exclusive. Classes must be constructed so that an observation is placed in one and only one class.

vi. Exclusive method is more suitable than inclusive to ensure continuity and get the correct class size. If inclusive method has been adopted, it is necessary to make an adjustment for further statistical treatment.

2.8 CUMULATIVE FREQUENCY DISTRIBUTION

If we are interested to know the total number of observation getting a value "less than" or "more than" a particular value of variable, it must consider the cumulative frequency distribution. There are two types of cumulative frequency distribution.

2.8.1 Less than Cumulative Frequency Distribution

Less than cumulative frequency for any value of the variable (or class) is obtained on adding successively the frequencies of all the previous values (or classes) including the frequency of value (class) against which the totals are written, provided that the values (classes) are arranged in an ascending order of magnitude taken upper limit of each class.

Let us consider the distribution of marks of 60 students in an examination.

Marks	Number of students
20 - 30	2
30 - 40	8
40 -50	10
50 - 60	20
60 - 70	12
70 - 80	8

If we want to know how many students get less than 40 marks, less than 50 marks, less than 60 marks we construct the following 'Less than' cumulative frequency table.

Marks	Frequency	Less than c.f.
20 - 30	2	2
30 - 40	8	10
40 - 50	10	20
50 - 60	20	40
60 - 70	12	52
70 - 80	8	60

Then, 'Less than' cumulative frequency distribution is:

Marks	Frequency
Below 30	2
Below 40	10
Below 50	20
Below 60	40
Below 70	52
Below 80	60

2.8.2 More than Cumulative Frequency Distribution

Similarly, the values classes are arranged in a descending order of magnitude taken lower limit of each class. In this case frequencies are added from bottom to top.

Let us consider the distribution of marks of 60 students in an examination.

'More than' cumulative frequency table

Marks	Frequency	More than c.f.
20 - 30	2	60
30 - 40	8	58
40 - 50	10	50
50 - 60	20	40
60 - 70	10	20
70 - 80	8	8
Total	60	

So, more than cumulative frequency distribution is:

Marks	Frequency
Above 20	60
Above 30	58
Above 40	50
Above 50	40
Above 60	20
Above 70	8

Example 3

Convert the following distribution into 'more than' frequency distribution.

Weekly wages below (Rs)	20	40	60	80	100
Number of workers	41	92	156	194	201

Solution

The given distribution is less than cumulative frequency distribution. So, we first convert it into simple frequency distribution.

Wage (Rs)	Less than c.f.	Frequency (f)	More than c.f.
0-20	41	41	160 + 41 = 201
20-40	92	92 - 41 = 51	109 + 51 = 160
40-60	156	156 - 92 = 64	45 + 64 = 109
60-80	194	194 - 156 = 38	7 + 38 = 45
80-100	201	201 - 194 = 7	7
		N = $\sum$f = 201	

Thus, the more than cumulative frequency distribution is

Weekly wages (Rs)	Number of workers
Above 0	201
Above 20	160
Above 40	109
Above 60	45
Above 80	7
Up to 100	201

Example 4

In a survey it was found that 64 families bought milk in the following quantities (Liters) recorded as:

19	16	22	9	22	12	39	19	14	91	23
23	6	24	16	18	7	17	20	25	17	16
28	18	10	24	20	21	10	7	18	27	17
28	24	20	14	23	25	34	22	5	15	31
33	23	26	29	13	36	11	26	1	9	12
37	30	13	8	15	22	21	32	21		

Using Struges' rule converts the above data into a frequency distribution by inclusive method and converts it into exclusive frequency distribution.

Solution

Here, the total observations is 64

By sturges' rule,

$K = 1 + 3.322 \text{ Log}_{10} 64 = 1 + 3.322 \times 1.8062$

$= 1 + 6.000196 = 7.000196 \approx 7$

Now, Range = Maximum value – Minimum value = 39 - 5 = 34

Class size, $h = \frac{39-5}{K} = \frac{34}{7} = 4.857 \approx 5$

Now, frequency distribution by inclusive method is:

Milk (Liter)	Tally bar	No. of families
5 - 9	𝍸 II	7
10 - 14	𝍸 𝍸	10
15 - 19	𝍸 𝍸 III	13
20 - 24	𝍸 𝍸 𝍸 III	18
25 - 29	𝍸 III	8
30 - 34	𝍸	5
35 – 39	III	3

Correction factor (C_f) = $\frac{10-9}{2} = 0.5$

Lower limit of each class = lower limit - C_f

Upper limit of each class= upper limit + C_f

Thus, exclusive frequency distribution as given below:

Milk (Liter)	No. of families
4.5 - 9.5	7
9.5 - 14.5	10
14.5 - 19.5	13
19.5 - 24.5	18
24.5 - 29.5	8
29.5 - 34.5	5
34.5 - 39.5	3
Total	N = 64

Example 5

If the marks (mid values) in a frequency distribution of marks obtained by student are 40, 48, 56, 64, 72, 80, 88

Find **a.** Class interval size **b.** The class limits **c.** The class boundaries.

Solution

a. Class interval = 48 – 40 = 8

Half of class interval = 4

b. Lower limit of first class = 40 – 4 = 36

Lower limit of second class = 48 – 4 = 44, and so on

Upper limit of first class = 40 + 4 = 44

Upper limit of second class = 48 + 4 = 52 and so on

c. Class boundaries are,

36 – 44, 44 – 52, 52 – 60, 60 – 68, 68 – 76, 76 – 84 and 84 – 92

Example 6

The following are the weekly production in units of 60 workers in a manufacturing company.

23	48	51	64	82	19	33	50	39	72
35	88	77	25	39	52	48	64	49	57
41	72	62	49	32	54	67	46	55	52
82	44	75	56	51	63	59	69	53	57
75	85	68	55	52	45	40	57	20	42
46	51	50	16	62	56	54	40	55	75

The management has decided to give bonus of Rs 5, 10, 15, 20 and 25 to each worker in the respective output group of 40 -50 , 50-60 and so on weekly outputs. Construct the suitable frequency distribution.

Solution

Frequency distribution of output and bonus

Output	Tally bar	No of workers	Bonus (Rs)
10 – 20	||	2	-
20 – 30	|||	3	-
30 – 40	𝍸	5	-
40 – 50	𝍸 𝍸 ||	12	5
50 – 60	𝍸 𝍸 𝍸 𝍸	20	10
60 – 70	𝍸 |||	8	15
70 – 80	𝍸 |	6	20
80 – 90	||||	4	25
		N = 60	

2.9 BIVARIATE FREQUENCY DISTRIBUTION

If we are interested to study the two variables of the same population simultaneously, it is called bivariate. And its distribution is called bivariate frequency distribution.

In bivariate frequency distribution the values of one variable are in rows and values of other variable are in columns; the values can be discrete or continuous. The frequencies in each class are kept in the body of the table.

The theoretical frame of a two–way bivariate frequency table is as shown below.

BI-VARAITE TABLE

Y / X	y_1	y_2	…	…	y_n	Total
x_1	f_{11}	f_{12}	…	…	f_{1n}	Σf_{1i}
x_2	f_{21}	f_{22}	…	…	f_{2n}	Σf_{2i}
…	…	…	…	…	…	…
x_m	f_{m1}	f_{m2}	…	…	f_{mn}	Σf_{mi}
Total	Σf_{j1}	Σf_{j2}	…	…	Σf_{jn}	$N = \Sigma f$

Where, N is the total frequency $x_1, x_2, \ldots x_m$ and $y_1, y_2, \ldots y_n$ are the values of variables X and Y respectively and $f_{11}, f_{12}, \ldots f_{mn}$ are their cell frequencies.

Example 7

The following figures give the ages in the years of newly married husbands and wives. Represent the data by a bivariate frequency distribution.

(Husband's age, wife's age): (25, 17), (26, 18), (27, 19), (25, 17), (20, 20), (24, 18), (27, 18), (28, 19), (25, 18), (26, 19), (25, 17), (26, 18), (29, 19), (25, 19), (27, 20), (26, 19), (25, 17), (26, 20), (26, 17), (26, 18)

Solution

Let, X = age of husband and Y = age of wife

We observe that the variable X takes the values from 24 to 28 and Y takes the values from 17 to 20.

We obtain the bivariate frequency distribution as given below:

Bivariate frequency distribution of the ages (in years)

Y	X	Age of Husbands (X)					
		24	25	26	27	28	Row Total
Age of wives (Y)	17	| (1)	||| (3)	| (1)	-	-	5
	18	| (1)	| (1)	||| (3)	| (1)	-	6
	19	-	| (1)	|| (2)	|| (2)	| (1)	6
	20	-	-	| (1)	| (1)	| (1)	3
Column Total		2	5	7	4	2	N = 20

2.10 TABULATION

Tabulation is the process of arrangement of statistical data in the systematic presentation of the information contained in rows and columns.

Prof. A.L. Bowley's manual of statistics refers tabulation as "The intermediate process between the accumulation of data in whatever from they are obtained and the final reasoned account of the result shown by the statistics".

It is one of the most important and ingenious devices for:

i. Presenting the data in a condensed form.

ii. Furnishing the maximum information contained in the data in the minimum possible space, without sacrificing the quality and useful ness of the data.

iii. It is final stage in collection and compilation of the data and forms the gateway for further statistical analysis and interpretations.

iv. It is suitable for comparisons, helps to obtain averaging, correlation of statistical data.

v. It is suitable for further diagrammatic and graphic representations. It is time consuming to comprehend.

vi. Tabulated data are attractive and impressive as compared to the data in the textual form.

2.10.1 Parts of Table

The various parts of a table vary from problem depending on the nature of the data and the purpose of investigation.

The common parts of a table are as follows:

a. Table number b.Title of the tables c. Stubs

d. Captions e. Body of the table f. Head note

g. Foot note h. Source

Table number

A table should be numbered for identification. The number may be put at the center, above the title or at the bottom of the table.

Title of the table

Every table should have a title. It should be clear, brief and self–explanatory. The title should be set in bold type so as to give it prominence at the top of the table.

Stubs

Each row of the table must have a heading. The designations of the rows are called stubs or row headings. Stubs clarify the figures in the rows. They should be brief.

Captions

A table has many columns. Sub– headings of the column are called captions or column headings. They should be well defined and brief.

Body of the table

It is the most vital part of the table. It contains the numerical information. It should be made as comprehensive as possible. The data presented in the body are arranged according to the description classified in captions and stubs.

Head note

Head note is given just below the title enclosed in brackets. The unit of measurement should always be stated along with the title if this is uniform throughout the body of table. If different units have been adopted, then they should be stated along stubs or captions. For example, the units of measurements are usually expressed as head note such as 'in rupees', 'in quintals' etc.

Footnote

It is always placed at the bottom of the table. It is a statement that contains explanation of some specific items that cannot be understood by the reader from the title, captions and stubs.

Source note

A note at the bottom of the table should always be given to indicate the primary source as well as the secondary source from which the data has been taken.

Format of a standard table

Title

Table number ..... Head notes (if any)

Stubs	Captions	Row total
	BODY OF TABLE	
Column total		Grand total

Footnote:

Source:

2.10.2 Purpose of Tabulation

Tabulation is carried out for numbers of reasons, however, some of the main purposes of tabulation are as follows:

i. To present the classified data into the arranged and ordered form.

ii. To facilitate comparisons.

iii. To help in reference.

iv. To simplify the complex data.

v. To give identity to the data.

vi. To facilitate in further statistical analysis.

2.10.3 Types of Tabulation

Generally, there are three types of tables which are classified as the following basis

On the Basis of Purpose

On the basis of purpose, the table may be of two types.

a. General purpose b. Special purpose

General purpose table

These types of tables are also known as reference tables or repository tables. They provide information for general use as reference. They usually contain detailed information but this information is not recorded for special purpose i.e. not for comparative studies, relationship or significance of the figures. Most of the tables prepared by government agencies are of this type.

Special purpose table

These are prepared for the comparative studies to know the relationship and the significance of the figures provided by the data for the analysis of ratio, percentages, degree are used to facilitate comparison. Such tables provide information for special purpose or particular objective.

On the Basis of Nature of Enquiry

On the basis of the nature or originality of the data, the tables may be classified into two classes.

a. Primary

b. Derived or derivative

Primary table

In a primary table, the statistical facts or figures are expressed in the original form. It contains absolute and actual figures and are not rounded numbers or percentages given table.

Derived tables

They contain facts and figures and results derived from the original primary data. They express the information in terms of ratios, percentages, aggregates or statistical measures like average dispersion, etc.

For example, the time series data is expressed in a primary table but a table expressing the trend values, short time fluctuations, seasonal variation and cyclic variation is derived table.

On the Basis of Coverage

On the basis of coverage of the data the table may be of two types:

i. Simple table ii. Complex table

If the data are classified with respect to a single characteristic, it is called simple table. If the data are grouped into different classes with respect

to two or more characteristics or criteria simultaneously, then it is called complex or manifold table.

Example 8

Draw a blank table to show the number of candidates sex-wise appearing in the Certificate, Bachelors and Masters level examinations of a university in the faculties of Humanities, Science and Management in the year 2006.

Solution

Sex-wise and faculty wise distribution of students in 2006

Table No. 1

Faculty / Sex / Level		Humanities			Science			Management		
		M	F	Sub total	M	F	Sub total	M	F	Sub total
Certificate	I									
	II									
Bachelors	I									
	II									
	III									
Masters	I									
	II									

Note : M indicates Male F indicates Female

I = First year, II = Second year, III = Third year

Example 9

In a simple study about the smoking habits in two towns the following data were derived. Tabulate the Following data.

	Town A	Town B
Male in total population:	48%	49%
Smokers	24%	26%
Male smokers	18%	21%

Solution

Total population = 100 % (suppose)

% of male in town A = 48% % of female in A = 100-48 =52%

% of smoker in town A= 24%

% of male smoker in town A = 18%

% of female smoker in town A=(24 – 18)% = 6%

% of Non smoker in town A = (100 -24)% = 76%

% of male non smokers in town A =(48 – 18)% = 30%

% of female non smoker in town A= (52 – 6)% = 46%

Similarly for town B,

General purpose table

These types of tables are also known as reference tables or repository tables. They provide information for general use as reference. They usually contain detailed information but this information is not recorded for special purpose i.e. not for comparative studies, relationship or significance of the figures. Most of the tables prepared by government agencies are of this type.

Special purpose table

These are prepared for the comparative studies to know the relationship and the significance of the figures provided by the data for the analysis of ratio, percentages, degree are used to facilitate comparison. Such tables provide information for special purpose or particular objective.

On the Basis of Nature of Enquiry

On the basis of the nature or originality of the data, the tables may be classified into two classes.

a. Primary

b. Derived or derivative

Primary table

In a primary table, the statistical facts or figures are expressed in the original form. It contains absolute and actual figures and are not rounded numbers or percentages given table.

Derived tables

They contain facts and figures and results derived from the original primary data. They express the information in terms of ratios, percentages, aggregates or statistical measures like average dispersion, etc.

For example, the time series data is expressed in a primary table but a table expressing the trend values, short time fluctuations, seasonal variation and cyclic variation is derived table.

On the Basis of Coverage

On the basis of coverage of the data the table may be of two types:

i. Simple table ii. Complex table

If the data are classified with respect to a single characteristic, it is called simple table. If the data are grouped into different classes with respect to two or more characteristics or criteria simultaneously, then it is called complex or manifold table.

Example 8

Draw a blank table to show the number of candidates sex-wise appearing in the Certificate, Bachelors and Masters level examinations of a university in the faculties of Humanities, Science and Management in the year 2006.

Solution

Sex-wise and faculty wise distribution of students in 2006

Table No. 1

Faculty / Sex / Level		Humanities			Science			Management		
		M	F	Sub total	M	F	Sub total	M	F	Sub total
Certificate	I									
	II									
Bachelors	I									
	II									
	III									
Masters	I									
	II									

Note : M indicates Male F indicates Female

I = First year, II = Second year, III = Third year

Example 9

In a simple study about the smoking habits in two towns the following data were derived.

	Town A	Town B
Male in total population:	48%	49%
Smokers	24%	26%
Male smokers	18%	21%

Tabulate the above data.

Solution

Total population = 100 % (suppose)

% of male in town A = 48% % of female in A = 100-48 =52%

% of smoker in town A= 24%

% of male smoker in town A = 18%

% of female smoker in town A=(24 – 18)% = 6%

% of Non smoker in town A = (100 -24)% = 76%

% of male non smokers in town A =(48 – 18)% = 30%

% of female non smoker in town A= (52 – 6)% = 46%

Similarly for town B,

Table showing smoking habits of town A and B in percentage:

Attribute	Town A			Town B		
	M	F	Sub total	M	F	Sub total
Smokers	18	6	24	21	5	26
Non Smokers	30	46	76	28	46	74

Total	48	52	100	49	51	100

Note : M stands for male
F stands for female

Exercise – 3

1. Define classification of data. Write different modes of classification of statistical data.
2. What are the objectives of classification of data?
3. Discuss the problems in the construction of frequency distribution from raw data with particular reference to the choice of number of classes and class limits.
4. Explain, in brief, the basic principles and rules involved for forming a grouped frequency distribution.
5. Distinguish between:
   i. Discrete and continuous frequency distribution
   ii. Inclusive and exclusive frequency distribution
   iii. Less than and more than cumulative frequency distribution
   iv. Classification and tabulation of data
6. What are the terms involved in the continuous frequency distribution?
7. Define tabulation of statistical data. What are the advantages of tabulation? What are the different parts of statistical table?
8. What do mean by bi-variate frequency distribution?
9. Explain the function and importance of tabulation.
10. Form a frequency distribution table for the following data.

Marks Above	0	10	20	30	40	50
No. of students	30	22	21	14	10	0

Construct a less than cumulative frequency distribution table.

11. Construct a frequency distribution for the daily wages (Rs) of 30 laborers in a factory using equal class intervals one of them being (20 - 22)

14	16	16	14	22	13	15	24
12	23	14	20	17	21	18	18
19	21	17	16	15	11	12	21
20	17	18	11	22	23		

12. Number of books issued from a library per day is given below for 40 days.

84	80	91	81	88	82	87	97
93	99	75	85	72	92	84	90
87	78	83	98	86	80	93	86
88	83	82	101	89	82	85	95
80	89	84	76	81	103	92	94

Using class interval of 5, classify the number of books issued in the form of a frequency distribution.

13. Using Sturges' rule, $K = 1 + 3.322 \log_{10} N$, classify in equal intervals the following data of hours worked by 50 workers for a period of a month in certain factory.

195	110	175	161	157	155	108	164	128	114
178	165	133	151	71	94	97	42	30	62
138	156	167	124	164	116	149	104	141	103
150	162	149	79	113	69	121	93	143	140
144	187	184	197	87	40	122	203	148	117

14. Convert the following into an ordinary frequency distribution.
5 students get less than 3 marks, 12 students get less than 6 marks, 25 students get less than 9 marks and 33 students get less than 12 marks.
15. 25 observations of two variables X and Y are given below. Form a two-way frequency table showing the relationship between the two variables. Take class intervals of X as (10 -20), (20 -30) and so on that of Y as (100 - 200), (200 - 300) and so on.

X:	12	34	33	22	44	26	36	55	48	27
Y:	140	266	360	470	380	480	315	420	390	440
X:	37	21	69	51	27	42	43	43	52	57
Y:	390	590	590	250	550	360	360	570	290	416
X:	44	48	48	52	41					
Y:	380	492	370	312	330					

16. Prepare a suitable frequency distribution of the weight (Kg) of 33 students of a class given below. Convert it into exclusive frequency distribution.

42	74	40	60	72	15	41	61	75	63	63
53	70	76	64	50	67	65	78	77	56	55
68	69	71	60	79	79	54	73	59	71	70

17. The following are the weekly productions of X in units of 60 workers in a manufacturing company.

23	48	51	64	82	19	33	50	39	72
49	57	64	48	52	39	25	77	88	35
41	72	62	49	32	46	54	67	46	55
52	57	53	69	63	54	56	51	75	85

68	55	52	45	40	57	20	42	75	55
40	54	56	62	16	50	51	44	82	80

The management has decided to provide five bonus of Rs 5, 10, 15, 20, 25 and 30 to each of the workers in the respective output group of 10 or over up to 70 weekly outputs. Construct the frequency distribution of bonus and find the total bonus received by the workers.

18. Construct an ordinary frequency table from the following data:

Daily wages (Rs)	Number of workers
Below 20	10
Below 30	34
Below 40	61
Below 50	78
Below 60	89
Below 70	97
Below 80	100

19. Prepare a bivariate frequency distribution of marks obtained by 18 students in statistics and English.

Marks in Statistics	70	75	80	70	85	70	90	90	80
Marks in English	55	57	60	45	52	57	65	65	57
Marks in Statistics	85	75	90	70	85	80	70	85	75
Marks in English	61	55	45	55	60	45	52	57	57

20. Draw a blank table to show the number of students enrolled in a campus according to sex and faculty (Science, Commerce and humanities) in each of the years 2062 – 2064.

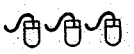
21. Out of 2000 voters in a place 1000 live in urban area . Out of 1170 voters in the favour of candidate A, 550 vote from urban area. Construct a table to represent this data.

22. In a simple study about drinking habit of milk in three towns, the following information was received.

Town A: Females were 45%, total milk drinkers were 40% and males non-milk drinkers were 20%.

Town B: Males were 50% males non- milk drinkers were 25% and females milk drinkers were 15%.

Town C: Males were 40%, Females non-milk drinkers 25% and male non-milk drinkers 40%.



Chapter 4

Diagrammatic and Graphic Presentation of Data

Diagrams and graphs are another important, convincing, appealing and easily understood method of presentation of statistical data other than classification and tabulation. Graphs and diagrams are nothing but the presentation of statistical data in the form of geometrical figures like points, lines, bars, rectangles, circles etc.

Importance of Diagram and Graph

Following are the important advantages of Diagrammatic and Graphic presentation of data.

i. Diagrams and graphs gives bird eye view of complex data.

ii. They can easy to understand even by a common man.

iii. They have long lasting impression.

iv. They save time and labour.

v. They facilitate comparison.

Difference between diagrams and graphs

Difference between diagrams and graphs are as follows.

i. Diagrams are constructed on plane paper whereas the graphs are on graph paper.

ii. Diagrams may be one, two, and three dimensional whereas graphs are of only two dimensional in general.

iii. The numerical data in diagrams are presented by bars, rectangles, circles, cubes etc whereas in graphs these data are presented in terms of points and lines.

iv. Graphs help in studying the further mathematical and statistical relationship but diagrams are only for the comparison.

v. Construction of graphs is easier in comparison to diagrams.

vi. Diagrams are more attractive to eye and common to all individual whereas graphs are only common to statistician and researcher.

General rules for constructing diagrams

Before constructing the diagrams we consider the following points.

Title

An attractive title must be written to each diagram, which should be brief, clear and self explanatory. It should be written either at the top or at the bottom of the diagram.

3.2 GENERAL RULES FOR CONSTRUCTING DIAGRAMS

Before constructing the diagrams we consider the following points.

i. **Title**

An attractive title must be written to each diagram, which should be brief, clear and self explanatory. It should be written either at the top or at the bottom of the diagram.

ii. **Proportion between width and height**

An appropriate proportion between width and height of the diagram should be used. A general rule suggested by Lutz is that 1 for short side and $\sqrt{2}$ for long side.

iii. **Selection of Scale**

A suitable scale should be adopted on the basis of horizontal and vertical scale. The scale showing the values should be in even number or in multiple of 5. For example, "income in thousand", "size of population in million" etc.

iv. **Selection of a diagram**

A suitable diagram must be selected based on nature of data, magnitude of observation etc.

v. **Neatness and cleanliness**

Diagrams should be draw neatly with the help of geometrical instruments. It should be clean and attractive.

vi. **Footnote**

The footnote should be written at the bottom of the diagram.

vii. **Index**

An index should be given for the clear understanding of the diagram.

viii. **Simplicity**

Diagrams should be as simple as possible so that they are easily understood.

3.3 TYPES OF DIAGRAMS

The various types of diagrams are given below.

i. Simple bar diagram.
ii. Sub divided bar diagram.
iii. Percentage bar diagram.
iv. Multiple bar diagram.
v. Pie- chart or pie- diagram

3.3.1 Simple Bar Diagram

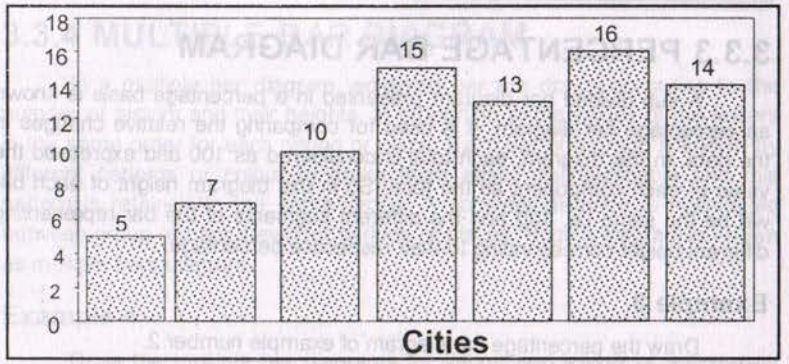
It is a diagram used to represent only one variable. It is used to comparative study of two or more values of a single variable. The different bars are drawn for the different values of the single variable on the same baseline. The height of the bars is used to represent the value of variable and the width of the bar is used to make the diagram attractive and understandable. The different bars should be the same width.

Example 1

Present the following data by a simple bar diagram

Cities	1	2	3	4	5	6	7
Population in Millions	5	7	10	15	13	16	14

Simple bar diagram



3.3.2 Sub-Divided Bar Diagram

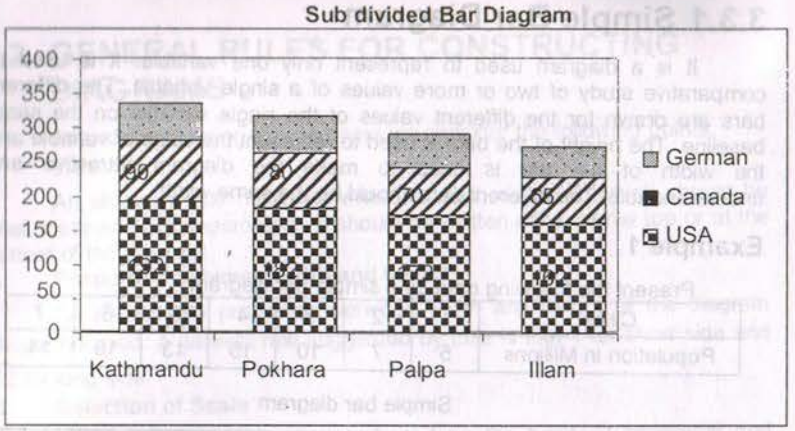
A bar diagram represents the magnitude of a single factor according to time periods, places, items etc .But when the magnitude of the factor is given with its sub-factors, each bar is further sub divided into components in proportion to the magnitude of the sub factors . Such a diagram is known as sub divided bar diagram.

Example 2

Draw sub divided bar diagram of the following data.

Countries wise tourists in various cities of Nepal

Country / cities	Kathmandu	Pokhara	Palpa	Illam
USA	192	182	172	162
Canada	90	80	70	65
German	55	54	50	45



3.3.3 PERCENTAGE BAR DIAGRAM

A sub divided bar diagram presented in a percentage basis is known as percentage bar diagram. It is used for comparing the relative changes in the data. In this diagram, each total is considered as 100 and expressed the value of each component as the total. So in this diagram height of each bar will be the same i.e. 100 and the different segments of the bar representing different height corresponding to their respective percentage.

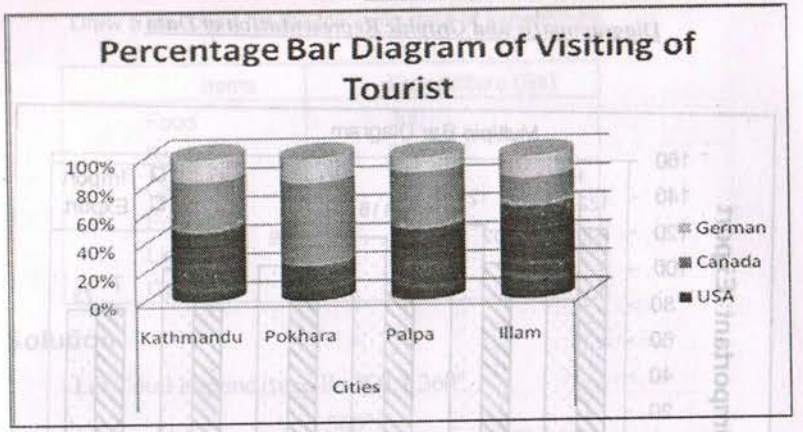
Example 3

Draw the percentage bar diagram of example number 2.

Solution

Percentage of total numbers of tourists

Country	Cities			
	Kathmandu	Pokhara	Palpa	Illam
USA	49	25	50	65
Canada	35	58	40	20
German	16	17	10	15
Total	100	100	100	100



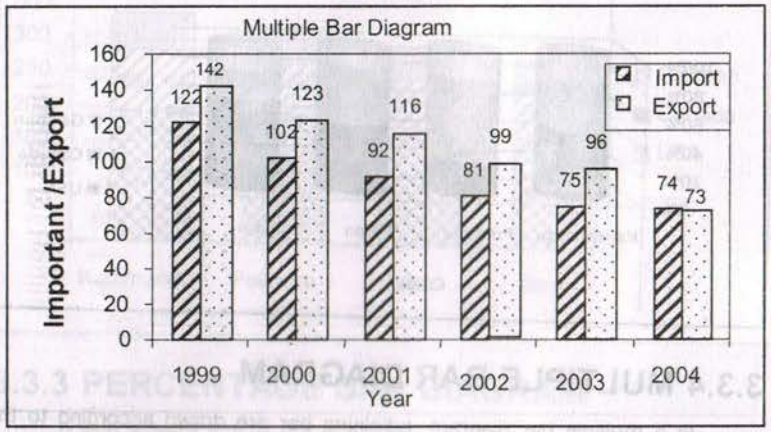
3.3.4 MULTIPLE BAR DIAGRAM

In a multiple bar diagram, adjoining bar are drawn according to the number of factors and their heights, In proportion to the values of the factors in the same order for each period or place. Each bar of a group is shown by different patterns or colour to make them easily distinguishable and this pattern is retained in the entire group. A constant distance is maintained between group of bars drawn for periods or places. Such a diagram is known as multiple bars diagram.

Example 4

Draw the multiple bar diagrams for import and export of a company given below.

Year	Import ('000 Rs.)	Export ('000 Rs.)
1999	122	142
2000	102	123
2001	92	116
2002	81	99
2003	75	96
2004	74	73



3.4 PIE-CHART

A pie-char is a circular diagram which is usually used for depicting the components of a single factor. Pie chart is an angular two dimensional diagram in which area of different sectors of circle can represent items of statistical data. The circle is divided into segments, which are in proportion to the size of the components. They are shown by different patterns of colors to make them attractive. In this diagram we compare the area of different components. The main limitation of a pie chart is that negative data can't be depicted.

To calculate the angle of each component, we first equate the total observation equal to 360° i.e. angle at center of the circle. Then the corresponding angle for any given value = $\frac{360°}{\text{Total value}} \times$ Given value. For a single pie chart, the radius may be any value which makes the chart attractive. But in case of two pie charts at a time radii of the circle of the pie chart must be in the proportion of square root of the total observations.

Therefore,

$r_1 : r_2 = \sqrt{\text{total of first sample}} : \sqrt{\text{total of second sample}}$

The comparison of the pie diagrams is to be made on the basis of the areas of the circles and to various sectors that are difficult to be ascertained visually with precision. Generally sub-divided or percentage bar diagrams are preferred to pie diagrams for studying the changes in the total and component parts. Moreover, pie diagrams are difficult to construct and compare with bars. If the number of component parts is more than 6, a pie chart is not preferable to construct.

Example 5

Draw a pie–chart for the following data of expenditure.

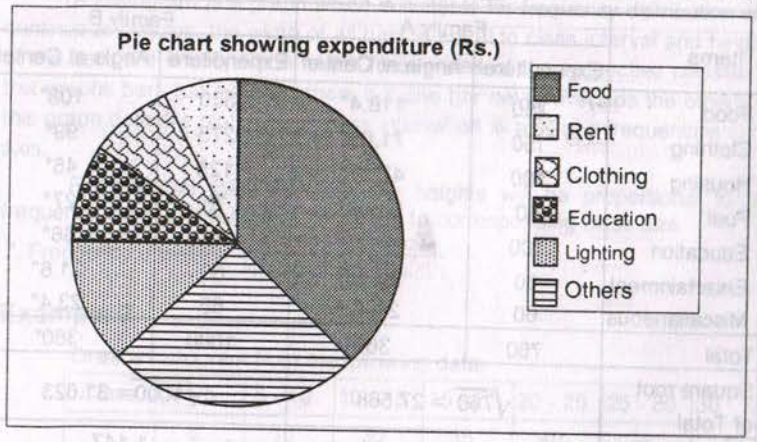
Items	Expenditure (Rs)
Food	300
Rent	200
Clothing	100
Education	75
Lighting	65
Others	60

Solution

Let Total expenditure Rs 800 = 360°

Then angle for Food = $\frac{360°}{800} \times 300 = 135°$ and so on

Items	Expenditure (Rs)	Angle at center (degree)
Food	300	135
Rent	200	90
Clothing	100	45
Education	75	33.75
Lighting	65	29.25
Others	60	27
Total	800	360



Example 6

Construct a pie-diagram for the following data of weekly expenditure of two families from Kathmandu.

Expenditure on	Family A	Family B
Food	250	300
Clothing	150	275
Housing	100	125
Fuel	50	75
Education	100	100
Entertainment	50	60
Miscellaneous	60	65
Total	760	1000

Solution

For Family A ,Taking Rs 760 = 360°

Rs 1 = $\frac{360°}{760}$ ∴ angle at center for food = $\frac{360°}{760}$ ×250 = 118.42°

In similar manner, other angles are calculated.

Similarly for family B, Taking Rs 1000 = 360°

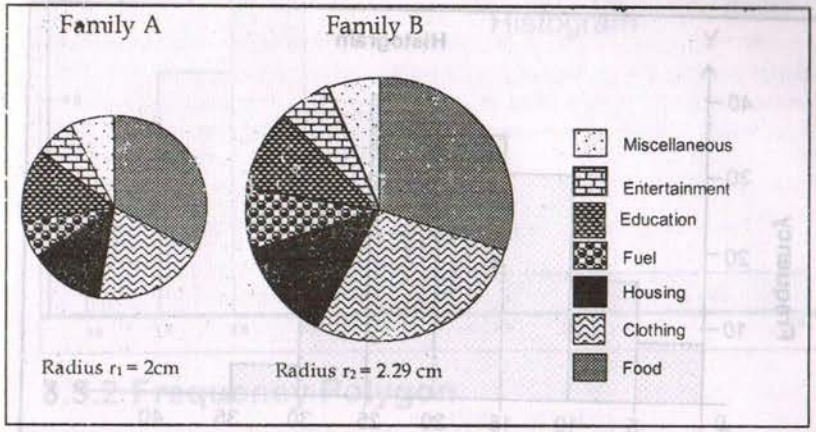
Rs 1 = $\frac{360°}{1000}$ Therefore, angle at center for food = $\frac{360°}{1000}$ ×300 = 108°

In similar manner, other angles are calculated.

Calculation table for pie-diagram

Items	Family A		Family B	
	Expenditure	Angle at Center	Expenditure	Angle at Center
Food	250	118.4°	300	108°
Clothing	150	71.05°	275	99°
Housing	100	47.37°	125	45°
Fuel	50	23.68°	75	27°
Education	100	47.37°	100	36°
Entertainment	50	23.68°	60	21.6°
Miscellaneous	60	28.42°	65	23.4°
Total	760	360°	1000	360°
Square root of Total	$\sqrt{760}$ = 27.568		$\sqrt{1000}$ = 31.623	
Radii	1		1.147	

Pie-chart showing expenditure of two families from Kathmandu



3.5 GRAPHICAL REPRESENTATION OF DATA

The most common graphical representation of date are:

i. Histogram
ii. Frequency Polygon
iii. Frequency Curve
iv. Ogives or Cumulative frequency Curve.

3.5.1 Histogram

A Histogram is a graph which is suitable for frequency distribution with continuous classes, the width of all bars is equal to class interval and heights of the bars are in proportion to the frequencies of the respective classes. In this graphs bars touch each others but one bar never overlaps the others. In this graph data are plotted with class interval on x- axis and frequencies on y-axis.

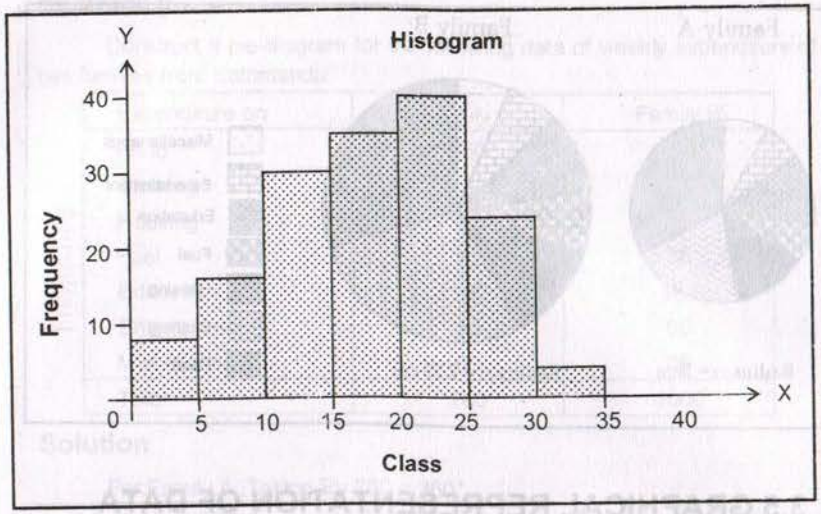
For unequal class intervals the heights will be proportional to the frequency density. It is ratio of frequency to corresponding class size.

∴ Frequency density = $\frac{\text{Class frequency}}{\text{length of class width}}$

Example 7

Draw a histogram from the following data.

Class	0 - 5	5 - 10	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35
f	8	16	30	35	40	24	4



Example 8

Draw a histogram from the following data

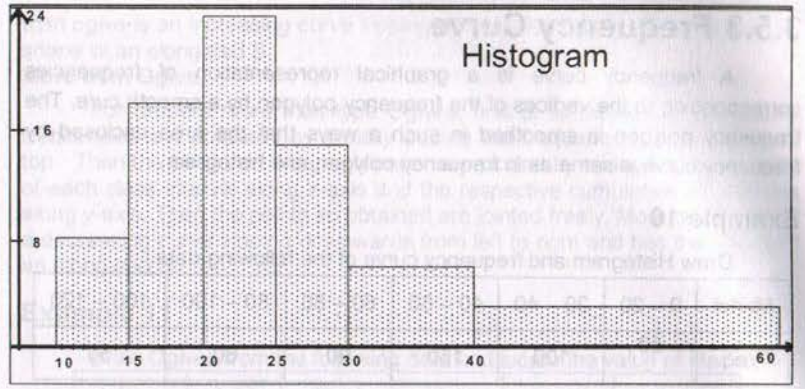
Wage	10 – 15	15 – 20	20 – 25	25 – 30	30 – 40	40 – 60
frequency	8	18	25	15	12	12

Solution

It is a case of Unequal Class interval therefore we first calculate frequency density of each class as,

$$\text{Frequency density of a class} = \frac{\text{Class frequency}}{\text{Length of class width}}$$

Wage	frequency	Frequency density
10-15	8	8
15-20	18	18
20-25	25	25
25-30	15	15
30-40	12	12/2 = 6
40-60	12	12/4 = 3



3.5.2 Frequency Polygon

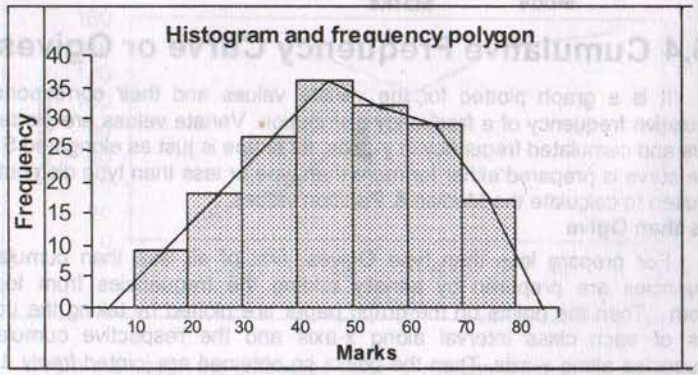
It is another method of graphical representation of frequency distribution .When the distribution is discrete the frequency polygon is obtained by plotting the points with values of the variable on the x- coordinate and the corresponding frequencies on the y-coordinate and joining the points by straight line.

If the frequency distribution is continuous then frequency polygon is drawn by joining the mid points of the top of the adjacent bars of a histogram in order. The main purpose of frequency polygon is to find the mode.

Example 9

Draw the histogram and frequency polygon of a following data.

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	9	18	27	36	32	29	17



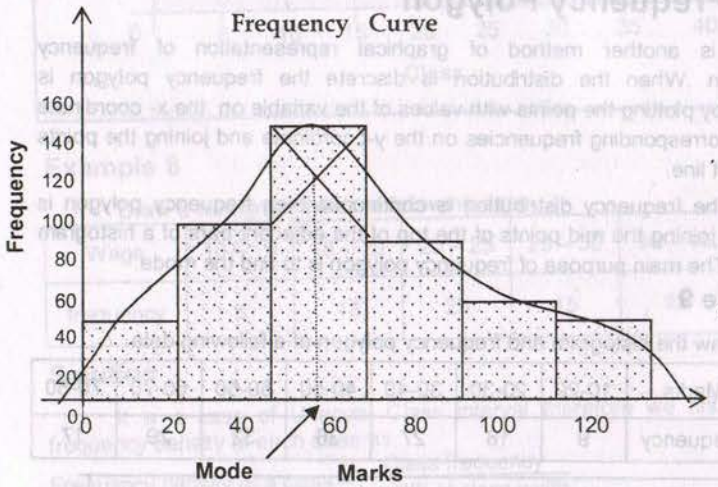
3.5.3 Frequency Curve

A frequency curve is a graphical representation of frequencies corresponding to the vertices of the frequency polygon by a smooth cure. The frequency polygon is smoothed in such a ways that the area enclosed by frequency curve is same as in frequency polygon and histogram.

Example 10

Draw Histogram and frequency curve of the following data.

Marks	0 – 20	20 – 40	40 – 60	60 – 80	80 – 100	100 – 120
Freq.	50	100	150	90	60	50



3.5.4 Cumulative Frequency Curve or Ogives

It is a graph plotted for the variate values and their corresponding cumulative frequency of a frequency distribution. Variate values are plotted in x-axis and cumulated frequency in y- axis. Its shape is just as elongated S. An ogive curve is prepared either for more than type or less than type distribution. It is used to calculate the Median & Partition values.

Less than Ogive

For prepare less than type Ogives, first of all less than cumulative frequencies are prepared by serially adding the frequencies from top to bottom. Then the points on the graph paper are plotted by taking the upper limits of each class interval along x-axis and the respective cumulative frequencies along y-axis. Then the points so obtained are jointed freely. Less than ogive is an increasing curve sloping upward from left to right and has the shape of an elongated S.

More than Ogive

For prepare more than type Ogives, first of all more than cumulative frequencies are prepared by serially adding the frequencies from bottom to top. Then the points on the graph paper are plotted by taking the lower limits of each class interval along x-axis and the respective cumulative frequencies along y-axis. Then the points so obtained are jointed freely. More than ogive is a decreasing curve sloping downwards from left to right and has the shape of an elongated S.

Example 11

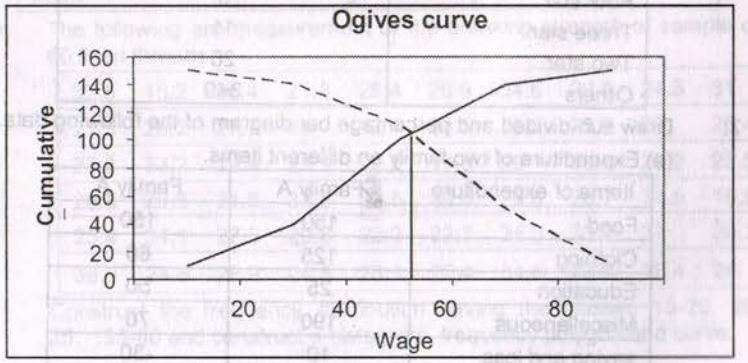
Draw Ogives from the following data and locate the value of Median.

Wage	0-20	20-40	40-60	60-80	80-100
Freq.	10	30	60	40	10

Solution

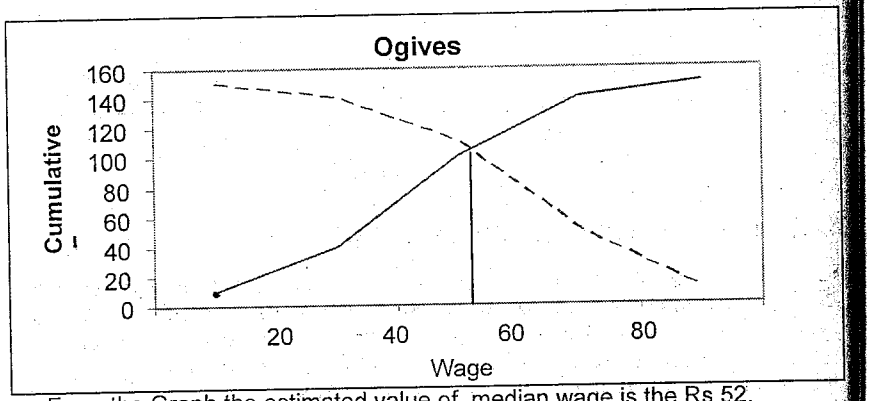
First of all prepare more than and less than cumulative frequency as,

Less than Method		More than Method	
wage	frequency	wage	frequency
Less than 20	10	More than 0	150
Less than 40	40	More than 20	140
Less than 60	100	More than 40	110
Less than 80	140	More than 60	50
Less than 100	150	More than 80	10



From the Graph the estimated value of median wage is the Rs 52.

Less than Method		More than Method	
wage	frequency	wage	frequency
Less than 20	10	More than 0	150
Less than 40	40	More than 20	140
Less than 60	100	More than 40	110
Less than 80	140	More than 60	50
Less than 100	150	More than 80	10



From the Graph the estimated value of median wage is the Rs 52.

Exercise - 4

1. What are graphs and diagrams? What are the importance and limitation of diagrammatic and graphical representation of statistical data?
2. Distinguish between "Histogram and Bar diagram", Pie chart and sub-divided bar diagram" and less than and more than Ogives"
3. Explain the different types of graphs used for representation of frequency distribution.
4. Draw the simple bar diagrams of the following data.

Department	Civil	Computer	Electronic	Architecture
No. of students	52	48	44	26

5. The following are the number of automobile accidents that occurred at 60 major intersections in a certain city during the first of July weekend.

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

Group these data into frequency distribution showing how often each of the values occurs and draw a bar diagram.

6. The following data regarding the number of different categories of hotel in Kathmandu. Represent the data by simple bar diagram.

Categories	No. of hotels
Five star	5
Four star	8
Three star	11
Two star	26
Others	349

7. Draw sub-divided and percentage bar diagram of the following data.

(a) Expenditure of two family on different items.

items of expenditure	Family A	Family A
Food	150	150
Clothing	125	60
Education	25	50
Miscellaneous	190	70
saving and loss	10	-30
Total income	500	300

(b) Following data show the percentage of total population in rural and urban are of Nepal

	Percentage of total Population in	
	Rural	Urban
Infant and young children	13.7	12.9
Boys and girls	25.1	23.2
Young men and women	32.3	36.5
Middle-aged men and women	20.4	20.1
Elderly persons	8.5	7.3

8. Following data gives the numbers of student admitted and passed in final exam of the past 4 semesters, represent it by multiple bar diagram.

Programs	Number of students	
	Admitted	Passed
Civil	45	32
Computer	48	36
Electronics	49	40
Architecture	26	16

9. Display the following data of profit and expenses in different year by the suitable bar diagram.

Year	Profit (million Rs.)	Expense (million Rs.)
1996	30	28
1997	32	27
1998	28	21
1999	27	26
2000	31	28

10. The following are measurement of the breaking strength of sample of 60 linen threads

32.5	15.2	35.4	21.3	28.4	26.9	34.6	29.3	24.5	31

21.5	28.3	27.7	25	32.7	29.5	30.2	23.9	23	26.4
27.3	33.7	29.4	21.9	29.3	17.3	29	36.8	29.2	23.5
20.6	29.5	21.5	37.5	33.6	29.6	21.8	25.7	34.8	18.6
25.4	34.1	27.5	29.6	22.2	22.7	31.3	33.2	37	28.3
36.9	24.6	28.9	24.8	28.1	25.4	34.5	23.6	38.4	24

Construct the frequency distribution having the classes 15-20, 20-25....35-40 and construct a histogram, frequency polygon and curve.

11. The following data relate to the expenditure of two families per month.

Items of expenditure	Family A	Family B
Food	300	200
Rent	200	100
Clothing	150	100
Education	200	50
Lighting	50	25
Others	100	25

Represent the above data by angular diagram.

12. Draw the histogram for the following data.

i)

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80
frequency	10	15	32	42	26	12	9

(ii)

Mid values	20	30	40	50	60	70
no. of workers	4	12	15	18	30	16

13. Draw histogram and a frequency polygon for the following data given below. (i)

Marks in statistics	10-20	20-30	30-40	40-50	50-60	60-70	70-80
students	18	23	37	47	26	16	5

(ii)

Class intervals	0-5	5-10	10-15	15-25	25-30	30-40
frequency	8	10	30	40	24	12

14. Draw histogram, frequency polygon and frequency curve of the following data:

Daily wage (Rs.)	10-15	15-20	20-25	25-30	30-40	40-60
no. of workers	8	18	25	15	12	12

15. Draw less then and more than Ogives from the following data given below and estimate the value of median.

Marks	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100
students	50	100	150	90	60	50	25	13	5

Chapter 5

Measure of Central Tendency

Introduction

After the collection of data, the analysis is done with the help of different statistical tools like measure of central tendency, dispersion, skewness, correlation and regression etc.

The measure of central tendency is defined as an average. It is a single value within which the range of data tends to cluster which represents a groups of individual values in a simple and concise manner so that the mind can get a quick understanding of the general idea of the individual in the group. Since the value lies within the range of data, it is known as measure of central tendency. The objectives of averages are.

i. To facilitate comparison.
ii. To know about population from a sample.
iii. To present the silent features of mass of complex data.
iv. To trace statistical relations.
v. To helps in decision making.

Properties of good measure of central tendency

The desire properties or requisites of a good measure of central tendencies are as follows.

i. It should be based on all the observations of a set of values.
ii. It should be rigidly defined and its values should be definite.
iii. It should be easily computable.
iv. It should be least affected by extreme values.
v. It should fluctuate least from sample to sample drawn from the same population.
vi. It should be suitable for further statistical and mathematical treatments.

Function of central tendencies

The various functions of central tendencies or averages are as follows.

i. To present a simple and coincide picture of large and complicated set of data.
ii. It makes possible and easier to compare two and more groups of data.
iii. To provides a study of different groups of data.
iv. It facilitates the interpretation of data.
v. It helps in taking decisions. For example one can say whether the income of a family is above or below the average income.

Limitations of Central tendencies

The general limitations of an average are as follows.

4.1.3 Limitations of Central Tendencies

The general limitations of an average are as follows.

i. An average represents a group as a whole not an individual.
ii. Often an average is a value which does not exist in the set of data.
iii. Some times an average comes out a value which is not feasible for example average size of family is 3.36

4.2 TYPES OF CENTRAL TENDENCIES

The following are the important types of averages.

i. Arithmetic average
ii. Geometric Mean
iii. Harmonic Mean
iv. Median
v. Mode

4.2.1 Arithmetic Average

Arithmetic mean of set of observation may be defined as the sum of observation divided by the total number of value in the set. It is denoted by $\bar{x}$

i.e. $\bar{x} = \frac{\text{Sum of observation}}{\text{Number of observation}}$

$$\therefore \bar{x} = \frac{\Sigma x}{n}$$

Calculation of mean

For Individual series

Let $x_1, x_2, x_3 \ldots x_n$ be the values of the variables .then mean is defined as

$\therefore \bar{x} = \frac{\Sigma x}{n}$ This is called "direct method"

And $\bar{x} = A + \frac{\Sigma d}{n}$ This is called short cut method,

Where 'A' is assumed mean, d = x – A, deviation taken from assumed mean. When the sizes of the values are very big and the numbers of observations are too large, we find mean by taking the deviation of the items from the assumed mean.

Example 1

Find the arithmetic mean from the data of expenditure of 5 students as 10, 20, 30, 40, and 50.

Solution

Here number of observations n = 5

For direct method

Total expenditure = $\Sigma x = 10 + 20 + 30 + 40 + 50 = Rs.150$

$\therefore \bar{x} = \frac{\Sigma x}{n} = \frac{150}{5} = Rs.30.$

And for short cut method,

Let assumed mean (A) = 40

Calculation table

Expenditure (X)	d = X – 40
10	-30
20	-20
30	-10
40	0
50	10

Therefore, Mean is $\bar{x} = A + \frac{\Sigma d}{n} = 40 - \frac{50}{5} = Rs.\ 30$

Note that assumed mean is any number between lowest and highest number, however middle number is suitable.

For Discrete Series and continuous series

For discrete and continuous data also known as frequency data, Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n$ be n observation with corresponding frequency $f_1, f_2, \ldots\ldots\ldots\ldots f_n$. Then the arithmetic mean or simply mean is defined as,

$\bar{x} = \frac{\Sigma fx}{N}$, $N = \Sigma f$ is total frequency, This is for direct method.

$\bar{x} = A + \frac{\Sigma fd}{N}$ where $d = x - A$, This is for short cut method.

$\bar{x} = A + \frac{\Sigma fd'}{N} \times h$ where $d' = \frac{x-A}{h}$, This is for step deviation method

Here 'A' is assumed mean and 'h' is class width. [In continuous series × is the mid values of class]

Example 2

Calculate the mean for the following data.

Wage	10	20	30	40	50
Freqqency	2	4	10	8	6

Solution

Calculation table of mean

Wage (Rs.)	Frequency (f)	fx
10	2	20
20	4	80
30	10	300
40	8	320
50	6	300
150	**30**	**1020**

Here N = 30 and $\Sigma fx = 1020$

Therefore mean $\bar{x} = \frac{\Sigma fx}{N} = \frac{1020}{30}$ = Rs 34.

Example 3

Calculate mean of the following data by using

i) direct method, ii) short cut method, iii) step deviation method.

Income(Rs)	50-100	100-150	150-200	200-250	250-300
Persons	5	8	15	16	6

Solution

Let assumed mean A = 175 and h = 10 d = x-175 and $d' = \frac{x - 175}{50}$

wage	f	x	f x	d	fd	D^l	Fd^l
50 – 100	5	75	375	-100	-500	-2	-10
100 – 150	8	125	1000	-50	-400	-1	-8
150 – 200	15	175	2625	0	0	0	0
200 – 250	16	225	3600	50	800	1	16
250 – 300	6	275	1650	100	600	2	12
Total	**50**	**875**	**9250**	**0**	**500**	**0**	**10**

Here N = 50, A = 175 and h = 100-50 = 50

i. For direct method

$\bar{x} = \frac{\Sigma fx}{N} = \frac{9250}{50}$ = Rs 185,

ii. for short cut method

$\bar{x} = A + \frac{\Sigma fd}{N} = 175 + \frac{35000}{50} = 175 + 10$ = Rs 185

iii. For step deviation method

$\bar{x} = A + \frac{\Sigma fd'}{N} \times h = 175 + \frac{10}{50} \times 50$ = Rs. 185

Example 4

Find the arithmetic mean of the data of monthly income of 585 families.

Monthly income	No. of families
1000 – 2000	50
2000 – 3000	75
3000 – 4000	120
4000 – 5000	150
5000 – 6000	100
6000 – 7000	65
7000 – 8000	25

Solution

Here number of families N = 585, let $d = \frac{x - A}{h} = \frac{x - 4500}{1000}$

Monthly income	no. of families	Mid values(x)	fx	$d = \frac{x - a}{h}$	fd
1000-2000	50	1500	75000	-3	-150
2000-3000	75	2500	187500	-2	-150
3000-4000	120	3500	420000	-1	-120
4000-5000	150	4500	675000	0	0
5000-6000	100	5500	550000	1	100
6000-7000	65	6500	422500	2	130
7000-8000	25	7500	187500	3	75
Total	**585**	**31500**	**2517500**	**0**	**-115**

(i) For direct method $\bar{x} = \frac{\Sigma fx}{N} = \frac{2517500}{585}$ = Rs 4303.42

(ii) for step deviation method

$\bar{x} = A + \frac{\Sigma fd'}{N} \times h = 4500 + \frac{-115}{585} \times 1000$ = Rs. 4303.42

Properties of Arithmetic Mean

The important properties of the arithmetic mean are as follows:-

i. The sum of the deviations of the given observation from their mean is always zero i.e. $\Sigma (x - \bar{x}) = 0$

Proof: $\Sigma (x - \bar{x}) = \Sigma x - n\bar{x} = \Sigma x - \Sigma x = 0$

ii. The sum of the square of the deviation of the given observation from their mean is minimum, i.e. $\Sigma (x - \bar{x})^2$ is minimum

Proof: Let, if possible $\Sigma(x - a)^2$ is minimum , for this let

$y = \Sigma (x - a)^2$,

For minima of function $y = \Sigma (x - a)^2$ we must have $\frac{dy}{da} = 0$ and $\frac{d^2y}{da^2} > 0$

Therefore, $\frac{dy}{da} = 2\Sigma(x - a)(-1)$

For maxima and minima $\frac{dy}{da} = 0$

Or, $2\Sigma(x - a)(-1) = 0$

Or, $\Sigma x - n a = 0$

$\Rightarrow a = \bar{x}$ and $\frac{d^2y}{da} > 0$.

Hence the sum of the square of the deviation of the given observation from their mean is minimum.

iii. Since $\bar{x} = \frac{\Sigma x}{n}$ then $\Sigma x = n\bar{x}$

iv. **Combined mean:** Let n_1 and n_2 be the size of two samples with means $\bar{x}_1$, $\bar{x}_2$ Then the combined mean of the whole group of size n_1+n_2 is denoted by $\bar{x}_{12}$ and defined as

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

Example 5

Find the average GPA of B.E. student of Civil and computer students of a college taken together, when GPA of 48 civil students is 3.56 and that of 42 computer students is 3.47.

Solution

Given number of student in civil $n_1 = 48$ and average

GPA $\bar{x}_1 = 3.56$

Number of student in computer $n_2 = 42$ and average GPA $\bar{x}_2 = 3.47$

We have combined mean

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2} = \frac{48 \times 3.56 + 42 \times 3.47}{48 + 42} = 3.518$$

Thus the combined GPA of whole students taken together is 3.518.

Example 6

The mean age of a combined group of men and women is 30 years. If the mean age of the group of men is 32 and that of a group of women is 27. Find out the percentage of men and women in the group.

Solution

Given mean age of men $\bar{x}_1$ =32 years. Mean age of women $\bar{x}_2 = 27$ years, the mean age of men and women taken together is $\bar{x}_{12} = 30$ years

Percentage of men (n_1) =? And Percentage of women (n_2) =?

Total number of persons = $n_1 + n_2 = 100$ (i)

We have combined mean $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

$$\text{or, } 30 = \frac{32\,n_1 + 27n_2}{100} = 3.518$$

$$\text{or, } 32\,n_1 + 27\,n_2 = 3000 \quad \text{(ii)}$$

On solving equation (i) and (ii) we get percent of men = 60% and percentage of women = 40%.

4.2.2 Weighted Mean

In general all the items under consideration don't have an equal importance; therefore we assign proper weights to the items according to their relative importance. Then computed mean is called weighted mean. Let $x_1, x_2, x_3, \ldots x_n$ be the n observations with corresponding weights $w_1, w_2, \ldots w_n$. Then the weighed mean $\bar{x}_w$ is given by

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \ldots.w_nx_n}{w_1 + w_2 + \ldots. + w_n} = \frac{\Sigma wx}{\Sigma w}$$

Example 7

Calculate weighted mean for the following data.

Items	Index No.	Weight (%)
Food	225	52
Clothing	175	8
Fuel and lighting	155	10
Housing	280	14
Others	150	16

Solution

CALCULATION

Items	x	w	wx
Food	225	52	11700
Clothing	175	8	1400
Fuel and lighting	155	10	1550
Housing	280	14	3920
Others	150	16	2400
Total	985	100	20970

Here total weight $\Sigma w = 100$, and $\Sigma wx = 20970$

Therefore weighted mean $\bar{x}_w = \frac{\Sigma wx}{\Sigma w} = \frac{20970}{100}$

Merit of Arithmetic mean

i. It is easy to calculate.
ii. It is rigidly defined.
iii. It is based on all the observations.
iv. Observations need not to be arranged in order to calculation of mean.
v. It provides sound basis for comparison of two or more series.
vi. It can be calculated if partial calculations are available.
vii. It is less susceptible to sampling fluctuations.
viii. It is most suitable for further statistical treatment.

Demerit of Arithmetic mean

i. It is affected too much by extreme values.
ii. Mostly it doesn't correspond to any value of the set of observations.

iii. It can't be calculated for frequency distribution with open end class.
iv. It doesn't convey any information about the spread or trend of data.
v. It is not a suitable measure of central value in case of highly skewed distribution.

4.2.3 Median

Median is a positional average i.e. its value depends on the position occupied by a value in the frequency distribution. Median is the value of the variable that divides the ordered set of values in to two equal halves i.e. 50% value are to the left of median and 50% are to the right of median. It is most preferable measure of location for asymmetric distribution.

Calculation of median

For individual series

To calculate median in case of individual series we follows the following points.

i. Arrange the data in either ascending or descending order of magnitude.
ii. If the number of observations is odd, the middle value gives the median and if number of observations is even there will be two middle values so the arithmetic mean of two middle values gives the median.
iii. Formula for calculating the median is as,

Median = Value of $\left(\frac{n+1}{2}\right)^{th}$ item. Here n is the number of observations.

Example 8

Find the median marks of the 9 students as: 35, 85, 39, 42, 53, 62, 71, 82, 48

Solution

First arrange the data in ascending order of magnitude as: 35, 39, 42, 48, 53, 62, 71, 82, and 85.
Here number of observation n = 9

Therefore Median = values of $\left(\frac{n+1}{2}\right)^{th}$ items = 5^{th} item.

Therefore Median = 53

Example 9

Find the median wages (Rs.) of 10 workers of a factory as: 250, 410, 267, 350, 280, 500, 750, 300, 450 and 200.

Solution

First arrange the data in ascending order of magnitude as: 200, 250, 267, 280, 300, 350, 410, 450, 500 and 750

Here number of observation n = 10

Therefore Median = values of $\left(\frac{n+1}{2}\right)^{th}$ items = 5.5^{th} item.

Therefore Median = $\frac{\text{Values of } 5^{th} + \text{Values of } 6^{th}}{2}$ items

$= \frac{300 + 350}{2} = 350.$

Hence the median wage is Rs. 350

For discrete series

To find the median in case of discrete series, we use the following procedure.

i. Arrange the data according to their ascending order of magnitude,
ii. Form the cumulative frequency distribution.
iii. To calculate median we use the following formula as

Median = Value of $\left(\frac{N+1}{2}\right)^{th}$ items, Where 'N' is the total number of observation.

Example 10

Find the median from the following data of rainfall.

Rainfall (in mm)	24	30	45	60	100
No. of days	15	16	20	24	15

Solution

Calculation of median is as follows

Rainfall (in mm)	No. of days (f)	c f
24	15	15
30	16	31
45	20	51
60	24	75
100	15	90

Here N = 90, Therefore Median = Value of $\left(\frac{N+1}{2}\right)^{th}$ items

$= \text{Value of } \left(\frac{90+1}{2}\right)^{th} = 45.5^{th}$ items

Cumulative frequency corresponding to 45.5 is 45.

Hence Median = 45 mm.

For Continuous series

To find the median in case of continuous series we use the following steps.

i. Prepare the less than cumulative frequency distribution.

ii. Find the value of $\frac{N}{2}$

iii. The class with cumulative frequency just equal or greater than $\frac{N}{2}$ is median class and this contains median.

iv. The value of median is calculated by using following formula.

$$\text{Median} = L + \frac{\frac{N}{2} - cf}{f} \times h$$

Where L is lower limit of median class,

$N = \Sigma f$, total frequency,

C.f. is cumulative frequency proceeding to the median class, 'f' is the frequency of median class, and 'h' is size of median class

Merit of median

Points in favour of median are as follows:

i. Median is not influenced by extreme values because it is a positional average.
ii. It can be calculated in case of distribution with open end intervals.
iii. Median can be located even if the data are incomplete.
iv. It can be located even for qualitative factor such as ability, honesty, beauty etc.

Demerit of median

Following are the demerits of median

i. A slight change in the series may bring drastic change in median value.
ii. In case of even number of observations or continuous series, median is an estimated value other than any value in the series.
iii. It is not suitable for further mathematical or statistical treatment except its use in mean deviation.

Example 11

Find the median from following data

Wage in Rs.	No. of workers
Below 1000	50
1000-1999	500
2000-2999	555
3000-3999	100
4000-4999	30
5000 and above	15

Solution

Calculation of cumulative frequency

Wage in Rs.	f	C f
Below 1000	50	50
1000 – 1999	500	550
2000 – 2999	555	1105
3000 – 3999	100	1205
4000 – 4999	30	1235
5000 and above	15	1250

Here value of $\frac{N}{2} = \frac{1250}{2} = 625$, therefore 2000 – 2999 is median class but the real limit of median class is 1999.5-2999.5

$$\text{Then Median} = L + \frac{\frac{N}{2} - cf}{f} \times h$$

Here L=1999.5, c f =550, f = 555 and h = 1000

$$\therefore M_d = 1999.5 + \frac{625 - 550}{555} \times 1000 = 2134.635$$

Hence the median wage is = Rs. 2134.634

4.2.4 Mode

The mode of the values of variable in a series of data is that value of a series which appear most frequently than any other and around which the other items of the set concentrate densely. It is denoted by M_o.

Computation of mode in case of continuous series

For frequency distribution with classes in ascending order of magnitude, the class corresponding to the maximum frequency is called model class and we use the following formula for calculation of mode.

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where L is lower limit of model class

$\Delta_1 = f_1 - f_0$ and $\Delta_2 = f_1 - f_2$

f_1 is maximum frequency

f_0 is frequency preceding to model class

f_2 is frequency succeeding to model class

h is size of model class

Merit of mode

i. It is easy to calculate and easy to understand.
ii. It is not at all affected by extreme observations.
iii. It can be obtained even in case of open-end class.
iv. It can be obtained by inspection or by graph.

Demerit of mode

i. It is not based on all the observations.
ii. It is not rigidly defined.
iii. It is not suitable for further mathematical and statistical treatment.

Example 12

Calculate the mode of the variate values as: 2, 3, 5, 3, 4, 6, 7, 3, 1, 5 and 3.

Solution

Here variate values 3 repeated maximum no of times, hence mode of the given data is 3.

Example 13

Find the model size of shoes sales in a departmental store from the following data.

Size of shoes	1	3	4	5	6
No of shoes	40	50	80	60	30

Solution

Here maximum frequency is 80, whose corresponding variate values is 4 Hence model sizes of shoes = 4.

Method of grouping

If the frequency distribution is regularly distributed, mode can be easily determined by inspection. If the maximum frequency occur more than one time or the maximum frequency occur at the very beginning or at the end of the observations and the given frequency is not in the regular order, then mode is calculated by using method of grouping. It consists of 2 tables grouping tables & analysis table.

The method of grouping is applied in order to make the frequency distribution regular. In real sense, the aim of grouping is to remove the bias prevent in the types of means discussed before and thereby minimizing their irregularities.

In grouping table, there are six columns except the column of the size of item or class interval's column. Their columns are completed in this ways.

Column I : Write the original or given frequencies in first column.

Column II : Add the given frequencies two by two and write the total corresponding to the middle of the item or class interval.

Column III : Leave the first frequency and then add two by two.

Column IV : Add the given frequency three by three.

Column v : Leave the first frequency and then add three by three.

Column VI : Leave the first two frequencies and then add three by three.

Then analysis table is prepared by using the grouping table in order to count the maximum frequency in each column. In analysis table the values of the variables are kept either in row or column and the column number are kept in another side of the table and their corresponding maximum frequencies for each column are listed. The maximum frequency in each column is entered by means of vertical bars. The value of the variables in the column, which occurrence maximum times is the mode.

Example 14

Find the model marks of the 55 B.E. students

Marks	30	35	40	45	50	55	60	65	70	75
No. of students	6	9	4	2	10	8	7	5	1	3

Solution

Here the frequencies are not in the regular order. So mode is calculated by grouping method

GROUPING TABLE

Marks	Column I	Column II	Column III	Column IV	Column V	Column VI
30	6	15	-		-	-
35	9		13	13		-
40	4	6			15	
45	2		12			16
50	**10**	**18**		**20**		
55	8		**15**		**25**	
60	7	12				**20**
65	5		6	13		
70	1	4			9	
75	3					

ANALYSIS TABLE

Marks → / Column no.	30	35	40	45	50	55	60	65	70	75
I					I					
II					I	I				
III						I	I			
IV				I	I	I				
V					I		I			
VI						I		I		
Total				1	4	5	3	1		

Here, Marks 55 repeated more than the other variates values. So, the value of modal marks is 55.

Example 15

Find the Model income from the following data.

Income(Rs)	50 -100	100 -150	150 -200	200 -250	250 -300	300 -350	350 -400
No. of persons	5	8	15	16	9	6	5

Solution

Here the class 200-250 is the model class, since this class contains the maximum frequency. And the mode is calculated by using the following formula.

$$M_o = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h = L + \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \times h$$

Where L=200, lower limit of model class

$\Delta_1 = f_1 - f_0$ and $\Delta_2 = f_1 - f_2$

f_1 =16, maximum frequency

f_0 =15, frequency preceding to model class

f_2 =9, frequency succeeding to model class

h =50, size of model class

$$\text{Therefore Mode} = 200 + \frac{16 - 15}{2 \times 16 - 15 - 9} \times 50 = 206.25.$$

Hence model income of the given data is Rs. 206.25

4.2.5 Partition Values

Partition values are variate values of variable, say X, which divide the total number of observations into equal number of parts. The equal number of parts may be of four, eight, ten, twelve, one hundred etc.

Quartiles, Deciles and Percentiles

Three variate values of the variable, say X, which divide the series or frequency in to four equal parts, are called quartiles for the corresponding distribution of X. There are three quartiles namely Q_1, Q_2, Q_3. Q_1 is called the lower quartiles and Q_3 is called the upper quartile, 25% of values are less than Q_1 and 25% values are larger than Q_3 and the rest 50% values lie between Q_1 and Q_3. Q_2divide the series in to two equal parts and hence it is the same as median.

Nine variate values of the variable, say X, which divide the series or frequency in to ten equal parts, are called deciles for the corresponding frequency distribution of X. There are nine deciles namely, $D_1, D_2, \ldots\ldots, D_9$ D_1 is called the lower deciles and D_9 is called the upper deciles, the items before D_1 and after D_9 are 10% Also the number of items that lies between two deciles is also 10% .the fifth deciles divide the series in to two equal parts and hence it is the same as median.

The variate values of the variable, say X, which divide the series or frequency in to one hundred equal parts, are called percentiles for the corresponding frequency distribution of X. There are in all ninety nine percentiles namely $P_1, P_2, \ldots\ldots P_{99}$. P_1 is called the lower percentile and P_{99} is called the upper percentiles. The area between any two percentiles, before $P_1$and after P_{99} is 0.01 i.e. 1%. Fifty percentiles divide the series in to two equal parts and hence it is the same as median.

Calculation of Partition values

For Individual Series

Let $x_1, x_2, \ldots\ldots x_n$ be n observation arranged in their ascending order of magnitude, then the quartiles, deciles and percentiles can be computed as

$$Q_i = \text{Value of } i\left(\frac{(n+1)}{4}\right)^{th} \text{ items, where } i = 1, 2, 3.$$

$$D_i = \text{Value of } i\left(\frac{(n+1)}{10}\right)^{th} \text{ items, where } i = 1, 2, \ldots..9.$$

$$P_i = \text{Value of } i\left(\frac{(n+1)}{100}\right)^{th} \text{ items, where } i = 1, 2, \ldots..99.$$

For discrete series

Let $x_1, x_2, \ldots\ldots x_n$ be n arranged data with corresponding frequency $f_1, f_2, \ldots\ldots f_n$ then partition values are calculated as,

$$Q_i = \text{Value of } i\left(\frac{(N+1)}{4}\right)^{th} \text{ items where } i = 1, 2, 3.$$

$$D_i = \text{Value of } i\left(\frac{(N+1)}{10}\right)^{th} \text{ items where } i = 1, 2, \ldots..9.$$

$$P_i = \text{Value of } i\left(\frac{(N+1)}{100}\right)^{th} \text{ items where } i = 1, 2, \ldots..99.$$

Where N = Σf is total frequency

For Continuous Series

In case of continuous data calculation of partition values is same as in case of median. But here we first find corresponding partition class and partition values is calculated as follows.

$$Q_i = L + \frac{\left(\frac{iN}{4} - cf\right)}{f} \times h, \text{ where } i = 1, 2, 3.$$

$$D_i = L + \frac{\left(\frac{iN}{10} - cf\right)}{f} \times h, \text{ where } i = 1, 2, \ldots, 9.$$

$$P_i = L + \frac{\left(\frac{iN}{10} - cf\right)}{f} \times h, \text{ where } i = 1, 2, \ldots \ldots 99.$$

Where L is lower limit of the class in which the partition value lies.

cf is the cumulative frequency preceding to the class in which the partition values lie.

h is size of class or class width.

f is the frequency of the class in which the partition value lies.

4.3 RELATIONSHIP BETWEEN MEAN, MEDIAN AND MODE

For symmetrical distribution arithmetic mean, median and mode coincide i.e., Mean = Median = Mode.

For moderately skewed distribution, mean, median and mode follow the following empirical relationship.

Mean – Mode = 3(Mean – Median)

Mode = 3 Median – 2Mean

4.4 CONDITIONS OF COMPUTING MEASURES OF CENTRAL TENDENCY

For Median and Mode

i. The distribution should consist of exclusive classes.

ii. The distribution should be of equal classes.

iii. The distribution should consist of simple frequency (without cumulative frequency).

For Mean

i. The distribution should consist of close end classes.

ii. The distribution should consist of simple frequency (without cumulative frequency).

Note: For quartiles, deciles, and percentiles the conditions should be the same as that of median and mode.

4.5 CHOICE OF AN AVERAGE

It is obvious that no single average is suitable for all practical problems from the discussion of various measures of averages. However, the selection of averages depends upon the nature of distribution, purpose of inquiry etc.

The following are the situations for the selection of an average.

4.5.1 Arithmetic Mean

Arithmetic mean is recommended when,

i. Data is quantitative in nature.

ii. Distribution is less skewed.

iii. The distribution consists of close end classes.

4.5.2 Median

Median is suggested when

i. Data is qualitative in nature.

ii. Data is highly skewed i.e., there exists extreme observation.

iii. The distribution consists of open–end classes.

4.5.3Mode

Mode is used while dealing with business viz. consumer preference on the certain goods to know the most common item among the items.

4.5.4 Geometric Mean

Geometric mean is used to compute average rates, ratios and percentage etc.

4.5.5 Harmonic Mean

Harmonic mean is used in computing the average relating to rates and ratios, where time factor is the variable.

Among the above averages, arithmetic mean is the commonly used measure of central tendency.

Example 16

Calculate Q_1, D_7, and P_{85} for the following data

Value	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	8	10	15	20	15	10	8

Solution

Calculation of partition values is as follows

Values(X)	frequency	Cumulative frequency
0 – 10	8	8
10 – 20	10	18
20 – 30	15	33
30 – 40	20	53
40 – 50	15	68
50 – 60	10	78
60 – 70	8	86

Here $N = \Sigma f = 86$, total frequency

Therefore, $Q_1 = \text{Value of } \left(\frac{N}{4}\right)^{th} = \text{items } \frac{86}{4} = 21.5^{th}$ item,

Then first quarter lies in class 20 – 30

$$Q_1 = L + \frac{\left(\frac{iN}{4} - cf\right)}{f} \times h$$

Where, L = 20, h = 10, cf = 18, and f = 15

$$\therefore Q_1 = 20 + \frac{(21.5 - 18)}{15} \times 10 = 22.34$$

Now, $D_7 = \text{Value of } 7\left(\frac{N}{10}\right)^{th} \text{ items} = 7 \times \frac{86}{10} = 60.2^{th}$ Item

Then seventh decile lies in class 40 – 50

$$D_7 = L + \frac{\left(\frac{7N}{10} - cf\right)}{f} \times h$$

Where, L = 40, h = 10, cf = 53, and f = 15

$$D_7 = 40 + \frac{(60.2 - 52)}{15} \times 10 = 44.8$$

Again, for $P_{85} = \text{Value of } 85\left(\frac{N}{100}\right)^{th} \text{ items} = 85 \times 86/100 = 73.1^{th}$ item

Then eighty fifth percentiles lies in class 50 – 60

$$P_{85} = \frac{\left(\frac{85N}{100} - cf\right)}{f} \times h$$

Where, L = 50, h = 10, cf = 68, and f = 10

$$\therefore P_{85} = 50 + \frac{(73.1 - 68)}{10} \times 10 = 55.1$$

EXERCISE - 4

1. What are the requisites of a good measure of Central tendency?
2. What are the types of Averages? Define simple mean and weighted mean?
3. What do you understand by median and what are the steps to calculate the median in continuous data.
4. What is mode and when we use the grouping method to calculate the mode.
5. The following are 15 measurement of the boiling point of a silicon compound(in degree Celsius) : 166, 141, 136, 153, 170, 162, 155, 146, 183, 157, 148, 132, 160, 175 and 150 calculate the average boiling point
6. Following are the marks obtained by 12 B.E. students in probability and statistics: 90, 91, 82, 63, 41, 38, 64, 74, 82, 70, 71, and 62. Find the average marks of students.
7. In a frequency distribution, the mid values of the first two classes are 10 and 20 and each class has equal class intervals. The cumulative frequencies of the classes are 3, 18, 37, 43, 48 and 62 give the original frequency distribution and fined the value of arithmetic mean.
8. The following table gives the basic salaries of persons in a company. Calculate the mean basic salaries of the persons.

salary(Rs)	3500	4400	5100	7500	9000	10380
No. of persons	5	7	10	15	13	16

9. Find the missing frequency from the data given below when arithmetic mean of the data is 27.

Mid values	5	15	25	35	45
frequency	5	8	?	16	6

cf is the cumulative frequency preceding to the class in which the partition values lie.

h is size of class or class width.

f is the frequency of the class in which the partition value lies.

Example 16

Calculate Q_1, D_7, and P_{85} for the following data

Value	0-10	10-20	20-30	30-40	40-50	50-60	60-70
frequency	8	10	15	20	15	10	8

Solution

Calculation of partition values is as follows

Values(X)	frequency	Cumulative frequency
0 – 10	8	8
10 – 20	10	18
20 – 30	15	33
30 – 40	20	53
40 – 50	15	68
50 – 60	10	78
60 – 70	8	86

Here $N = \Sigma f = 86$, total frequency

Therefore, Q_1 = Value of $\left(\frac{N}{4}\right)^{th}$ = items $\frac{86}{4} = 21.5^{th}$ item,

Then first quarter lies in class 20 – 30

$$Q_1 = L + \frac{\left(\frac{iN}{4} - cf\right)}{f} \times h \qquad \text{where, } L = 20, h = 10, cf = 18, \text{ and } f = 15$$

$$\therefore Q_1 = 20 + \frac{(21.5 - 18)}{15} \times 10 = 22.34$$

Now, D_7 = Value of $7\left(\frac{N}{10}\right)^{th}$ items = $7 \times \frac{86}{10} = 60.2^{th}$ Item

Then seventh decile lies in class 40 – 50

$$D_7 = L + \frac{\left(\frac{7N}{10} - cf\right)}{f} \times h \qquad \text{; where, } L = 40, h = 10, cf = 53, \text{ and } f = 15$$

$$\therefore D_7 = 40 + \frac{(60.2 - 52)}{15} \times 10 = 44.8$$

Again, for P_{85} = Value of $85\left(\frac{N}{100}\right)^{th}$ items = $85 \times 86/100 = 73.1^{th}$ item

Then eighty fifth percentiles lies in class 50 – 60

$$P_{85} = \frac{\left(\frac{85N}{100} - cf\right)}{f} \times h \qquad \text{; where, } L = 50, h = 10, cf = 68, \text{ and } f = 10$$

$$\therefore P_{85} = 50 + \frac{(73.1 - 68)}{10} \times 10 = 55.1$$

Exercise - 5

1. What are the requisites of a good measure of Central tendency?
2. What are the types of Averages? Define simple mean and weighted mean?
3. What do you understand by median and what are the steps to calculate the median in continuous data.
4. What is mode and when we use the grouping method to calculate the mode.
5. The following are 15 measurement of the boiling point of a silicon compound(in degree Celsius): 166, 141, 136, 153, 170, 162, 155, 146, 183, 157, 148, 132, 160, 175 and 150 calculate the average boiling point
6. Following are the marks obtained by 12 B.E. students in probability and statistics: 90, 91, 82, 63, 41, 38, 64, 74, 82, 70, 71, and 62. Find the average marks of students.
7. In a frequency distribution, the mid values of the first two classes are 10 and 20 and each class has equal class intervals. The cumulative frequencies of the classes are 3, 18, 37, 43, 48 and 62 give the original frequency distribution and fined the value of arithmetic mean.
8. The following table gives the basic salaries of persons in a company. Calculate the mean basic salaries of the persons.

salary(Rs)	3500	4400	5100	7500	9000	10380
No. of persons	5	7	10	15	13	16

9. Find the missing frequency from the data given below when arithmetic mean of the data is 27.

Mid values	5	15	25	35	45
frequency	5	8	?	16	6

10. Following data represent the wage ('00 Rs.) received by workers in a construction site. Find the average salaries received by the workers.

wage	25-30	30-35	35-40	40-45	45-50	50-55	55-60	60-65	65-70

workers	10	13	18	21	24	28	20	11	8

11. An enquiry into budgets of the middle class families in a Kathmandu gave the following data. Compute the weighted arithmetic mean.

Expense on	Food	Fuel	Clothing	Rent	Others
Percentage	35	10	20	15	20
Index no.	96.67	92	86.67	100	112.5

12. Find the average of the following data by (i) direct method, (ii) short cut method, (iii) step deviation method

expenditure(Rs.)	No. of Families
200 – 300	10
300 – 400	15
400 – 500	25
500 – 600	16
600 – 700	14

13. An examination was held to decide about the award of a scholarship in P. U. the weight of various subjects were different. The marks obtained by 4 candidates (out of 100 in each subject) were given below.

Subject:	Weight	Marks			
		A	B	C	D
Probability	5	70	80	85	92
Physics	3	90	75	75	55
Math	3	50	60	55	70
English	1	50	45	65	50
Chemistry	2	70	50	55	60

Who gets the scholarship and why?

14. From the information given below (i) which company pays larger amount as daily wage. (ii) What is the average daily wage for the workers of the two companies?

	Company A	Company B
No. Of workers	250	200
Average daily wage	20	25

15. The mean salary of 100 workers was found to be Rs. 40. Later on it was discovered that a salary 53 was misread as 83. Find the correct mean corresponding to the correct salary. What will be the mean salary when the wrong salary is omitted?

16. An experiment shows the height of 51 plants given below.

Height (cm)	10	20	30	40	50	60

No. of plant	2	3	–	21	–	5

If average heights of all the 51 plants are 40 cm find the missing frequencies corresponding to the height 30 and 50 cm.

17. The following are the numbers of twists that were required to break 12 togged bars : 33, 24, 39, 48, 26, 35, 38, 54, 23, 34, 29 and 37

Find (i) mean and (ii) median

18. Calculate the median and mode of the following distribution.

X	20-30	30-40	40-50	50-60	60-70
f	7	10	20	18	7

19. Calculate the appropriate measure of central tendencies from the following distribution and support your choice.

Wage in Rs.	No. of workers
Below 1000	50
1000-1999	500
2000-2999	555
3000-3999	100
4000-4999	30
5000 and above	15

20. Dr. Singh is the director of the student s financial aid office at a college. He has used available data on the summer earnings of all students who have applied to his office for financial aid. The frequency distribution is given below.

summer earnings (Rs.)	no. of students
0-500	231
500-1000	304
1000-1500	400
1500-2000	296
2000-2500	123
2500-3000	68
3000 and above	23

(i) Find the modal values of Dr. Singh's data

(ii) If students aids is restricted to those whose summer earnings were at least model summer earnings, how many of the applicants qualified ?

21. The expenditure of 1000 families is given as below.

Expenditure	40-59	60-79	80-99	100-119	120-139

No. of families	50 - 500 - 50

The median of the frequency distribution is Rs. 87. Calculate the missing frequencies.

22. Calculate the median, third deciles and 20th percentile from the following data.

Central value	2.5	7.5	12.5	17.5	22.5
frequency	7	18	25	30	20

23. From the following distribution of marks of 500 students of a college, find the minimum pass marks if only 20% of the students had failed and also find the minimum marks obtained by the top 25% of the students.

Marks	0-20	20-40	40-50	50-60	60-80	80-100
No. of students	50	100	150	90	60	50

24. Find the mode of the following data by the method of grouping.

X	1	2	3	4	5	6	7	8	9	10
f	6	9	4	2	10	8	7	5	1	3

25. In a moderately asymmetrical distribution the values of mean and median are 20 and 24 respectively. Find the value of mode.

Answers:

5. 146.81, **6.** 69. **7.** 32.46

8. 7498.182 **9.** 15 **10.** 4740

11. 97.87 **12.** Rs.461.25

13. D, largest weighted mean **14.** (i) Same, (ii) Rs. 22.22

15. 39.7, 39.57 **16.** 9 and 11

17. Mean = 35, Md = 34.5 **18.** Median 47.5, mode = 45.5

19. Since first and last classes are open-end, Md =2134.635

20. (i) 1240 (ii) 727 **21.** 263 and 137

22. Md = 15, D_3= 11 and P_{20} = 8.61

23. P_{20} = 30 and P_{75} = 58 **24.** Mode =6

25. Mode = 32, Use $(\bar{x} - M_o) = 3(\bar{x} - M_a)$

Measure of Dispersion

Introduction

Experience tells that in many situations, the spread of values is different but their central values are same for example.

								Mean	Median	Mode
Data A	30	31	32	32	32	33	34	32	32	32
Data B	5	15	23	32	32	32	85	32	32	32

Therefore, a central value provides no information about the scattering of values in a set of data, hence certain measures are evolved which reflect on the scattering of values in numerical terms are known as measure of dispersion.

Thus dispersion is defined as the scatter-ness of the items from the central values of data. It is also called the measurement of variability in the given set of data.

The main **objectives** of measuring variability or dispersion are as follows.

i. To determined the reliability of calculated averages.
ii. It makes possible to compare two series of data in respect of their variability.
iii. Measure of dispersion provides the basis for the quality control of produced items in industry.
iv. It has a wide application in almost all fields of statistics.

Requisites of a good measure of dispersion

The main characteristics of an ideal measure of dispersion can be given as follows.

i. It should be based on all observations.
ii. It should be rigidly defined.
iii. It should be simple to understand and easy to calculate
iv. Its unit should be same as the unity of measurement of items.
v. It should be least affected by fluctuation of sampling.
vi. It should not be affected by extreme values.

Absolute and Relative Measure of Dispersion

If the unit of a measure of dispersion is in same terms as that of the observations of a series, it is called absolute measure of dispersion, e.g. height in meters, weights in kilogram, income in rupees etc. In this case two series having different units of dispersion can't be compared. Hence for comparison of two series having different unit of measurements, one requires a unitless measure of dispersion, such measure are termed as relative measure or coefficient of dispersion.

a unitless measure of dispersion, such measure are termed as relative measure or coefficient of dispersion.

5.2 METHODS OF MEASURE OF DISPERSION

Following are the different measure of dispersion; in central tendencies we discusses the absolute measure only but here we discuss the relative measure also known as coefficient.

i. Range
ii. Quartiles Deviation
iii. Mean Deviation
iv. Standard deviation
v. Box Plot
vi. Lorenz curve

5.2.1 Range

It is simplest and easiest method of studying the variation in the set of observations. It is defined as the difference between the largest and smallest values of a set of data. If 'L' is largest and 'S' is the smallest value in the data then,

Range = L – S.

In a grouped frequency distribution, the range is the difference between the upper limit of the last class and the lower limit of the smallest class of ascending ordered data.

And relative measure of range called coefficient of Range is defined as Coefficient of Range = $\frac{L - S}{L + S}$, it is a unit less quantity. Lesser the coefficient of range better is the data.

Range has a special application in the quality control measures, it is a very useful measure to study fluctuation of data variation in the weather forecasts, stock market fluctuation etc.

Merit of Range

i. It is the easiest measure of dispersion.
ii. It is simple to understand and easy to calculate.
iii. Only minimum time is required to know the variability with the help of range.

Demerit of Range

i. It depends on two extreme values of a series. Thus it gives no information about the observations lying between smallest and largest values.
ii. It is highly susceptible to sampling fluctuations

iii. It is not suitable for further statistical treatment.
iv. Addition or deletion of a single value may change the entire complex of range.

Example 1

Find the Range and coefficient of range of the following data of marks of 5 students as: 63, 71, 55, 45, 81.

Solution

Here largest values of variable L = 81 and smallest value of variable is S = 45

Therefore Range = L-S = 81 – 45 = 36 and

Coefficient of range = $\frac{L - S}{L + S} = \frac{81 - 45}{81 + 45} = 0.286$

Example 2

Find the Range and coefficient of range of the following data.

X	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
f	2	3	5	4	2

Solution

Here largest values of variable L = 60 and smallest value of variable is S = 10

Therefore, Range = L-S = 60 – 10 = 50 and

Coefficient of range = $\frac{L - S}{L + S} = \frac{60 - 10}{60 + 10} = 0.714$

5.2.2 Quartile Deviation

The difference between the upper and lower quartiles is called inter quartile range; symbolically it is equal to $(Q_3 - Q_1)$. Half the inter quartiles range is known as semi- inter quartile range or quartile deviation (QD) and given as;

$$QD = \frac{Q_3 - Q_1}{2} = \frac{(Q_3 - M_d) + (M_d - Q_1)}{2}$$

So it gives the average amount by which the two quartiles differ from the median, since in a distribution 25% of observations will fall below Q_1 and 25% of observations will lie above Q_3. So 50% of the observation will lie between Q_1 and Q_3 Less the QD, more will be the uniformity or less will be the variability in the central 50% of the items and. again greater the QD. Less will be the uniformity or greater will be the variability in the central 50% of the

item. It is an absolute measure of dispersion. The relative measure of dispersion based on quartiles also known as coefficient of QD is defined as.

$$\text{Coefficient of QD} = \frac{(Q_3 - Q_1)/2}{(Q_3 + Q_1)/2} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It is a unit less quantity. Lesser the coefficient of QD better is the data.

Merits of Quartile deviation

i. It is easy to calculate and understand.
ii. It can be calculated in case of open end frequency distribution as well.
iii. It is not affected by 25% upper and 25% lower extreme values.

Demerits of Quartile deviation

i. It is based only on the central 50% observation and ignores the other 50%.
ii. It is not applicable to further algebraic treatment.
iii. It is not based on all the observations.
iv. It is affected by fluctuation of sampling.

Example 3

Find the semi- inter quartile range and coefficient of Quartile deviation from the following data.

Class Interval	0 – 10	10 – 20	20 – 30	30 – 40
frequency	1	4	3	2

Solution

We have semi- inter quartile range is $= \frac{Q_3 - Q_1}{2}$

And, coefficient of quartile deviation is $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

Calculation of quartile deviation and its coefficient

Class interval	Frequency	Cumulative frequency
0 – 10	1	1
10 – 20	4	5
20 – 30	3	8
30 – 40	2	10
Total	10	

For Q_1, values of N/4 = 2.5 therefore 10-20 is the first quartile class then

$$Q_1 = L + \frac{\left(\frac{N}{4} - cf\right)}{f} \times h;\ \text{where, } L = 10,\ h = 10,\ cf = 1,\ \text{and } f = 4$$

$$Q_1 = 10 + \frac{2.5 - 1}{4} \times 10 = 13.75$$

For Q_3, values of 3N/4 = 7.5 therefore 20-30 is the third quartile class then

$$Q_3 = L + \frac{\left(\frac{3N}{4}\right) - cf}{f} \times h;\ \text{where, } L = 10,\ h = 20,\ cf = 5,\ \text{and } f = 3$$

$$Q_3 = 20 + \frac{7.5 - 5}{3} \times 10 = 28.34$$

Therefore the semi inter quartile range is $= \frac{Q_3 - Q_1}{2}$

$$= \frac{28.34 - 13.75}{2} = 7.295$$

And the coefficient of Quartile deviation is $= \frac{Q_3 - Q_1}{Q_3 + Q_1}$

$$= \frac{28.34 - 13.75}{28.34 + 13.75}$$

$$= 0.3466 = 34.66\%$$

5.2.3 Mean Deviation

Range and QD are positional measures of dispersion, whereas mean deviation is a measure of dispersion which is based on all the values of set of data. It is defined as the average of the absolute deviation taken from an average, usually, the mean, median or mode. It is denoted by MD. TO clarify whether the average used in mean deviation is mean, median or mode, a suffix is attached to MD such as $MD_{\bar{x}}$, MD_{md}, MD_{Mo}

Calculation of Mean deviation

For individual data

Let $x_1, x_2, \ldots x_n$ be n observation. Then mean deviation is defined as

$MD = \frac{\Sigma|X - A|}{n}$, Here A may be any chosen constant out of mean, median and mode.

For discrete and continuous data

Let $x_1, x_2, \ldots x_n$ be n observation with corresponding frequency $f_1, f_2, \ldots, f_n$. Then mean deviation is defined as $MD = \frac{\Sigma f|X - A|}{N}$,

Here N = Σf is total frequency.

For continuous data X is the mid- point of classes. Here, it is worth emphasizing that mean deviation is minimum about the median. That is why median is commonly used as an average value about which the mean deviation is calculated.

Mean deviation has the same unit of measurement as that of the variable X. if two series have different units of measurement, the series can't be compared. Hence for comparing any two series, a unitless measure is given known as coefficient of mean deviation defined as

$$\text{Coefficient of MD} = \frac{\text{Mean deviation}}{\text{average used}}$$

Example 4

Find the mean deviation about mean and its coefficient from the following data.

Class	10-20	20-30	30-40	40-50	50-60	60-70	70-80
f	18	30	40	55	38	20	16

Solution

We have the formula for mean deviation about mean is

$$MD_{\bar{x}} = \frac{\sum f|x - \bar{x}|}{N} \text{ and its coefficient is } = \frac{MD_{\bar{x}}}{\bar{x}}$$

Calculation of MD

C.I.	f	Mid value(x)	f x	$x - \bar{x}$	$\lvert x - \bar{x} \rvert$	$f\lvert x - \bar{x} \rvert$
10-20	18	15	270	-28.71	28.71	516.77
20-30	30	25	750	-18.71	18.71	561.29
30-40	40	35	1400	-8.71	8.71	348.39
40-50	55	45	2475	1.29	1.29	70.97
50-60	38	55	2090	11.29	11.29	429.03
60-70	20	65	1300	21.29	21.29	425.81
70-80	16	75	1200	31.29	31.29	500.65
Total	217		9485			2852.90

Therefore, mean $\bar{x} = \frac{\sum fx}{N} = \frac{9485}{217} = 43.71$

And mean deviation about mean = $MD_{\bar{x}} = \frac{\Sigma f|x - \bar{x}|}{N} = \frac{2852.9}{217} = 13.15$

And the coefficient of mean deviation taken about mean is

$$= \frac{MD_{\bar{x}}}{\bar{x}} = \frac{13.15}{43.71} = 0.3007$$

Merits of Mean Deviation

i. It utilizes all the observations of the series.
ii. It is simple to calculate and understand
iii. It is least affected by extreme values.

Demerits of Mean deviation

i. The foremost weakness of mean deviation is that in its calculation negative differences are considered as positive without any sound reasoning.
ii. It is not an amenable to further statistical treatment.
iii. It can't be calculated in case of open end frequency distribution if mean is used as average.

5.2.4 Standard Deviation

Standard deviation is the absolute measure of dispersion. It fulfills all the requisites of a good measure of dispersion except that it is sensitive to extreme values. That is why it is known as standard deviation. It is defined as the positive square root of the mean of the square of the deviations taken from mean ($\bar{x}$). It is denoted by σ. It is also known as mean error, mean square error or root mean square deviation from mean.

Calculation of standard deviation
For individual series

Let $\bar{x}$ be the mean of 'n' observations $x_1, x_2, \ldots x_n$. Then their standard deviation is given by

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$, which is standard deviation by direct method. When variate values are large, we use short cut method as,

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$, Where $d = x - a$, a is assumed mean

For discrete frequency distribution

For a frequency distribution of $x_1, x_2, \ldots x_n$ with corresponding frequency $f_1, f_2, \ldots, f_n$, then the standard deviation is defined as

$$\sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{N}} = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$$, which is standard deviation by direct method.

And for short cut method $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$, $d = x - a$

For Continuous frequency distribution

Formula for calculation of standard deviation in case of continuous frequency distribution is same as in case of discrete frequency distribution but only difference is that 'x' is mid-value of the classes. Also in case of step deviation method standard deviation is defined as

$$\sigma = \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2} \times h$$, Where $d' = \frac{x - a}{h}$, h is the class width.

Properties of standard deviation

The following are the most important properties of standard deviation.

i. Standard deviation is the least possible root mean square deviation i.e. $\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$ is minimum.

Proof, σ is minimum if $\Sigma(x - \bar{x})^2$ is minimum and which is minimum since the sum of the square of the deviation of the items is minimum when taken from mean (Properties of mean)

ii. Standard deviation is not less than the mean deviation from mean; the relation between SD, MD and QD is as follows

4SD = 5 MD = 6 QD.

iii. **Combined Standard deviation:** The combined Standard deviation of two or more groups can be computed on the basis of their individual Standard deviation. If σ_1 and σ_2 are Standard deviation of two groups with size n_1 and n_2 and means $\bar{x}_1$ and $\bar{x}_2$. Then combined Standard deviation of two groups taken together is denoted by σ_{12} and given as $\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$

Where $d_1 = (\bar{x}_1 - \bar{x}_{12})$, $d_2 = (\bar{x}_2 - \bar{x}_{12})$ and $\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$ is combined mean. This formula can be extended for three or more groups.

iv. Standard deviation is independent of change of origin but not of scale.

v. If the Standard deviation is calculated from the sample then the formula is

$$\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n-1}}$$

Example 5

Find the standard deviation of the following data: 20, 18, 32, 41, 48, 61 and 70

Solution

X	20	18	32	41	48	61	71
X^2	400	324	1024	1681	2304	3721	5041

We have the calculating formula of standard deviation is

$$\sigma = \sqrt{\frac{1}{n}\Sigma(x - \bar{x})^2} = \sqrt{\frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2}$$

$$= \sqrt{\frac{14495}{7} - \left(\frac{291}{7}\right)^2} = 18.507$$

Example 6

Find the standard deviation of the data of the number of printing mistake per pages as:

No. of error per page	0	1	2	3	4	5	6	7
frequency	112	90	80	50	49	48	30	25

Solution: We have the formula of standard deviation for discrete data is

$$\sigma = \sqrt{\frac{\Sigma f(x - \bar{x})^2}{N}} = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma x}{N}\right)^2}$$

X	f	f X	fX^2
0	112	0	0
1	90	90	90
2	80	160	320
3	50	150	450
4	49	196	784
5	48	240	1200
6	30	180	1080
7	25	175	1225
Total	**484**	**1191**	**5149**

Therefore, the value of standard deviation is

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

$$= \sqrt{\frac{5149}{484} - \left(\frac{1911}{484}\right)^2} = 2.14 \text{ mistake per pages}$$

Example 7

Calculate the mean and standard deviation for the following data by direct and step deviation method.

Marks	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
frequency	7	12	24	10	7

Solution

Let, assumed mean a = 25 and common factor h = 10 and also

$$d' = \frac{x - a}{h}$$

Calculation of mean and standard deviation

C.I.	f	Mid value(x)	f x	f x^2	d	f d	f d^2
0-10	7	5	35	175	-2	-14	28
10-20	12	15	180	2700	-1	-12	12
20-30	24	25	600	15000	0	0	0
30-40	10	35	350	12250	1	10	10
40-50	7	45	315	14175	2	14	28
Total	**60**	**125**	**1480**	**44300**	**0**	**-2**	**78**

i. $\therefore$ Mean $\bar{x} = \frac{\Sigma fx}{N} = \frac{1480}{60} = 24.67$ and

Standard deviation $\sigma = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$

$$= \sqrt{\frac{44300}{60} - \left(\frac{1480}{60}\right)^2} = 11.40$$

ii. For step deviation method mean $\bar{x} = a + \frac{\Sigma fd'}{N} \times h$

$$= 25 + \frac{-2}{60} \times 10 = 24.67$$

And standard deviation = $h \times \sqrt{\frac{\Sigma fd'^2}{N} - \left(\frac{\Sigma fd'}{N}\right)^2}$

$$= 10 \times \sqrt{\frac{78}{60} - \left(\frac{-2}{60}\right)^2}$$

$$= 11.40$$

Example 8

The mean and standard deviation of two distributions of 100 and 150 items are 50, 5 and 40, 6 respectively. Find the standard deviation of all the 250 items taken together.

Solution

Here;

Sample size $n_1 = 100$ $n_2 = 150$

Mean $\bar{x}_1 = 50$ $\bar{x}_2 = 40$

SD $\sigma_1 = 5$ $\sigma_2 = 6$

$$\therefore \text{Combined mean } \bar{x}_{12} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{100 \times 50 + 150 \times 40}{100 + 150} = 44$$

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$$= \sqrt{\frac{100(5^2 + 6^2) + 150(6^2 + (-4)^2)}{100 + 150}} = 7.46$$

Here $d_1 = 50 - 44 = 6$ and $d_2 = 40 - 44 = (-4)$

Merit of standard deviation

i. It is based on all items of the series.
ii. It is rigidly defined.
iii. It is least affected by fluctuation of sampling.
iv. It is suitable for further statistical treatments.
v. It is suitable for comparing variability.

Demerit of standard deviation

i. It can't be calculated for open end classes.
ii. It gives greater weight to the extreme values and less to those which are near the mean.
iii. It is difficult to compute and understand for non mathematical person.

5.3 VARIANCE

Variance is nothing but just square of the standard deviation i.e. Variance = σ^2. It possesses all the requisites of a good measure of dispersion except that its unit is square of the unit of measurement of variate values. Hence many times it becomes difficult to actually adjudge the magnitude of variation. Also variance is sensitive to extreme values. Variance is the back bone of statistics.

i.e. $\sigma^2 = \frac{\Sigma f(x - \bar{x})^2}{N} = \frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2$

5.4 COEFFICIENT OF VARIATION

The relative measure of dispersion based on the standard deviation is defined as the ratio of the standard deviation and the mean and is called coefficient of standard deviation.

$$\text{Coefficient of SD} = \frac{SD}{\text{mean}} = \frac{\sigma}{\bar{x}}$$

When coefficient of SD is expressed in percentage is called coefficient of variation, it is denoted by C.V. and given as:

$$\text{C.V.} = \frac{\sigma}{\bar{x}} \times 100\ \%$$

It is independent of unit of measurement. So two distributions can batter compared, with the help of C.V. for their variability.

Less the C.V. more will be the uniformity, consistency, homogeneity, equitable etc. and more the C.V. less will be the uniformity, consistency, homogeneity etc.

Example 9

Calculate the Coefficient of variation for the following data of profits.

Profit	5-10	10-15	15-20	20-25	25-30	30-35	35-40	40-45
Freq.	9	17	23	30	22	16	9	4

Solution

Profit	Mid value(X)	f	fx	fx^2
5 – 10	7.5	9	67.5	506.25
10 – 15	12.5	17	212.5	2656.25
15 – 20	17.5	23	402.5	7043.75
20 – 25	22.5	30	675	15187.5
25 – 30	27.5	22	605	16637.5
30 – 35	32.5	16	520	16900
35 – 40	37.5	9	337.5	12656.25
40 – 45	42.5	4	170	7225
Total	**200**	**130**	**2990**	**78812.5**

Here $N = \Sigma f = 130$, $\Sigma fx = 2990$ and $\Sigma fx^2 = 78812.5$

Then $\bar{x} = \frac{\Sigma fx}{N} = \frac{2990}{130} = 23$

$$\therefore \sigma = \sqrt{\frac{\Sigma fx^2}{N} - \left(\frac{\Sigma fx}{N}\right)^2}$$

$$= \sqrt{\frac{78812.5}{130} - \left(\frac{2990}{130}\right)^2}$$

$$= 8.79$$

Then the C.V. $= \frac{\sigma}{\bar{x}} \times 100\ \%$

$$= \frac{8.79}{23} \times 100\%$$

$$= 38.214\%$$

5.5 FIVE NUMBER SUMMARY

The five number summaries is a descriptive statistics that provides information about a set of observations. It consists of the five most important descriptive measures.

i. **The minimum**: this is the smallest value of the set of data.

ii. **The first quartile (lower quartile)** : this number is denoted by Q_1 and 25% of the data falls below the Q_1.

iii. **The median**: this is the midway point of the data 50% of data falls below the median.

iv. **The third quartile (upper quartile)** : this number is denoted by Q_3 and 75% of the data falls below the Q_3.

v. **The maximum**: this is the largest value of the set of data.

In order for these statistics to exit the observations must be from a univariate variable that can be measured on an ordinal, interval or ratio scale. It is possible to quickly compare several sets of observations by comparing their five number summaries, which can be represented graphically using Box and whisker plots.

The five number summary gives information about the location (from the median), dispersion (from the quartiles) and range (from the minimum and maximum value) of observations. The five number summary also gives information about the shape of the distribution.

a. The distribution is left skewed (Negatively Skewed) if

i. The distance from smallest value to the median is greater than the distance from the median to largest value .

ii. The distance from smallest value to Q_1 is greater than the distance from Q_3 to the largest value.

iii. The distance from Q_1 to the median is greater than the distance from median to Q_3.

b. The distribution is right skewed (Positively Skewed) if

i. The distance from smallest value to the median is less than the distance from the median to largest value.

ii. The distance from smallest value to Q_1 is less than the distance from Q_3 to largest value.

iii. The distance from Q_1 to median is less than the distance from median to Q_3.

c. The distribution is symmetrical if

Both distances (which are stated above) are the same.

Example 10

Calculate the five number summary for the following set of observations: 15, 10, 5, 20, 35, 30, 25, 40, 45. Also comment about the shape of the distribution.

Solution

Arranging the data into ascending order of magnitude, 5, 10, 15, 20, 25, 30, 35, 40, 45.

Here, Total number of data (N) = 9

Calculation of Q_1 (lower quartile):

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

$$= \text{value of } \left(\frac{9+1}{4}\right)^{th} \text{ item}$$

$$= \text{value of 2.5th item}$$

$$= \text{2nd term} + 0.5\,(\text{3rd term} - \text{2nd term})$$

$$= 10 + 0.5(15 - 10)$$

$$= 12.5$$

Calculation of median:

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

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

$$= \text{value of 5th item}$$

$$= 25$$

Calculation of Q_3 (upper quartile):

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

$$= \text{value of of } \left\{\frac{3(9+1)}{4}\right\}^{th} \text{ item}$$

$$= \text{value of 7.5th item}$$

$$= \text{7th term} + 0.5\,(\text{8th term} - \text{7th term})$$

$$= 35 + 0.5(40 - 35)$$

$$= 37.5$$

Smallest value = 5

Largest value = 45

Hence,

Minimum = 5, $Q_1 = 12.5$, $Q_2 = 25$, $Q_3 = 37.5$ and maximum = 45

Now,

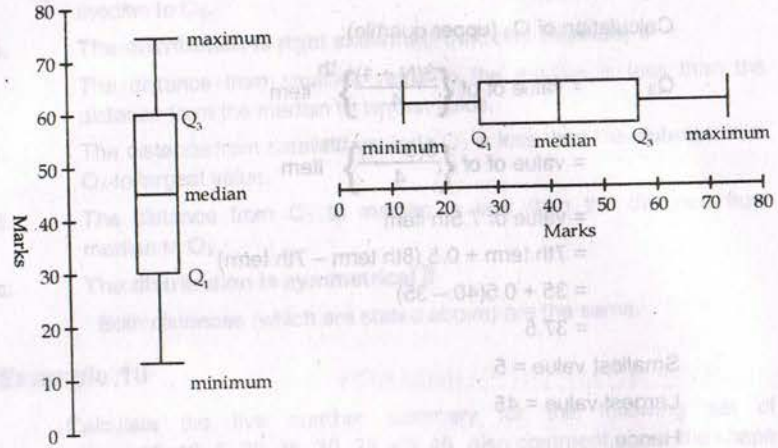
The distance from smallest value to median = (25 – 5) = 20

The distance from median to largest value = (45 – 25) = 20

Here, both the distances are the same so, the distribution is symmetric.

5.6 BOX-AND-WHISKER PLOTS

Box-and-whisker plot is graphical representation of five number summaries which consists of the smallest and largest value, the median and the first and third quartile of given data set. Immediate visual of a box-and-whisker plot are the centre, the spread and the overall range of the distribution. The bottom line (or the vertical line of the left side) of the box represents first quartile and the top line (or the vertical line at the right side) of the box represents third quartile. The line inside the box represents the median. Lines extending vertically or horizontally from the boxes indicating variability outside the upper and lower quartiles called whiskers. Whiskers represent lower 25% of the data and upper 25% of the data. The 50% data lies in the middle box. The following are figures of box-and-whisker plots.



Example 11

The following set of numbers is the marks obtained by fifteen students in statistics.

27, 18, 59, 54, 82, 91, 87, 100, 85, 34, 52, 61, 78, 68,

Solution,

First of all arranging the data into ascending order of magnitude,

18, 27, 34, 52, 54, 59, 61, 68, 78, 82, 85, 87, 91, 93, 100

Here,

Number of data (N) = 15

Calculation of Q_1 (lower quartile):

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

$$= \text{value of } \left(\frac{15+1}{4}\right)^{th} \text{ item}$$

$$= \text{value of } 4^{th} \text{ item}$$

$$= 52$$

Calculation of median,

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

$$= \text{value of } \left(\frac{15+1}{2}\right)^{th} \text{ item}$$

$$= \text{value of } 8^{th} \text{ item}$$

$$= 68$$

Calculation of Q_3

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

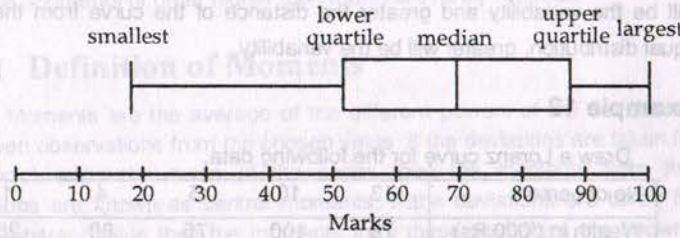
$$= \text{value of } \left\{\frac{3(15+1)}{4}\right\}^{th} \text{ item}$$

$$= \text{value of } 12^{th} \text{ item} = 87$$

Smallest value = 18

Largest value = 100

Hence the required box plot is as



5.6.1 Outlier

The box plot is used to find out the outliers. Outliers are the values of variables which are exceptionally extreme values. We use the following formula to find out the outlier.

First of all calculate the value of Q_1 and Q_3, then the inter quartile Range (IQR)is calculate as

$IQR = Q_3 - Q_1$

If the Values of variables lies below $Q_1 - 1.5(IQR)$ and the above the $Q_3 + 1.5(IQR)$ then these values are considered as the exceptionally extreme and known as outlier

5.7 LORENZ CURVE

It is the graphical method of studying variations in the distribution. This method was for the first time used by Max. O. Lorenz for measuring the variations in the distribution of incomes and wealth. It is a special types of cumulative frequency curve which is used to portray the data to indicate whether a factor is equally distributed in relation to the other factor for certain segment of the population.

Following steps are involved while drawing a Lorenz curve:

i. The variate and the frequencies are both cumulated. Express these cumulated variate values and frequencies in terms of percentages by taking the totals for each as 100.

ii. Paired cumulative percentages are plotted on a graph paper choosing same scale along axes from 0 to 100 by keeping frequency cumulative percentage in x-axis and cumulative value percentage in y-axis.

iii. Plotted points are jointed by a smooth free hand curve. This curve always starts from the origin and terminated at the end point (100,100).

Less the distance of the curve from the line of equal distribution, less will be the variability and greater the distance of the curve from the line of equal distribution, greater will be the variability.

Example 12

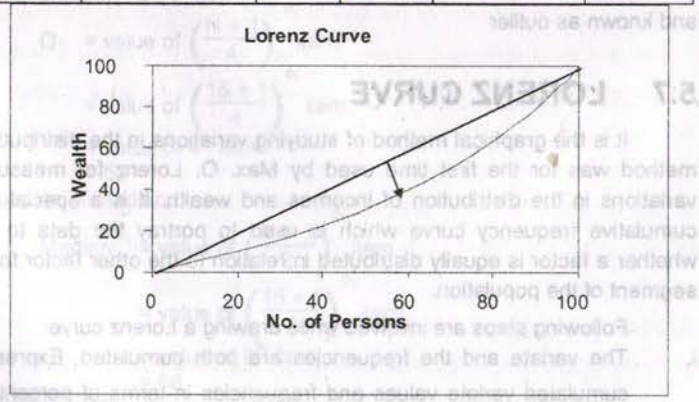
Draw a Lorenz curve for the following data.

No of person	13	10	5	4	1
Wealth in ('000 Rs)	78	100	75	80	25

Solution

To calculation the Variation of distribution by Lorenz's curve first we prepare table and make the graph of Lorenz's curve as

Persons	Cum. Persons	%	wealth	Cum. wealth	%
13	13	39.4	78	78	21.8
10	23	69.7	100	178	49.7
5	28	84.8	75	253	70.7
4	32	97.0	80	333	93.0
1	33	100.0	25	358	100.0



5.8 MOMENTS

Moments are useful in order to describe the frequency distribution. Each frequency distribution is specified by its moments especially the first and second moments. They also help in determining the shape of distribution. Since skewness and kurtosis are usually measured with the help of moments. Mean and variance are nothing but first and second moments respectively. In practice, moments of order higher than fourth are rarely required.

5.8.1 Definition of Moments

Moments are the average of the different powers of the deviations of the given observations from the chosen value. If the deviations are taken from the exact arithmetic mean of the series then the moments from these deviations are known as central moments, If the deviations are taken from some arbitrary value then the moments from these deviations are known as raw moments.

5.8.2 Central Moments

For ungrouped data, the moments about the A.M is denoted by μ_r and defined as

$$\mu_r = \frac{1}{n} \sum (x - \bar{x})^r, \; r = 1, 2 \ldots\ldots\ldots$$

In particular, the first four central moments are,

$$\mu_1 = \frac{1}{n} \sum (x - \bar{x}) = 0, \text{ since } \sum (x - \bar{x}) = 0$$

$$\mu_2 = \frac{1}{n} \sum (x - \bar{x})^2, \quad \mu_3 = \frac{1}{n} \sum (x - \bar{x})^3, \; \mu_4 = \frac{1}{n} \sum (x - \bar{x})^4, \text{ where}$$

n is total number of observations.

For frequency distribution, the first four central moments are

$$\mu_1 = \frac{1}{N} \sum f(x - \bar{x}) = 0, \text{ since } \sum f(x - \bar{x}) = 0$$

$$\mu_2 = \frac{1}{N} \sum f(x - \bar{x})^2; \quad \mu_3 = \frac{1}{N} \sum f(x - \bar{x})^3;$$

$$\mu_4 = \frac{1}{N} \Sigma f(x - \bar{x})^4; \quad \text{where } N = \Sigma f \text{ is total frequency.}$$

Note that: μ_1 is always zero and μ_2 is variance (σ^2)

5.5.4 Raw Moment

For ungrouped data, the r^{th} moment about any arbitrary value "a" is denoted by μ_r and defined as

$$\mu_r = \frac{1}{n} \Sigma (x - a)^r, \; r = 1, 2 \ldots\ldots\ldots$$

In particular, the first four raw moments are,

$$\mu_1 = \frac{1}{n} \Sigma (x - a)$$

$$\mu_2 = \frac{1}{n} \Sigma (x - a)^2, \quad \mu_3 = \frac{1}{n} \Sigma (x - a)^3 \quad , \quad \mu_4 = \frac{1}{n} \Sigma (x - a)^4$$

where n is total number of observations.

For frequency distribution, the first four raw moments are

$$\mu_1' = \frac{1}{N} \Sigma f(x - a) = \frac{1}{N} \Sigma fU \qquad \mu_2' = \frac{1}{N} \Sigma f(x - a)^2 = \frac{1}{N} \Sigma fU^2$$

$$\mu_3' = \frac{1}{N} \Sigma (x - a)^3 = \frac{1}{N} \Sigma fU^3 \qquad \mu_4' = \frac{1}{N} \Sigma f(x - a)^4 = \frac{1}{N} \Sigma fU^4$$

where $U = X - a$

For continuous data, if $U = \frac{X - a}{h}$ then first four raw moments are,

$$\mu_1' = h \frac{1}{N} \Sigma fU' \qquad \mu_2' = h^2 \frac{1}{N} \Sigma fU'^2$$

$$\mu_3' = h^3 \frac{1}{N} \Sigma fU'^3 \qquad \mu_4' = h^4 \frac{1}{N} \Sigma fU'^4$$

Note: here μ_1' is used to calculate the value of mean as

$\mu_1' = \bar{X} + a$ then $\bar{X} = a + \mu_1'$

5.8.5 Relation between Raw Moments and Central Moments

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

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

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

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

5.8.6 Measure of Skewness and Kurtosis Based on Moments

Coefficient of skewness is defined as $\beta_1 = \frac{\mu_3^2}{\mu_3}$ and β_1 is always show positive skewness which is not true. In order to over come this limitation, Karl Pearson gave another coefficient of skewness based on moments as

$$_1 = \sqrt{\beta_1} = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}}$$

If $_1 = 0$, then the distribution is symmetrical.

If $_1$ = positive then the distribution is positively skewed or right tailed distribution.

If $_1$ = negative then the distribution is negatively skewed or left tailed distribution.

Similarly, the coefficient of kurtosis based on moments are defined as,

$\beta_2 = \frac{\mu_4}{\mu_2^2}$ and $\gamma_2 = \beta_2 - 3$

If $\beta_2 = 3$ i.e. $\gamma_2 = 0$, then the distribution is normal or Mesokurtic.

If $\beta_2 > 3$ i.e. $\gamma_2 > 0$, then the distribution is Leptokurtic.

If $\beta_2 > 3$ i.e. $\gamma_2 < 0$, then the distribution is Platykurtic.

Some workout Examples

Example 13

The running capacity of two horses is given below. Which horse is more consistent, why?

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

Solution

Let X = running capacity of horse A

Y = running capacity of horse B

Let, A = 290 and B = 290 are assumed mean for two series

For horse A let , $d = X - A = X - 290$, $d' \; \frac{d}{h} \; \frac{d}{5}$

For horse B let, $d = Y - B = Y - 290$

Calculation of coefficient of variation

Horse A				Horse B		
X	d	d'	d'^2	Y	d	d^2
250	-40	-8	64	280	-10	100
255	-35	-7	49	282	-8	64
280	-10	-2	4	290	0	0
290	0	0	0	295	5	25
295	5	1	1	298	8	64
300	10	2	4	295	5	25
Total		$\Sigma d' = -14$	$\Sigma d'^2 = 122$		$\Sigma d = 0$	$\Sigma d^2 = 278$

For horse A,

$$\bar{X} = A + \frac{\Sigma d'}{n} \times h = 290 + \frac{(14)}{6} \times 5$$

$$= 290 - 11.67$$

$$= 278.33$$

$$\sigma = \sqrt{\frac{\Sigma d'^2}{n} - \left(\frac{\Sigma d'}{n}\right)^2} = \sqrt{\frac{122}{6} - \left(\frac{-14}{6}\right)^2} \times 5$$

$$= \sqrt{20.33 - 5.43} \times 5$$

$$= \sqrt{14.9} \times 5$$

$$= 19.3$$

Coefficient of variation $(CV_A) = \frac{\sigma}{\bar{X}} \times 100\ \%$

$$= \frac{19.3}{278.33} \times 100\% = 6.93\%$$

For horse B,

$$\bar{Y} = B + \frac{\Sigma d}{n} = 290 + \frac{0}{6} = 290$$

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

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

$$= 6.81$$

Acme Library 01917

And coefficient of variation $(CV_B) = \frac{\sigma}{\bar{X}} \times 100\%$

$$= \frac{6.81}{290} \times 100\% = 2.35\%$$

From the above calculation, the CV of running capacity of horse B is less than CV of horse A. Hence, we conclude that horse B is more consistent than horse A regarding its running capacity.

Example 14

Students' age in the BE Computer program and BE Civil program of ABC University are described by two samples. If the homogeneity in age of the class is a positive factor in learning, make suggestion, with reason, which of the two groups will be easier to teach?

BE Computer	Age	23	29	28	22	30	21	25	26	27	24
	No of students	9	2	5	10	1	4	11	6	3	9
BE Civil	Age	27	31	30	29	28	33	34	35	36	32
	No of students	10	8	5	4	6	5	5	11	2	4

Solution

We have to calculate the coefficient of variation (CV) of age for each program to compare the homogeneity of age of students.

For BE Computer program

Let, X = age of students, A = 25, d = X-A = X - 25

Calculation of mean and standard deviation

X	f	d	fd	fd^2
23	9	-2	-18	36
29	2	4	8	32
28	5	3	15	45
22	10	-3	-30	90
30	1	5	5	25
21	4	-4	-16	64
25	11	0	0	0
26	6	1	6	6
27	3	2	6	12
24	9	-1	-9	9
Total	N=60		$\Sigma fd = -33$	$\Sigma fd^2 = 319$

Now, arithmetic mean, $\bar{X} = A + \frac{\Sigma fd}{N}$

$$= 25 + \frac{(-33)}{60} = 24.45$$

Standard deviation, $\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{319}{60} - \left(\frac{-33}{60}\right)^2}$

$= 2.24$

Coefficient of variation (CV) $= \frac{\sigma}{\bar{X}} \times 100\%$

$= \frac{2.24}{24.45} \times 100\%$

$= 9.16\%$

For BE Civil program

Let, A = 31, d =X -A=X - 31

Calculation of mean and standard deviation

X	f	d	fd	fd^2
27	10	-4	-40	160
31	8	0	0	0
30	5	-1	-5	5
29	4	-2	-8	16
28	6	-3	-18	54
33	5	2	10	20
34	5	3	15	45
35	11	4	44	176
36	2	5	10	50
32	4	1	4	4
Total	N = 60		$\Sigma fd = 12$	$\Sigma fd^2 = 530$

Now, $\bar{X} = A + \frac{\Sigma fd}{N} = 31 + \frac{12}{60}$

$= 31.2$

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{\frac{530}{60} - \left(\frac{12}{60}\right)^2} = 2.97$$

Then, C.V. $= \frac{\sigma}{\bar{X}} \times 100\% = \frac{2.97}{31.2} \times 100\% = 9.52\%$

Since the C.V. of age of students of BE Computer is less than C.V. of BE Civil program, it is easier to teach Computer program than Civil program.

Example 15

Compute the combined standard deviation for the following numerical information.

Number of administrative staff = 75

Number of teaching staff = 50

Average monthly salary of administrative staff = Rs 4500

Average monthly salary of teaching staff = Rs 6000

Standard deviation of the salary of administrative staff = Rs 800

Standard deviation of the salary of teaching staff = Rs 900

Solution

Given,

Administrative staff	Teaching staff
$n_1 = 75$	$n_2 = 50$
$\bar{X}_1 = \text{Rs } 4500$	$\bar{X}_2 = \text{Rs } 60$
$\sigma_1 = \text{Rs } 800$	$\sigma_2 = \text{Rs } 900$

According to question we have to find out Combined standard deviation, σ_{12}

Now, combined mean, $\bar{X}_{12} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$

$= \frac{75 \times 4500 + 50 \times 6000}{75 + 50}$

$= \text{Rs } 5100$

Let, $d_1 = \bar{X}_1 - \bar{X}_{12} = 4500 - 5100 = -600$

$d_2 = \bar{X}_2 - \bar{X}_{12} = 6000 - 5100 = 900$

So, combined standard deviation is

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

$$= \sqrt{\frac{75 \times (800)^2 + 50 \times (900)^2 + 75 \times (-600)^2 + 50 \times (900)^2}{75 + 50}}$$

$$= \sqrt{\frac{75 \times 640000 + 50 \times 810000 + 75 \times 360000 + 50 \times 810000}{125}}$$

= 1117.14

Therefore, required combined standard deviation is Rs 1117.14.

Example 16

The information regarding the incomes of two villages A and B are given below.

	Village A	Village B
Number of people	600	500
Average income (Rs)	175	186
Variance of income	100	81

a. What is the total income of both villages put together?

b. In which village is the variation of income greater?

c. What is the average income of people in A and B villages put together?

d. What is the combined standard deviation?

Solution

Given,

Village A		Village B	
n_1	= 600	n_2	= 500
$\overline{X}_1$	= Rs175	$\overline{X}_2$	= Rs186
σ_1^2	=100	σ_2^2	= 81

a. We have the relation $\overline{X}_1 = \frac{\sum X}{n_1}$

For village A, total income, $\Sigma X = n_1 \overline{X}_1$ = 600×175 = Rs 10,5000

For village B, total income, $\Sigma X = n_2 \overline{X}_2$ = 500×186 = Rs 93,000

Therefore, the total income of both villages put together

= 105000 + 93000 = Rs 198000

b. For village A,

$$CV_A = \frac{\sigma_1}{\overline{X}_1} \times 100\% = \frac{10}{175} \times 100\%$$

= 5.71%

For village B,

$$CV_B = \frac{\sigma_2}{\overline{X}_1} \times 100\ \% = \frac{9}{186} \times 100\ \%$$

= 4.84 %

As, $CV_A > CV_B$ village A has the greater variation in income.

c. Average income of both villages combined together is

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

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

= Rs 180

d. Here, $d_1 = \overline{X}_1 - \overline{X}_{12}$ = 175 - 180 = -5

$d_2 = \overline{X}_2 - \overline{X}_{12}$ = 186 - 180 = 6

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

$$= \sqrt{\frac{600\times100 + 500\times81 + 600\times(-5)^2 + 500\times(6)^2}{600+500}}$$

$= \sqrt{121.36}$

= 11.02

Therefore, required combined standard deviation of income is Rs 11.02.

Measure of Skewness and Kurtosis based on moments

Coefficient of skewness is defined as $\beta_1 = \frac{\mu_3^2}{\mu_3}$ and β_1 is always show positive skewness which is not true. In order to over come this limitation, Karl Pearson gave another coefficient of skewness based on moments as

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

If $\gamma_1 = 0$, then the distribution is symmetrical.

If γ_1 = positive then the distribution is positively skewed or right tailed distribution.

If γ_1 = negative then the distribution is negatively skewed or left tailed distribution.

Similarly, the coefficient of kurtosis based on moments are defined as,

$\beta_2 = \frac{\mu_4}{\mu_2^2}$ and $\gamma_2 = \beta_2 - 3$

If $\beta_2 = 3$ i.e. $\gamma_2 = 0$, then the distribution is normal or Mesokurtic.

If $\beta_2 > 3$ i.e. $\gamma_2 > 0$, then the distribution is Leptokurtic.

If $\beta_2 > 3$ i.e. $\gamma_2 < 0$, then the distribution is Platykurtic.

Exercise 6

1. What do you mean by dispersion and what are the requisites of good measure of dispersion?
2. Define range and its relative measure.
3. What do you understand by Quartile deviation and its coefficient?
4. Define standard deviation and coefficient of variation
5. What is the relation between variance and standard deviation?
6. Find the range of the data :22, 25, 64, 82, 91, 76, 18 and 30
7. Find the range and the coefficient of range from the following data:

Temp(^{0}C)	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
No. of days	5	12	25	18	2

8. Calculate the quartile deviation and the coefficient of quartile deviation from the following data of marks of students as:

Marks	10 – 19	20 – 29	30 – 39	40 – 49	50 – 59
frequency	2	6	22	13	7

9. Find the mean and standard deviation of the following miles per gallon obtained in 20 test runs performed on Urban roads with an intermediate car:

19.7	21.9	22.8	22	21.5	20.5	23.2	23	22.5	19.3
21.4	21.1	22.2	19.9	20.8	20.9	22.6	21.7	19.4	21.3

10. Following data represent the breaking strength (in ounces) of a sample of 60 linen threads. Find the Mean deviation about mean, median and mode

Classes	15 – 19.9	20 – 24.9	25 – 29.9	30 – 34.9	35 – 39.9
frequency	3	15	24	12	6

11. Calculate the mean deviation about mean from the following data.

X	2 – 4	4 – 6	6 – 8	8 – 10
Y	20	40	30	10

12. Find the mean and standard deviation from the following data of the sulfur oxides emission.

Sulfur oxides	4.95-8.95	8.95-12.95	12.95-16.95	16.95-20.95	20.95-24.95	24.95-28.95	28.95-32.95
frequency	3	10	14	25	17	9	2

13. The following are figures on a Well's daily production of oil in barrels: 214, 203, 226, 198, 243, 225, 207, 203, 208, 200, 217, 202, 208, 212, 205, and 220. Find the Mean, standard deviation and coefficient of variation.
14. In two companies A and B engaged in similar type of industry, the average weekly wage and standard deviation are given below;

	Company A	Company B
Average weekly wage (Rs.)	460	490
Standard deviation	50	40
No. of wage earners	100	80

a. Which company pays larger amount as weekly wages?
b. Which company show greater variability in the distribution weekly wages?
c. What is the mean and standard deviation of all the workers in two companies taken together?

15. Student's ages in the Bachelor of Engineering, Civil and Computer program of a Khwopa engineering college are described by these two samples data. If Homogeneity in age of the class is a positive factor in learning, suggest which of the two groups will be easier to teach

Computer Program	Civil Program
20	19
26	26
24	22
19	21
21	20
18	22
22	26
23	27
24	20

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

17. The combined mean and variance of salary of 250 workers of city A and B taken together are 560 and 5497 respectively. The mean and variance of the salary of 150 workers of city B are 500 and 81 respectively. Find the variable of salaries in city A

18. Scores of two golfers for 10 rounds were as follows.

Golfer A	74	75	78	78	72	77	79	78	81	76
Golfer B	86	84	80	88	89	85	86	82	82	79

Find which golfer may be considered to be a more consistent player

19. A purchasing agent obtained samples of incandescent lamps from two suppliers. He had the samples tested in his own laboratory for length of life with the following results.

Length of life (hours)	700 – 900	900 – 1100	1100 – 1300	1300 – 1500
Sample from Company A	10	16	26	8
Sample from Company B	3	42	12	3

i. Which company's bulb gives a higher average life?

ii. Which company's bulbs are more uniform?

20. From the prices of shares X and Y given below, state which share is more stable in value?

X	55	54	52	53	56	58	52	50	51	49
Y	108	107	105	105	106	107	104	103	104	101

21. From the data given below state which series is more variable.

Variable	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 70
Score A	10	18	32	40	22	18
Score B	18	22	40	32	29	10

22. The following are the sample data regarding the wealth (in millions of Rs.) of families in Kathmandu and Lalitpur. Show which city has greater inequality in wealth distribution.

wealth	Number of families	
	Kathmandu	Lalitpur
0 – 5	750	270
5 – 10	120	130
10 – 15	42	35
15 – 20	38	25
20 – 25	27	22
25 – 30	25	18

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

24. Suppose two groups of human males yield the following information

	Group A	Group B
Age	24 Years	15 years
Mean weight	145 lbs	80 lbs
variance	100 lbs	100 lbs

Find which group is more variable.

25. The running capacity of two horses is given below. State which is more consistent and why?

Horse A	280	255	280	290	295	300
Horse B	280	282	290	295	298	295

26. Cricketer A and B made the following runs in 10 innings. Who is more consistent player and who is a better run scorer.

Player A	42	17	83	59	72	76	64	45	40	32
Player B	28	70	31	10	59	108	82	14	2	95

27. Draw a Lorenz curve for the following data

Temp(^{0}C)	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50

No. of days	5	12	25	18	2

Answers:

6. Range = 73, and coefficient of range = 0.67
7. Range =50, coefficient of range = 1
8. QD = 6.86 and CQD = 0.18
9. Mean = 21.38 and S.D = 1.19
10. MD_{mean}=3.85, MD_{md}= 3.77 MD_{mo}= 3.913
11. 1.52
12. Mean = 18.85 and S.D = 5.51
13. Mean =211.9375, SD = 11.53 and C. V = 5.44%
14. (a) company A, (b) Company B, (c) Mean =473.31 and S.d = 48.19
15. C. V of Computer = 10.8 and C. V. of Civil = 12.2, Computer
16. 40.23, 11.82
17. Mean = 650 and variance = 21
18. C. V of A =3.2 & C. V of B =3.7; A is consistent
19. (i)Company A,(ii) Company B
20. Share B,
21. Score B
22. Lalitpur; $C.V_{Kath.}$ = 64.63 and $C.V._{Lalitpur}$ = 71.12
23. Mean = 40.23 and s.d =11.82
24. Group B, $C.V_A$=6.9 and $C.V_B$ =12.5
25. Horse B
26. Player A , Player A

Chapter 7

Correlation and Regression Analysis

Introduction

The various statistical methods discussed up to now have been concerned primarily with statistical summaries such as measure of central tendency, dispersion, Skewness, Kurtosis, which measures average variability and shape of data, which are usually sufficient when we have university data. But in practice, we may come across a number of problems consisting two or more variable. Distribution consisting two variables are said to be bi-variate distribution. In this chapter we discuss various method used to determine if there exist any relationship between two variables. As for example, quality of steel and life of bridge, quantity of bitumen and strength of road etc

There are two basic tools for summarizing bi-variate data correlation analysis summarizes the strength of relationship (if any) between two factors, while regression analysis shows how to predict of control one of the variables using the other one.

Correlation analysis is the statistical tool generally used to describe the degree to which one variable is linearly related to another. There are various phenomena which are related to each other. For instance, when quality of bitumen used in road reduces the condition of road becomes worse. When demand of certain commodities increases, then its price goes up and when its demand decreases then its pries comes down. The theory by means of which quantitative connection between two sets of phenomena are determined is called correlation theory. On basis of the theory of correlation one can study the comparative change occurring in two related phenomena and their cause effect analysis.

Types of correlation

The correlation can be classified in to three types.

a. Positive, negative and correlation.
b. Linear and non –linear correlation.
c. Simple. Partial and multiple correlations.

Positive, Negative and No Correlation

Two variables are said to be positively correlated if they tend to change together in the same direction, that is, if they tend to increase or decrease together. Such positive correlation is postulated by economic theory for the consumption and income. When the income increases the consumption increases, and conversely, when income decreases the consumption decreases. For example,

a.

Advertisement Cost (Rs)	10	9	8	7
Sales (Rs)	80	50	40	30

i.

Advertisement Cost (Rs)	10	9	8	7
Sales (Rs)	80	50	40	30

ii.

Income (Rs in '000)	10	12	15	19	20
Consumption (Rs in '000)	6	8	9	10	12

Two variables are said to be negatively correlated if they tend to change in the opposite direction: when X increases Y decreases, and vice versa. For example, the quantity of a commodity demanded and its price are negatively correlated. When price increases, demand for commodity decreases and when price falls demand increases. For example,

Demand (unit in '00)	10	9	8	6	4
Price (Rs in '000)	6	8	9	10	12

Two variables are uncorrelated when they tend to change with no connection to each other. For example one should expect zero correlation between the height of the inhabitants of a country and the production of steel.

The following table shows nature of correlation between two variables.

Change in one variable	Change in other variable	Nature of correlation
Increase (↑)	Increase (↑)	Positive (+)
Decrease (↓)	Decrease (↓)	Positive (+)
Increase (↑)	Decrease (↓)	Negative (-)
Decrease (↓)	Increase (↑)	Negative (-)

Linear and Non-linear Correlation

Correlation may be linear, when all points (X, Y) on a scatter diagram seem to cluster near a straight line, or nonlinear, when all points seem to lie near a curve. For example,

Linear correlation

X	4	5	6	7
Y	10	12	14	16

Non-linear correlation

X	4	5	6	7
Y	10	14	25	40

Simple, Partial and Multiple Correlations

In simple correlation, we measure the correlation between two variables (of which one is dependent and the other is independent). For example,

i. The correlation between amount of rainfall and production

ii. The correlation between number of unit sold and price

6.2 COVARIANCE

Before we study the correlation analysis we introduce the concept of covariance between two variables. Let $X_1, X_2, \ldots X_n$ be n number of observations of a random variable X and $Y_1, Y_2 \ldots Yn$ be the corresponding observations of another variable Y. Let the set of n pairs of observations be given by $(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)$. Then, the covariance between variables X and Y denoted by Cov (X, Y) is defined by

$$\text{Cov}(X, Y) = \frac{1}{n} \Sigma (X - \bar{X})(Y - \bar{Y})$$

$$= \frac{\Sigma XY}{n} - \frac{\Sigma X}{n} \frac{\Sigma Y}{n}$$

The value of covariance between two variables X and Y measures the simultaneous changes between two variables.

6.3 METHOD OF STUDYING CORRELATION

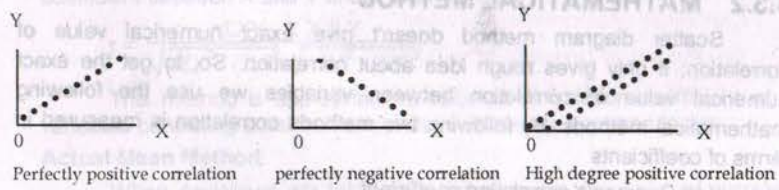
There are following three methods to study the correlation between two variables.

a. Graphical method or scatter diagram method

b. Mathematical Method

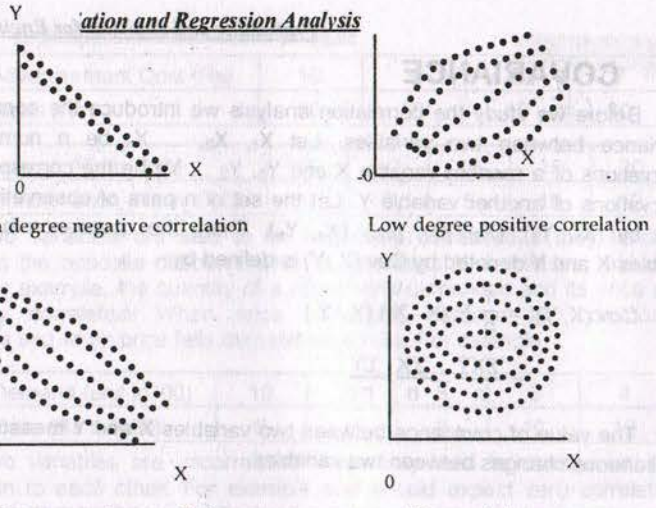
6.3.1 Scatter Diagram Method

The first step in determining whether there is a relationship between two variables is to examine the graph of the observed (or known) data. In this method, one variable's values are kept in X-axis and another variable's values are kept in Y-axis. The graph formed by plotting these pairs of values of X and Y is known as scatter diagram. If the plotted dots show some trend of upward or downward, then the two considering variables are said to be correlated. If the dots are close together and follow some trend of either increasing or decreasing then there is a strong relationship between them. The following scatter diagrams show different form of correlation.



Perfectly positive correlation perfectly negative correlation High degree positive correlation

Y



It is the simplest method of ascertaining the correlation between two variables and not influenced by the size of extreme items, but it gives only rough idea about the correlation and it doesn't give the exact extent to which the two variables are correlated.

Merits and demerits of scatter diagram

Merits

i. This is the simplest method to measure correlation.
ii. It is least affected by the size of the extreme values.
iii. Even a non-mathematical person can get idea about correlation.
iv. It establishes the base line for fitting the appropriate regression equation.

Demerits

i. This is graphical method. So, it does not give exact numerical value of correlation.

6.3.2 MATHEMATICAL METHOD

Scatter diagram method doesn't give exact numerical value of correlation; it only gives rough idea about correlation. So, to get the exact numerical value of correlation between variables we use the following mathematical methods. By following two methods correlation is measured in terms of coefficients

i. Karl Pearson's correlation coefficient.
ii. Spearman's Rank correlation coefficient.

6.4 KARL PEARSON'S CORRELATION COEFFICIENT

The Karl Pearson's correlation coefficient measures the degree of linear association between two variables. This method is popularly known as Pearsonian correlation coefficient. It is unitless measure of correlation. It is defined as the ratio of covariance between two variables to the product of the standard deviation of the two variables.

Let X and Y are two variables, then Karl Pearson's correlation coefficient between X and Y denoted by r_{xy} or r(x,y) or simply 'r' and is defined as

$$r = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$\text{Where, } Cov(X, Y) = \frac{1}{n}\Sigma(X - \bar{X})(Y - \bar{Y}) = \frac{\Sigma XY}{n} - \frac{\Sigma X}{n}\frac{\Sigma Y}{n}$$

$$Var(X) = \sigma_X^2 = \frac{\Sigma(X - \bar{X})^2}{n} = \frac{\Sigma X^2}{n} - \left(\frac{\Sigma X}{n}\right)^2$$

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

Where $\bar{X}$ = The Arithmetic mean of values of variables X and

$\bar{Y}$ = The Arithmetic mean of values of variables Y

and n = Number of pair of observations

6.4.1 Calculation of Karl Pearson's correlation coefficient

We use the following formulae to calculate Karl Pearson's correlation coefficient

Direct method

Let X and Y are two variables, then Karl Pearson's correlation coefficient between X and Y under this method is

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma x)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

This method is appropriate to calculate correlation coefficient of two variables consisting small numerical size without any common factor.

Actual Mean Method

When deviations are taken from actual mean, the value of correlation coefficient is defined by

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

Where, $x = (X - \bar{X})$ deviations taken from Arithmetic mean of X and

$y = (Y - \bar{Y})$ deviations taken from Arithmetic mean of Y

This formula particularly convenience if $\bar{x}$ and $\bar{y}$ comes out to be in whole numbers and known as product moment formula.

Short Cut Method and steps deviation method

In this method we use the following formula to calculate the correlation coefficient between two variables

$$r = \frac{n\Sigma uv - \Sigma u\,\Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

Where, n = number of pairs of observations

$u = X - A$ and $v = Y - B$

A = assumed mean of variable X

B = assumed mean of variable Y This method is appropriate to calculate correlation coefficient of two variables consisting large numerical size.

Also if $u' = \frac{X - A}{h}$ and $v' = \frac{Y - B}{k}$, h and k are common factor of series X and Y

$$r = \frac{n\Sigma u'v' - \Sigma u'\,\Sigma v'}{\sqrt{n\Sigma u'^2 - (\Sigma u')^2}\sqrt{n\Sigma v'^2 - (\Sigma v')^2}}$$

This method is appropriate when there is common factor in X series or Y series or both series.

6.4.2 Properties of Karl Pearson's correlation coefficient

Following are the important properties of Karl Pearson's correlation coefficient.

i. Correlation coefficient lies between – 1 and + 1 i.e. $-1 \le r \le +1$

ii. The formula of Correlation coefficient is symmetrical i.e. $r_{xy} = r_{yx}$.

iii. Correlation coefficient is geometric mean between two regression coefficient i.e. $r = \sqrt{b_{xy}\, b_{yx}}$

iv. Correlation coefficient is a relative statistical measure and has no unit.

v. Two independent variables are uncorrelated but the converse may not be true i.e. uncorrelated variables may not be independent.

vi. Correlation coefficient is independent of change of origin as well as of scale i.e. $r_{XY} = r_{uv}$.

6.4.3 Interpretation of calculated value of correlation coefficient

The calculate value of Karl Pearson's correlation coefficient may be interpreted in the following manner.

i. If r = 1, there is perfect positive correlation between the variables.

ii. If r = –1, there is perfect negative correlation between the variables.

iii. If r = 0, there is no correlation between the variables.

iv. if r lies between 0.700 to 0.999, there is high degree positive correlation.

v. If r lies between – 0.700 to – 0.999, there is high degree negative correlation.

vi. If r lies between, 0.500 to 0.699, there is moderate positive correlation.

vii. If r lies between, – 0.500 to – 0.699, there is moderate negative correlation.

viii. If r lies between, 0.001 to 0.499, there is low degree positive correlation.

ix. If r lies between – 0.001 to – 0.499, there is low degree negative correlation.

Example 1

If the covariance between the two variables X and Y is 6 and the standard deviation of X and Y are 2.45 and 2.61 respectively. Find the coefficient of correlation.

Solution

We have given Covariance (X, Y) = 6, $\sigma_x = 2.45$, $\sigma_y = 2.61$

Correlation coefficient (r) =?

Now, correlation coefficient between X and Y is given by

$$r = \frac{Cov(x, y)}{\sigma_x \sigma_y} = \frac{6}{2.45 \times 2.61} = 0.9383.$$

Thus, there is high degree positive correlation between X and Y.

Example 2

From the following data compute the Karl Pearson's coefficient of correlation between X and Y

	X series	Y series
No. of items:	6	6
sum of variable:	160	116
Sum of square of variable:	4480	2462

Sum of product of variable is 3275

Solution

Here, we are given in usual notation n= 6, $\Sigma X = 160$, $\Sigma Y = 116$, $\Sigma X^2 = 4480$, $\Sigma Y^2 = 2462$ and $\Sigma XY = 3275$

Now , Karl Pearson's correlation coefficient is given by

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma x)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{6 \times 3275 - 160 \times 116}{\sqrt{6 \times 4480 - (160)^2}\sqrt{6 \times 2462 - (116)^2}} = 0.84$$

Thus, the value of 'r' indicates that there is high degree positive correlation between two variables X and Y.

Example 3

Calculate the Karl Pearson's correlation coefficient between the data of driving speed and mileage of a car.

Driving speed	30	50	40	55	30	25	60	25	50	55
Mileage	28	25	25	23	30	32	21	35	26	25

Solution

Let, Driving speed = X and Mileage = Y

Calculation of Correlation coefficient

X	Y	X^2	Y^2	XY
30	28	900	784	840
50	25	2500	625	1250
40	25	1600	625	1000
55	23	3025	529	1265
30	30	900	900	900
25	32	625	1024	800
60	21	3600	441	1260
25	35	625	1225	875
50	26	2500	676	1300
55	25	3025	625	1375
420	**270**	**19300**	**7454**	**10865**

Here, n= 10, ΣX = 420, ΣY = 270, ΣX^2 = 19300, ΣY^2 = 7554 and ΣXY = 10865

Now, the value of Karl Pearson's correlation coefficient is given by

$$r = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma x)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$r = \frac{10 \times 10865 - 420 \times 270}{\sqrt{10 \times 19300 - (420)2}\sqrt{10 \times 7554 - (270)^2}} = -0.91$$

Thus, the value of 'r' indicates that there is high degree negative correlation between Driving speed and mileage.

Example 4

Calculate the Karl parson's correlation coefficient for the following data.

X	4	5	6	7	8
Y	9	11	12	15	18

Solution

Let $x = (X - \bar{X})$ and $y = (Y - \bar{Y})$

Table for the calculation of correlation coefficient

X	Y	x	y	x^2	y^2	xy
4	9	-2	-4	4	16	8
5	11	-1	-2	1	4	2
6	12	0	-1	0	1	0
7	15	1	2	1	4	2
8	18	2	5	4	25	10
30	**65**	**0**	**0**	**10**	**50**	**22**

Here, n = 5, ΣX = 30, ΣY = 65, Σx^2 = 10, Σy^2 = 50 and Σxy = 22

Then, means are as; $\bar{X} = \frac{\Sigma X}{n} = \frac{30}{5} = 6$ and $\bar{Y} = \frac{\Sigma Y}{n} = \frac{65}{5} = 13$

Therefore, the value of Karl Pearson's correlation coefficient is

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} = \frac{22}{\sqrt{10} \times \sqrt{50}} = 0.984$$

Therefore, there is high degree positive correlation between two variables X and Y.

Example 5

Find the correlation coefficient between age of marriage and number of women in a certain locality

Age of marriage	10 - 15	15 - 20	20 - 25	25 - 30	30 - 35	35 - 40
Number of women	10	22	46	25	5	2

Solution

Let, X- Age of marriage, Y –Number of women, A = 22.5 and B = 25

$u = \frac{X - 22.5}{5}$ and v = Y - 25

Calculation of Correlation coefficient

Age	X	Y	u	v	u^2	v^2	uv
10 – 15	12.5	10	– 2	– 15	4	225	30
15 – 20	17.5	22	– 1	– 3	1	9	3
20 – 25	22.5	46	0	21	0	441	0
25 – 30	27.5	25	1	0	1	0	0
30 – 35	32.5	5	2	– 20	4	400	– 40
35 – 40	37.5	2	3	– 23	9	529	– 69
Total			**3**	**– 40**	**19**	**1604**	**– 76**

Here,

$n = 6$, $\Sigma u = 3$, $\Sigma v = -40$, $\Sigma u^2 = 19$, $\Sigma v^2 = 1604$ and $\Sigma uv = -76$

Now, the value of Karl Pearson's correlation coefficient is given by

$$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u^2 - (\Sigma u)^2}\sqrt{n\Sigma v^2 - (\Sigma v)^2}}$$

$$r = \frac{6 \times (-76) - 3(-40)}{\sqrt{6 \times 19 - (3)^2}\sqrt{6 \times 1604 - (-40)^2}} = -0.366$$

Thus, The value of correlation coefficient indicates that there is low degree negative correlation between age of women and the responses in marriage.

6.4.4 Correlation for Bi-Variate Frequency Distribution

In bi-variate frequency distribution or in a two – way frequency table, the values of one variable are in rows and values of other variable are in columns, then values can be discrete or continuous. The frequencies in each class are kept in the body part of table. The theoretical frame of tow–way frequency table is as shown below.

X / Y	y_1	y_2	...	...	y_n	Total
x_1	f_{11}	f_{12}	...	...	f_{1n}	Σf_{1i}
x_2	f_{21}	f_{22}	...	...	f_{2n}	Σf_{2i}
...	...	...	...	...	...	...
x_n	f_{n1}	f_{n2}	...	...	f_{nn}	Σf_{ni}
Total	Σf_{j1}	Σf_{j2}	...	...	Σf_{ji}	N

Where N is the total frequency, in the table $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are the values of variables X and Y respectively and $f_{11}, f_{12} \ldots f_{nn}$ are their cell frequencies.

Then the value of Karl Pearson's correlation coefficient for bi-variate frequency distribution is given by

$$r = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

$$Cov(X, Y) = \frac{\Sigma fXY}{N} - \left(\frac{\Sigma fX}{N}\right)\left(\frac{\Sigma fY}{N}\right)$$

$$Var(X) = \sigma_X^2 = \frac{\Sigma fX^2}{N} - \left(\frac{\Sigma fX}{N}\right)^2 \text{ And } Var(Y) = \sigma_Y^2 = \frac{\Sigma fY^2}{N} - \left(\frac{\Sigma fY}{N}\right)^2$$

$$\therefore r = \frac{N\Sigma fXY - \Sigma fX \Sigma fY}{\sqrt{N\Sigma fX^2 - (\Sigma fX)^2}\sqrt{N\Sigma fY^2 - (\Sigma fY)^2}}$$

$$r = \frac{N\Sigma fuv - \Sigma fu \Sigma fv}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\sqrt{N\Sigma fv^2 - (\Sigma fv)^2}}$$

If $u = \frac{X - A}{h}$ and $v = \frac{Y - B}{K}$, where A and B are assumed mean of variables X and Y and h and k are common factor for variable X and Y.

Example 6

Calculate the correlation coefficient from the following bi- variate frequency distribution.

X \ Y	20	25	30	35
0	10	-	11	-
1	-	13	-	-
2	9	-	16	-
3	-	15	-	5

Solution

Calculation for correlation coefficient

X/Y	20	25	30	35	f	fX	fX^2	fXY
0	0		0					
	10	-	11	-	21	0	0	0
1		325						
	-	13	-	-	13	13	13	325
2	360		960					
	9	-	16	-	25	50	100	1320
3		1125		525				
	-	15	-	5	20	60	180	1650
f	19	28	27	5	**79**	**123**	**293**	**3295**
fY	380	700	810	175	**2065**			
fY^2	7600	17500	24300	6125	**55525**			
fXY	360	1450	960	525	**3295**			

Therefore, Karl Pearson's correlation coefficient is

$$r = \frac{N\Sigma fXY - \Sigma fX \Sigma fY}{\sqrt{N\Sigma fX^2 - (\Sigma fX)^2}\sqrt{N\Sigma fY^2 - (\Sigma fY)^2}}$$

$$r = \frac{79 \times 3295 - 123 \times 2065}{\sqrt{79 \times 293 - (123)^2}\sqrt{79 \times 55625 - (2065)^2}} = 0.195$$

Thus, $r = 0.195$ shows that there is low degree positive correlation between two variable X and Y.

Example 7

Following data gives the marks in Engineering mathematics and applied mechanics in B.E. final examination. Whether there is any association between the marks of two subjects

Marks in math	Marks in Applied mechanics				
	20 - 30	30 - 40	40 - 50	50 - 60	60 - 70
25 - 35	5	9	3	-	-
35 - 45	-	10	25	2	-
45 - 55	-	2	12	2	-
55 - 65	-	-	4	16	5
56 - 75	-	-	-	4	2

Solution

Let, X = Marks in Engineering mathematics, Y = Marks in applied mechanics and A = 40 and B = 45

Here $u = \frac{X - A}{h} = \frac{X - 40}{10}$ and $v = \frac{Y - B}{k} = \frac{Y - 45}{10}$

Mark in Math		Mark in applied Y	20-30	30-40	40-50	50-60	60-70				
			25	35	45	55	65				
	X	u \ v	-2	-1	0	1	2	f	fu	fu^2	fuv
25-35	30	-1	10, 5	9, 9	0, 3	-, -	-, -	17	-17	17	19
35-45	40	0	-, -	0, 10	0, 25	0, 2	-, -	37	0	0	0
45-55	50	1	-, -	-2, 2	0, 12	2, 2	-, -	16	16	16	0
55-65	60	2	-, -	-, -	0, 4	32, 16	20, 5	25	50	100	52
65-75	70	3	-, -	-, -	-, -	12, 4	12, 2	6	18	54	24
		f	5	21	44	24	7	**101**	**67**	**185**	**95**
		fv	-10	-21	0	24	14	**7**			
		fv^2	20	21	0	24	28	**93**			
		fuv	10	7	0	46	32	**95**			

Therefore, the value of Karl Pearson's correlation coefficient between marks in engineering mathematics and applied mechanics is

$$r = \frac{N\Sigma fuv - \Sigma fu\,\Sigma fv}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\sqrt{N\Sigma fv^2 - (\Sigma fv)^2}}$$

$$r = \frac{101 \times 95 - 67 \times 7}{\sqrt{101 \times 185 - (67)^2}\sqrt{101 \times 93 - (7)^2}} = 0.792$$

Thus, there is high degree positive correlation between marks in engineering mathematics and marks in applied mechanics.

6.4.5 Coefficient of Determination

The Square of the simple correlation coefficient is called the coefficient of determination. It is used for interpretation of the value of the calculated correlation coefficient.

If r is correlation coefficient then r^2 is coefficient of determination. If r = 0.8, the value of r^2 = 0.64. This shows that 64% of the total variation in dependent variable has been explained by the independent variable and other 36% is due to other unknown factors. 1 - r^2 is called coefficient of non-determination.

Example 8

If the correlation coefficient between two random variable is found to be 0.9. Find the coefficient of determination.

Solution

the correlation coefficient (r) = 0.9 therefore the coefficient of determination is $r^2 = (0.9)^2 = 0.81$ i.e. the total variation in the value of dependent variable is determined 81% by independent variable and remaining other 19% is due to other unknown variable.

6.4.6 Probable Error of Correlation Coefficient

Probable error of correlation coefficient is an old measure for testing the reliability or significance of an observed correlation coefficient. The probable error of correlation coefficient is abbreviated by PE(r). Let, r be the correlation coefficient calculated from n pair of sample observations, then the probable error of r is defined by

$$PE(r) = 0.6745\frac{1 - r^2}{\sqrt{n}}$$

It is used to interpret whether calculated value of r is significant or not.

i. If $|r| < PE(r)$, there is no evidence of correlation i.e. the correlation coefficient is not significant.

ii. If $|r| > 6\,PE(r)$, there is evidence of correlation i.e. correlation coefficient is significant.

iii. For other situations except above two cases, nothing can be stated.

Moreover, the probable error of a correlation coefficient enables us to find the limits within which the population correlation coefficient can be expected to lie. The limits are $r \pm PE(r)$.

Example 9

The correlation coefficient between the marks of 25 students in Statistics and Economics is found to be 0.86. Test the significance of the

calculated correlation coefficient. And calculate the limits of population correlation coefficient.

Solution

Here, sample size (n) = 25

Correlation coefficient (r) = 0.86

The probable error $PE(r) = 0.6745 \frac{1-(0.86)^2}{\sqrt{25}} = 0.0351$

Now, 6 PE(r) = 6 x 0.0351 = 0.2107

Here, $|r| = 0.86 \geq 6 \times PE = 0.2107$

Therefore, correlation coefficient is significant.

∴ The limits for population correlation coefficient (ρ)are

$r \pm PE(r) = 0.86 \pm 0.0351$

∴ Lower limit = 0.86 - 0.0351= 0.825 and

Upper limit = 0.86 + 0.0351 = 0.895

6.5 SPEARMAN'S RANK CORRELATION COEFFICIENT

Karl Pearson's correlation coefficient is especially useful when data are quantitatively measured; there are some variables like beauty, intelligence, knowledge, honesty etc, which can't be measured quantitative directly. These types of variables can be measured by assigning ranks or sorts of ratings. Ranks are the assignments of orders or priorities according to their status or importance, then the degree of association that exists between the two sets of ranks are known as rank correlation. There are some methods used to measure rank correlation one of these methods is Spearman's rank correlation coefficient. It is denoted by R

Let R_1 and R_2 be the ranks of individuals according to the attribute or characteristic A and B respectively. If we assume that not two individual have the same rank in a character then R_1 and R_2 takes numerical values from 1 to n, Now Spearman's rank correlation coefficient is given by following formula.

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Where n is numbers of observations and $d = R_1 - R_2$ is the difference of ranks.

It is to be noted that $-1 \leq R \leq 1$ And $\Sigma d = 0$.

We can calculate the Spearman's Rank correlation coefficient in following three cases.

Case I: When ranks are given

The formula for calculation of Spearman's rank correlation coefficient when the ranks are given is as

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Example 10

Two Housewives asked to express their preference for 8 different kinds of species gave the following replies. Find rank correlation coefficient.

Housewife	A	B	C	D	E	F	G	H
Prerana	2	1	4	3	7	8	6	5
Kusum	4	1	2	3	5	7	8	6

Solution

Calculation of rank correlation coefficient

Species	Rank by Prerana (R_1)	Rank by Kusum (R_2)	$d = R_1 - R_2$	d^2
A	2	4	-2	4
B	1	1	0	0
C	4	2	2	4
D	3	3	0	0
E	7	5	2	4
F	8	7	1	1
G	6	8	-2	4
H	5	6	-1	1
Total				**18**

Here n = 8 Σd^2= 18 and Spearman correlation coefficient R =?

$$\text{Now } R = 1 - \frac{6\Sigma d^2}{n(n^2-1)} = 1 - \frac{6 \times 18}{8(64-1)} = 0.785$$

Thus, Spearman rank correlation coefficient R indicates that there is a high degree association between preferences to spices of two housewives.

Case II: When ranks are not given

In this case, firstly it is necessary to rank the given observations either starting from the smallest value or from the largest value of both variables and then we use the following formula to calculate rank correlation coefficient.

$$R = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Example 11

Calculate the Spearman's Rank correlation coefficient between the Prices of potato and average annual rainfall in a region.

Year	1	2	3	4	5	6	7
Price	36	26	34	31	27	30	20
Rainfall	12	15	13	10	9	14	18

Solution

Let, Rank for price = R_1 and Rank for rainfall = R_2

Calculation of rank correlation coefficient

year	Price	R_1	Rainfall	R_2	$d = R_1 - R_2$	d^2
1	36	1	12	5	-4	16
2	26	6	15	2	4	16
3	34	2	13	4	-2	4
4	31	3	10	6	-3	9
5	27	5	9	7	-2	4
6	30	4	14	3	1	1
7	20	7	18	1	6	36
Total					**0**	**86**

Here, $n = 7$, $\Sigma d^2 = 86$ and Spearman's rank correlation coefficient $R = ?$

$$R = 1 - \frac{6\Sigma d^2}{n(n^2 - 1)} = 1 - \frac{6 \times 86}{7(49 - 1)} = -0.535$$

Thus there is negative correlation between price of potatoes and average annual rainfall

Case III: When ranks are repeated

For some cases, when more than one item has the same value within a series, a common rank is given to all such repeated items. Then common rank is the arithmetic mean of the ranks, which these items would have got if there were different from each other and the next item will get the rank next to them. For example, if fourth and fifth item have the same value, then the rank for each of them fourth and fifth item is 4.5 and the rank of next item is 6.

For this case, the general formula for calculating rank correlation coefficient is adjusted by adding a tie correlation factor $\frac{m(m^2 - 1)}{12}$ to Σd^2, and then the formula for calculating Spearman rank correlation coefficient, when more than one item has same value is as follows.

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{m_1(m_1^2 - 1)}{12} + \frac{m_2(m_2^2 - 1)}{12} + \ldots\right\}}{n(n^2 - 1)}$$

Where m_1, m_2, .. etc are the number of time that an item is repeated.

Merits of Spearman's Rank correlation Coefficient

i. It is appropriate for qualitative data.

ii. It is easy to compute.

iii. It can be used when the data are highly skewed.

Demerits of Spearman's Rank correlation Coefficient

i. It is not suitable to compute in case of large number of observations.

ii. It is not appropriate for calculating correlation coefficient of a frequency distribution.

iii. It does not consider original values i.e. it involves ranks which is not actual measure.

Example 12

From the marks obtained by 10 students of BE in statistics and Research methodology. Compute the rank correlation coefficient between marks of them

Roll No.	1	2	3	4	5	6	7	8	9	10
Statistics	52	50	55	50	60	70	72	78	80	63
Research	20	25	60	45	80	25	55	65	25	75

Solution

Let, R_1 and R_2 be ranks provide to marks in statistics and research methodology respectively and $d = R_1 - R_2$

Calculation of rank correlation coefficient

Roll no.	Statistics	R_1	Research	R_2	d	d^2
1	52	8	20	10	-2	4
2	50	9.5	25	8	1.5	2.25
3	55	7	60	4	3	9
4	50	9.5	45	6	3.5	12.25
5	60	6	80	1	5	25
6	70	4	25	8	-4	16
7	72	3	55	5	-2	4
8	78	2	65	3	-1	1
9	80	1	25	8	-7	49
10	63	5	75	2	3	9
Total					**0**	**131.5**

Here $n = 10$, $\Sigma d^2 = 131.5$ and Spearman's rank correlation coefficient $R = ?$

Here marks 50 in statistics is repeated 2 times therefore $m_1 = 2$ and the marks 25 in Research is repeated three times then $m_2 = 3$

Now Spearman rank correlation in repeated ranks is

$$R = 1 - \frac{6\left\{\Sigma d^2 + \frac{m_1(m_1^2 - 1)}{12} + \frac{m_2(m_2^2 - 1)}{12}\right\}}{n(n^2 - 1)}$$

$$= 1 - \frac{6\left[131.5 + \frac{2(4 - 1)}{12} + \frac{3(9 - 1)}{12}\right]}{10(100 - 1)}$$

$$= 0.188$$

Thus, there is low degree positive cor^n between marks in Statistics and Research methodology

6.6 PARTIAL AND MULTIPLE CORRELATION

So far we have discussed the theory of correlation between two variables only, but the value of a variable are often influence by several other variables, Now we shall discuses three or more than three correlated variables then there are two concepts one is partial correlation and other is multiple correlation.

6.6.1 Partial Correlation

In partial correlation we measure the correlation between a dependent variable and one particular independent variable keeping all other independent variables as constant. For example, the correlation between number of unit sold and advertising expenditure keeping price as constant.

The correlation between two variables (x_1, X_2) when the linear effect of the other variables (x_3, x_4, X_5, ...x_n) in them has been eliminated from both is called partial correlation between x_1 and x_2 .

It is often important to measure the correlation between a dependent variable and one particular independent variable when all other variables are kept constants; that is, when effects of all other variables are removed. This can be done by defining a coefficient of partial correlation. If we denote by "$r_{12.3}$" the coefficient of partial correlation between X_1 and X_2 keeping X_3 constant. We find that

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1- r_{13}^2)(1- r_{23}^2)}}$$, similarly

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1- r_{12}^2)(1- r_{32}^2)}}$$

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1- r_{21}^2)(1- r_{31}^2)}}$$

Similarly, $r_{12.34}$ is the coefficient of partial correlation between X_1 and X_2 keeping X_3 and X_4 constant then

$$r_{12.34} = \frac{r_{12.4} - r_{13.4}\, r_{23.4}}{\sqrt{(1- r_{13.4}^2)(1- r_{23.4}^2)}} = \frac{r_{12.3} - r_{14.3}\, r_{24.3}}{\sqrt{(1- r_{14.3}^2)(1- r_{24.3}^2)}}$$

The results are useful since by means of these formulae any partial correlation coefficient can ultimately be made to depend on the simple correlation coefficients r_{12}, r_{13} , r_{23} .

The Square of Partial correlation coefficient is known as the **coefficient of partial determination** and use to explain the total variation of dependent variable due to independent variable when the effect of other variable kept constant.

6.6.2 Multiple Correlations

Multiple correlation studies the relationship between a dependent variable and combine effect of two or more independent variables. For

example, the correlation between numbers of unit sold and joint effect of advertising expenditure and price.

In multiple correlations we study three or more than three variables at a time. Let X_1, X_2 and X_3 be three variables. Then the $R_{1.23}$ denoted the multiple correlation coefficient of the dependent variable X_1 on independent variables X_2 and X_3 is given as

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1- r_{23}^2}}$$, similarly

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21}r_{13}r_{23}}{1- r_{13}^2}}$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31}r_{32}r_{12}}{1- r_{12}^2}}$$

The Square of Multiple correlation coefficient is known as the **coefficient of multiple determination** and use to explain the total variation of dependent variable due to combine effect of independent variables

Example 13

The following correlation coefficients between three related variables X_1, $X_2$and $X_3$are given below

$r_{12} = 0.98$, $r_{13} = 0.44$, $r_{23} = 0.54$

Calculate the Three partial correlation coefficient and three multiple correlation coefficients

Solution

The three Partial correlation coefficients are calculated as follows

$$r_{12.3} = \frac{r_{12} - r_{13}\, r_{23}}{\sqrt{(1- r_{13}^2)(1- r_{23}^2)}} = \frac{0.98 - 0.44X0.54}{\sqrt{(1- 0.44^2)(1- 0.54^2)}} = 0.9822$$

$$r_{13.2} = \frac{r_{13} - r_{12}\, r_{32}}{\sqrt{(1- r_{12}^2)(1- r_{32}^2)}} = \frac{0.44 - 0.98X0.54}{\sqrt{(1- 0.98^2)(1- 0.54^2)}} = -0.53257$$

$$r_{23.1} = \frac{r_{23} - r_{21}\, r_{31}}{\sqrt{(1- r_{21}^2)(1- r_{31}^2)}} = \frac{0.54 - 0.98X0.44}{\sqrt{(1- 0.98^2)(1- 0.44^2)}} = 0.6088$$

The Three Multiple Correlation coefficients are calculated as follows

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1- r_{23}^2}}$$

$$= \sqrt{\frac{0.98^2 + 0.44^2 - 2X0.98X0.44X0.54}{1- 0.54^2}} = 0.9857$$

$$R_{2.13} = \sqrt{\frac{r_{21}^2 + r_{23}^2 - 2r_{21}r_{13}r_{23}}{1- r_{13}^2}}$$

$$= \sqrt{\frac{0.98^2 + 0.54^2 - 2X0.98X0.44X0.54}{1- 0.44^2}} = 0.9875$$

$$R_{3.12} = \sqrt{\frac{r_{31}^2 + r_{32}^2 - 2r_{31}r_{32}r_{12}}{1- r_{12}^2}}$$

$$= \sqrt{\frac{0.44^2 + 0.54^2 - 2X0.98X0.44X0.54}{1 - 0.98^2}} = 0.7018$$

6.7 REGRESSION ANALYSIS

In general decisions are based on the relationship between two or more variables. For example, after considering the relationship between quality of cement and strength of concrete an engineer might attempt to control the quality of cement production. An Economist might be interested to predict the demand of particular product using the relationship between price and the demand of the product. Sometimes a manager will rely on intuition to judge how two variables are related. However, if data can be obtained, a statistical procedure called regression analysis can be used to develop an equation showing how the variables are related.

Regression analysis helps one determine the probable form of the relationship between the variables. The objective of this method of analysis is usually to predict or estimate the value of one variable corresponding to a given value of another variable. The English scientist Sir Francis Galton (1822-1911) first proposed the ideas of regression in reports of his research in the area of heredity. Galton used the word regression to describe a tendency of adult offspring, even those with short or tall parents, to revert toward the average height of the general population.

In regression terminology, the variable being predicted is the called dependent variable. The variable or variables being used to predict the value of the dependent variable are called the independent variables. For example, in analyzing the effect of advertising expenditures on sales, a marketing manager's desire to predict sales would suggest making sales the dependent variable. Advertising expenditure would be independent variable used to predict sales. Economists might base their predictions of the annual gross domestic product, or GDP, on the final consumption spending within the economy. Thus, the final consumption spending is the independent variable and the GDP is the dependent variable. Generally, the dependent variable is denoted by Y and independent variable by X.

6.7.1 Linear or Non-Linear Regression

If the graph of dependent and independent variables shows a linear trend then it is called linear regression. But if it is not in a straight line then it is called non-linear or curvilinear regression.

6.7.2 A simple Probabilistic Model of Regression

Let, $Y = \beta_0 + \beta_1 X$ is the deterministic mathematical model of regression of dependent variable Y on independent variable X because a value of Y is determined by putting a value of X. where β_0 is Y-intercept of line and β_1 is the slope of line

But in reality for a given value of independent variable X, values of Y varies from actual values by a random amount, which is observed by the simple probabilistic model as

$Y = \beta_0 + \beta_1 X + e$

Where, e is an error terms, the difference between an observed value of Y and the mean value of Y. the error term is normally distributed with mean zero and variance σ^2.

6.7.3 Least square method of estimation of the parameters of simple linear model

Consider the simple linear model $Y = \beta_0 + \beta_1 X + e$. Let there be n pair of observations $(x_1, y_1), (x_2, y_2).....(x_n, y_n)$. In the above model Y is called the dependent variable and X, the independent variable and e is the random error. For the i^{th} pair of observation $Y_i = \beta_0 + \beta_i X_i + e_i$

Or $e_i = Y_i - \beta_0 + \beta_1 X_i$ may be positive or negative, to avoid the problem of sign of error, square both side and also take the sum over all pair then we have

$Q = \Sigma e_i^2 = \Sigma(Y_i - \beta_0 - \beta_1 x_i)^2$ to get the best estimate of β_0 and β_1 one has to minimize $Q = \Sigma e_i^2$, which amounts to minimize $\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2$. Thus value of β_0 and β_1 which minimize Q are known as least square estimator. To estimate parameters β_0 and β_1 differential Q partially with respect to β_0 and β_1 respectively and equate to zero

i.e. $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$ which gives

$\Sigma Y_i = n\beta_0 - \beta_i \Sigma x_i$ and

$\Sigma x_i Y_i = \beta_0 \Sigma X_i - \beta_1 \Sigma X_i^2$

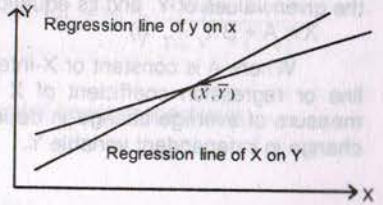
These two equations are known as normal equation. On solving these two equations we obtain as

$\beta_0 = \bar{Y} - \beta_1 \bar{X}$ and $\beta_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$

6.7.4 Lines of Regression

The lines of regression are the lines, which gives the best estimated value of dependent variables for the given value of independent variable's value. Let X and Y are two variables and there are two lines of regression.

i. Regression line of Y on X and
ii. Regression line of X on Y.



The derivation and assumption of the two regression lines are quite different. Where as these regression lines are not interchangeable and these lines are of average relationship but not of actual relationship. The regression lines of Y on X is obtained by minimizing error sums of square of estimation of Y while regression line of X on Y is obtained by minimizing the error sums of square for estimation X. This is the reason why there are two different regression equations for a given set of a data.

a. Regression line of Y on X

The regression line of Y on X gives the best estimated value of Y for the given values of X. The regression equation of Y on X is,

$Y = a + bX$ ... (i)

Where 'a' is constant or y-intercept and 'b' is the slope of the regression line or regression coefficient of Y on X which is denoted by b_{YX}. It is a measure of average change in dependent variable Y corresponding to a unit change in independent variable X.

In equation (i) knowing the value X, one can't estimate the value of Y without knowing the values of constants 'a' and 'b' so; first we should estimate the value of unknown regression parameters by using the Principle of least square estimation (LSE). Using the LSE for equation (i) we obtain the following two normal equations as;

$\Sigma Y = na + b\Sigma X$ ..... (ii)

$\Sigma XY = a\Sigma X + b\,\Sigma X^2$ ..... (iii)

Where 'n' is the number of pairs of observation, by solving these two normal equations, we get

$$a = \bar{Y} - b\bar{X} \quad \text{and} \quad b = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} = b_{YX}$$

Now, putting the values of 'a' and 'b' in equation (1) we get the required estimated regression equation of Y on X which is as follows.

$\hat{Y} = a + bX$

where $\hat{Y}$ is estimated or calculated value of Y for given value of X.

b. Regression line of X on Y

The regression line of X on Y gives the best estimated value of X for the given values of Y. and its equation is given by

$X = A + B\,Y$ ... (i)

Where A is constant or X-intercept and B is the slope of the regression line or regression coefficient of X on Y which is denoted by b_{XY} . It is a measure of average change in dependent variable X corresponding to a unit change in independent variable Y.

In equation (i) knowing the value Y, one can't estimate the value of X without knowing the values of constants A and B so, first we should estimate the value of unknown regression parameters by using the Principle of least square estimation (LSE). Using the LSE for equation (i) we obtain the following two normal equations as

$\Sigma X = nA + B\,\Sigma Y$ ... (ii)

$\Sigma XY = A\Sigma Y + B\,\Sigma Y^2$ ... (iii)

Where 'n' is the number of pairs of observation, by solving these two normal equations, we get

$$A = \bar{x} - B\,\bar{y} \quad \text{and} \quad B = \frac{n\Sigma XY - \Sigma XY}{n\Sigma Y^2 - (\Sigma Y)^2} = b_{XY}$$

Now, putting the values of 'A' and 'B' in equation (i) we get the required estimated regression equation of X on Y which is as follows.

$\hat{X} = a' + b'Y,$

Where $\hat{X}$ is estimated or calculated value of X for given value of Y.

6.7.5 Alternative Method of Fitting Regression Lines

The regression lines of Y on X and X on Y can be expressed respectively as

$Y - \bar{Y} = b_{YX}(X - \bar{X})$

$X - \bar{X} = b_{XY}(Y - \bar{Y})$

This shows that $(\bar{X}$ and $\bar{Y})$ is the common point of two regression lines. Where, $\bar{X}$ and $\bar{Y}$ are the arithmetic means of X and Y series. b_{YX} and b_{XY} are the regression coefficient of Y on X and regression coefficient of X on Y respectively. The regression coefficients are computed as

Direct method

$$b_{YX} = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma X^2 - (\Sigma X)^2} \qquad b_{XY} = \frac{n\Sigma XY - \Sigma X\Sigma Y}{n\Sigma Y^2 - (\Sigma Y)^2}$$

Shortcut method

If we take the deviation from assumed mean A and B then,

$$b_{YX} = b_{vu} = \frac{n\Sigma uv - \Sigma u\Sigma v}{n\Sigma u^2 - (\Sigma u)^2} \quad \text{and} \quad b_{XY} = b_{uv} = \frac{n\Sigma uv - \Sigma u\Sigma v}{n\Sigma v^2 - (\Sigma v)^2}$$

Where, $u = X - A$ and $v = Y - B$.

A and B are assumed mean of X and Y series respectively.

Step deviation method

Under this method,

$$b_{YX} = \frac{k}{h} \times \frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma u^2 - (\Sigma u)^2} = \frac{k}{h} b_{vu}$$

$$b_{XY} = \frac{h}{k} \times \frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma v^2 - (\Sigma v)^2} = \frac{h}{k} b_{uv}$$

Where, $u = \frac{X - A}{h}$ and $v = \frac{Y - B}{h}$, h and k are the common factor of X and Y series respectively.

It is to be notated that $b_{YX} = r \frac{\sigma_Y}{\sigma_X}$ and $b_{XY} = r \frac{\sigma_X}{\sigma_Y}$

Also, $r = \sqrt{b_{YX}\, b_{XY}}$, where r is the correlation coefficient.

Remarks

i. The two regression lines coincides if there is perfect correlation between the two variable i.e. $r = \pm 1$

ii. The two regression lines are perpendicular to each other if there is no correlation between the two variables i.e. $r = 0$.

6.7.6 Properties of Regression Coefficients

The following are the important properties of regression coefficients.

i. Both regression coefficients must have same sign and the sign of correlation coefficient is same as the sign of regression coefficients.

ii. The product of the two regression coefficient must be less than or equal to 1 i.e. $b_{YX}\, b_{XY} \leq 1$

iii. The correlation coefficient is the geometric mean between two regression coefficient i.e. $r = \sqrt{b_{YX}\, b_{XY}}$

iv. The two lines of regression intersect at the point $(\bar{X}, \bar{Y})$, where X and Y are the variables under study.

v. Regression coefficient are independent of change of origin but not of scale

i.e. $b_{XY} = b_{UV}$ if $u = X - A$ and $v = Y - B$ and

$b_{XY} \neq b_{UV}$ if $u = \frac{X - A}{h}$ and $v = \frac{Y - B}{K}$, Where A and B are assumed mean, h and k are the common factor of X and Y series respectively.

vi. Arithmetic mean of regression coefficient is greater than the correlation coefficient.

Proof: we have two regression coefficients as

$b_{YX} = r \frac{\sigma_Y}{\sigma_X}$ and $b_{XY} = r \frac{\sigma_X}{\sigma_Y}$. Then we have to show that

$$\frac{b_{YX} + b_{XY}}{2} > r$$

Or, $$\frac{r\frac{\sigma_Y}{\sigma_X} + r\frac{\sigma_X}{\sigma_Y}}{2} > r$$

Or, $r\left(\frac{\sigma_Y}{\sigma_X} + \frac{\sigma_X}{\sigma_Y}\right) > 2r$

Or, $\sigma^2_Y + \sigma^2_X > 2\sigma_Y \sigma_X$

Or, $(\sigma_Y - \sigma_X)^2 > 0$, which is always true.

6.7.7 Angle between Two Regression lines

We have two regression lines as

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

$$X - \bar{X} = r\frac{\sigma_X}{\sigma_Y}(Y - \bar{Y})$$

Now, slope of regression equation of Y on X $(m_1) = r\frac{\sigma_Y}{\sigma_X}$

Slope of regression equation of X on Y $(m_2) = \frac{\sigma_Y}{r\,\sigma_X}$

If θ be the angle between two regression lines then we have

$$\text{Tan}\,\theta = \pm\frac{m_1 - m_2}{1 + m_1 m_2} = \pm\frac{r^2 - 1}{r}\,\frac{\sigma_Y\sigma_X}{\sigma^2_X + \sigma^2_Y}$$

$$= \frac{1 - r^2}{r}\,\frac{\sigma_Y\sigma_X}{\sigma^2_X + \sigma^2_Y}$$ [taking negative sign since $r^2 \leq 1$]

$$\therefore \theta = \tan^{-1}\left\{\frac{1 - r^2}{r}\,\frac{\sigma_Y\sigma_X}{\sigma^2_X + \sigma^2_Y}\right\}$$

Case I : When $r = 0$, $\theta = \pi/2$, it means if two variables are uncorrelated, the two regression becomes perpendicular to each other.

Case II : If $r = \pm 1$, $\theta = 0$ and π, it means if two variables are perfectly correlated then two regression equations are either parallel or coincide.

Case III : If $\tan\theta > 0$ then θ is acute angle and If $\tan\theta < 0$ then θ is obtuse angle

Case IV : Correlation coefficient lies between – 1 to 1

Proof: The regression equation of Y on X is given as

$$Y - \bar{Y} = r\frac{\sigma_Y}{\sigma_X}(X - \bar{X})$$

Or, $\sigma_X (Y - \bar{Y}) = r\,\sigma_Y (X - \bar{X})$, on squaring both sides

Or, $\sigma^2_X (Y - \bar{Y})^2 = r^2\sigma^2_Y (X - \bar{X})^2$, taking expectation on both sides

Or, $\sigma^2_X E(Y - \bar{Y})^2 = r^2\sigma^2_Y E(X - \bar{X})^2$

Or, $\sigma^2_X \sigma^2_Y = r^2\sigma^2_Y \sigma^2_X$

Or, $r^2 = 1$

Or, $|r| = 1 \qquad \therefore -1 \le r \le 1$

6.8 CURVILINEAR REGRESSION

There are various relatonship between variables X and Y namely linear and non-linear. The non linear relationship between variables X and Y is aslo known as curvilinear regression. Some of commonly used crvilinear regressions in engineering are as follows

i. Quadratic or Parabolic Regression Equation

ii. Exponential Regression Equation

iii. Reciprocal regression equation

6.8.1 Fitting of Quadratic Regression Equation

A quadratic or second degree or parabolic regression is the simplest of the curvilinear models. If future estimation of regresson is not properly drawn by straight line regression, we use this method.

Let, $y = a + bX + cX^2$ .....(i), be a quadratic or parabolic line equation of the given data.

Where a, b and c are constant to be estimated by using the method of least square for this we have to solve the following three normal equations as

$\Sigma Y = na + b\Sigma X + c\Sigma X^2$..... (ii)

$\Sigma XY = a\Sigma X + b\Sigma X^2 + c\Sigma X^3$.....(iii)

$\Sigma X^2Y = a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4$.....(iv)

By solving the above equations, (ii), (iii) and (iv), we get value of a, b and c. Then the fitted second degree regression equation is

$Y_c = a + bX + cX^2$

6.8.2 Fitting of Exponential Regression Equation

An exponential regression equation is defined as

$Y = a\,b^X$

Taking log on both sides we have

$\text{Log}_{10}Y = \text{Log}_{10}a + \text{Log}_{10}b\,X$.

This is the equation of straight line, $\log_{10}Y$ is dependent on X and $\text{Log}_{10}a$ and $\log_{10}b$ are constant obtained by method of least square.

Another form of exponential relationship is

$Y = a\,X^b$

Taking log on both sides we have

$\text{Log}_{10}Y = \text{Log}_{10}a + b\,\text{Log}_{10}X$.

This is the equation of straight line, $\log_{10}Y$ is dependent on $\log_{10}X$ and $\text{Log}_{10}a$ and b are constant obtained by method of least square.

6.8.3 Fitting of Reciprocal Regression Equation

We have the reciprocal relationship between two variables X and Y commonly used in engineering as

$$Y = \frac{1}{a + bX}$$

Or, $a + bX = \frac{1}{Y}$.

Here for the points $(X, \frac{1}{Y})$ the graph is approximately straight line. $\frac{1}{Y}$ depends on Variable X and constant a and b are obtained by method of least square by solving following two normal equations as

$\Sigma\frac{1}{Y} = n\,a + b\,\Sigma X$ and

$\Sigma\frac{X}{Y} = a\,\Sigma X + b\,\Sigma X^2$

Example 14

The following measurements show the respective height in inches of 10 fathers and their eldest sons

father (X)	67	63	66	71	69	65	62	70	61	72
son (Y)	68	66	65	70	67	67	64	71	62	63

i. Find the regression line of son's height on father's heights

ii. Estimate the height of son for the given height of father as the 70 inches.

iii. Find the regression line of X on Y.

iv. At what point the two line of regression intersect.

v. Fine the correlation between variables X and Y

Solution

i. The mean values of X and Y

ii. The standard deviation of Y.

Solution

i. Since the lines of regression pass through the average of X and Y i.e.

$(\bar{X}, \bar{Y})$ so, $4\bar{x} - 5\bar{y} + 33 = 0$ and $20\bar{X} - 9\bar{Y} - 107 = 0$

On solving two equations we get $\bar{X} = 13$ and $\bar{y} = 17$

ii. From equation $4X - 5Y + 33 = 0$

$Y = \frac{4}{5}X + \frac{33}{5}$; which is regression equation of Y on X

so, $b_{YX} = \frac{4}{5}$

Again from equation $20X - 9Y - 107 = 0$

$X = \frac{9}{20}Y + \frac{107}{20}$ which is the regression equation of X on Y

So, $b_{XY} = \frac{9}{20}$

Now, we have correlation coefficient is the geometric mean of regression coefficient thus

$$r = \sqrt{b_{YX} \times b_{XY}} = \sqrt{\frac{4}{5} \times \frac{9}{20}} = 0.6$$

Again we have $b_{YX} = \frac{4}{5}$

Or, $r\frac{\sigma_Y}{\sigma_X} = \frac{4}{5}$ which implies $\sigma_Y = 4$

Therefore, the value of standard deviation of Y is 4.

Example 17

The following data gives the experience of machine operators in years and their performance as given by the number of good parts turned out per 100 pieces.

Operator	I	II	III	IV	V	VI	VII	VIII
Experience (X)	16	12	18	4	3	10	5	12
Performance(Y)	87	88	89	68	78	80	75	83

Calculate the regression equation of performance rating on experience and estimate the probable performance if an operator has 8 years experience.

Solution

Here we have to find the regression equation of performance rating (Y) on experience (X) i.e. Y on X. The regression equation of Y on X is given by

$$Y - \bar{Y} = b_{YX}(X - \bar{X}) \quad \ldots\ldots(i)$$

Where $\bar{Y}$ is mean of Y, $\bar{X}$ mean of X series, and b_{YX} is the regression coefficient of Y on X.

Let, $u = X - 12$ and $v = Y - 83$

Calculation of regression equation

X	Y	u	v	u^2	v^2	uv
16	87	4	4	16	16	16
12	88	0	5	0	25	0
18	89	6	6	36	36	36
4	68	-8	-15	64	225	120
3	78	-9	-5	81	25	45
10	80	-2	-3	4	9	6
5	75	-7	-8	49	64	56
12	83	0	0	0	0	0
	Total	**-16**	**-16**	**250**	**400**	**279**

Here $n = 8$, $\Sigma u = -16$, $\Sigma v = -16$, $\Sigma u^2 = 250$, $\Sigma v^2 = 400$ and $\Sigma uv = 279$

$$\therefore \bar{X} = a + \frac{\Sigma u}{n} = 12 + \frac{-16}{8} = 10 \text{ and}$$

$$\bar{Y} = b + \frac{\Sigma v}{n} = 83 + \frac{-16}{8} = 81$$

$$b_{YX} = \frac{n\Sigma uv - \Sigma u \Sigma v}{n\Sigma u^2 - (\Sigma u)^2} = \frac{8 \times 279 - (-16)(-16)}{8 \times 250 - (-16)} = 1.133$$

Now, Substituting the value of $\bar{X}$, $\bar{Y}$ and b_{YX} on equation (i) then

$Y - 81 = 1.133(X - 10)$

$Y = 1.133X + 69.67$ which is the required regression equation of Y on X

For estimation when $X = 8$ year $Y = ?$

Therefore, $Y = 1.133 \times 8 + 69.67 = 78.734$.

Thus, if an operator has 8 years experience then his performance is 78.74 good parts turned out per 100 pieces.

6.9 MULTIPLE REGRESSION

Multiple regressions generally explain the relationship between multiple independent or multiple predictor variables and one dependent or criterion variable. A dependent variable is modeled as a function of several independent variables with corresponding coefficients, along with the constant term. Multiple regressions require two or more predictor variables, and this is why it is called multiple regressions.

The multiple regression equation explained above takes the following form

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$$

Here, b_i's (i=1,2...n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes. The constant a is the value of dependent variable Y when the values of all independent variable is zeros.

As an example, let's say that the test score of a student in an exam will be dependent on various factors like his focus while attending the class, his intake of food before the exam and the amount of sleep he gets before the exam. Using this test one can estimate the appropriate relationship among these factors.

Let us consider three variables X_1, X_2 and X_3. The multiple regression equation of dependent variable X_1 on two independent variables X_2 and X_3 is an equation for estimating a dependent variable X_1 from two independent variables X_2 and X_3 which is given below

$$X_1 = a + b_1X_2 + b_2X_3 \qquad (1)$$

Where b_1 and b_2 are known as coefficients of regression. The three constants are obtained by solving following three normal equations simultaneously obtained by the method of least squares

$$\sum X_1 = na + b_1\sum X_2 + b_2\sum X_3$$

$$\sum X_1 X_2 = a\sum X_2 + b_1\sum X_2^2 + b_2\sum X_3 X_2$$

$$\sum X_1 X_3 = a\sum X_3 + b_1\sum X_2 X_3 + b_2\sum X_3^2$$

These equations can be obtained by multiplying both sides of equation (1) By 1, X_2 and X_3 successively and summing on both sides

Similarly the multiple regression equation of dependent variable X_2 on two independent variables X_1 and X_3 is an equation for estimating a dependent variable X_2 from two independent variables X_1 and X_3 which is given below

$$X_2 = a + b_1X_1 + b_2X_3 \qquad (2)$$

Where b_1 and b_2 are known as coefficients of regression. The three constants are obtained by solving following three normal equations simultaneously obtained by the method of least squares

$$\sum X_2 = na + b_1\sum X_1 + b_2\sum X_3$$

$$\sum X_1 X_2 = a\sum X_1 + b_1\sum X_1^2 + b_2\sum X_1 X_3$$

$$\sum X_2 X_3 = a\sum X_3 + b_1\sum X_1 X_3 + b_2\sum X_3^2$$

Also, the multiple regression equation of dependent variable X_3 on two independent variables X_1 and X_2 is an equation for estimating a dependent variable X_2 from two independent variables X_1 and X_3 which is given below

$$X_3 = a + b_1X_1 + b_2X_2 \qquad (3)$$

Where b_1 and b_2 are known as coefficients of regression. The three constants are obtained by solving following three normal equations simultaneously obtained by the method of least squares

$$\sum X_3 = na + b_1\sum X_1 + b_2\sum X_2$$

$$\sum X_1 X_3 = a\sum X_1 + b_1\sum X_1^2 + b_2\sum X_1 X_2$$

$$\sum X_2 X_3 = a\sum X_2 + b_1\sum X_1 X_2 + b_2\sum X_2^2$$

Example 18

Following data reveals the sales of a company due to the number of sales persons and the years of experience

Sales(Rs'000)	20	30	25	20	40	60	15
No.of sales person	2	3	5	4	2	1	4
years of experience	5	7	11	10	8	7	8

i. Estimate the best line of fit

ii. Estimate the sales of the company using one sales person with experience of nine years.

Solution

To find the sales of the company, we should assume sales as dependent variable. The multiple regression equation of dependent variable X_1 on two independent variables X_2 and X_3 is given below

$$X_1 = a + b_1X_2 + b_2X_3 \qquad (1)$$

Where b_1 and b_2 are known as coefficients of regression, the three constants are obtained by solving following three normal equations simultaneously obtained by the method of least squares

$$\sum X_1 = na + b_1\sum X_2 + b_2\sum X_3$$

$$\sum X_1 X_2 = a\sum X_2 + b_1\sum X_2^2 + b_2\sum X_3 X_2$$

$$\sum X_1 X_3 = a\sum X_3 + b_1\sum X_2 X_3 + b_2\sum X_3^2$$

Calculation Table

X_1	X_2	X_3	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
20	2	5	4	25	40	100	10
30	3	7	9	49	90	210	21
25	5	11	25	121	125	275	55
20	4	10	16	100	80	200	40
40	2	8	4	64	80	320	16
60	1	7	1	49	60	420	7
15	4	8	16	64	60	120	32
210	**21**	**56**	**75**	**472**	**535**	**1645**	**181**

Substituting these values in above equations, we get

$7a + 21\, b_1 + 56\, b_2 = 210$

$21a + 75\, b_1 + 181\, b_2 = 535$

$56a + 181\, b_1 + 472\, b_2 = 1645$

On solving these three equations we get

$a = 21.22$, $b_1 = -15.34$ and $b_2 = 6.85$.

i. Thus the required estimated best line of fit is

$X_1 = 21.22 - 15.34X_2 + 6.85X_3$

ii. The estimated value of sales of company using one sales person $X_2 = 1$ and Nine years of experience $X_3 = 9$ is given as

$X_1 = 21.22 - 15.34 \times 1 + 6.85 \times 9 = 67.43$

Thus the estimated sales is Rs 67.43 (Rs '000)

SOME WORKED OUT EXAMPLES

Example 19

Find the coefficient of correlation between the age and playing habit of the following students.

Age (year)	15	16	17	18	19	20
No. of students	125	100	75	60	50	40
No. of regular players	100	75	45	24	15	6

Solution

In order to calculate the correlation coefficient between age and playing habit of the students, let us convert the data of number of players with respect to total number of students in each age group with fixed base for this purpose, percentage of players is calculated. Then the correlation between Age (X) and percentage of player (Y) is calculated as follows. Percentage of player for year 15 $= \frac{100}{125} \times 100 = 80\%$ and so on.

Let, $u = X - 17$, $v = \frac{Y - 60}{5}$

Calculation of correlation coefficient

X	Y	u	v	uv	u^2	v^2
15	80	-2	4	-8	4	16
16	75	-1	3	-3	1	9
17	60	0	0	0	0	0
18	40	1	-4	-4	1	16
19	25	2	-7	-14	4	49
20	15	3	-9	-27	9	81
Total		$\Sigma u = 3$	$\Sigma v = -13$	$\Sigma uv = -56$	$\Sigma u^2 = 19$	$\Sigma v^2 = 171$

Now, correlation coefficient

$$r = \frac{n\Sigma uv - \Sigma u \Sigma v}{\sqrt{n\Sigma u2 - (\Sigma u)2} \times \sqrt{(n\Sigma v2 - (\Sigma v)2)}}$$

$$= \frac{6\times(-56) - 3\times(-13)}{\sqrt{6\times19 - (3)2} \times \sqrt{6\times171 - (-13)2}}$$

$$= \frac{-297}{10.246 \times 29.27} = -0.991$$

This shows that there is high degree of negative correlation between age and playing habits of students.

Example 20

While calculating correlation coefficient by using Karl Pearson's correlation method the pairs of observations (20, 65) and (18, 50) were taken wrongly instead of (25, 60) and (28, 59) and the information while calculating correlation are given as,

$r = 0.793$, $n = 10$, $\Sigma X = 220$, $\Sigma Y = 590$

$\Sigma X^2 = 4954$, $\Sigma Y^2 = 35232$ and $\Sigma XY = 13214$

Calculate the corrected correlation coefficient.

Solution

To calculate the corrected value of r we firstly calculate the corrected values as follows,

Corrected $\Sigma X = 220 - 20 - 18 + 25 + 28 = 235$

Corrected $\Sigma Y = 590 - 65 - 50 + 60 + 59 = 594$

Corrected $\Sigma X^2 = 4954 - (20)^2 - (18)^2 + (25)^2 + (28)^2 = 5639$

Corrected $\Sigma Y^2 = 35232 - (65)^2 - (50)^2 + (60)^2 + (59)^2 = 35588$

Corrected $\Sigma XY = 13214 - (20 \times 65) - (18 \times 50) + (25 \times 60) + (28 \times 59)$

$= 12314 - 360 - 3250 + 700 + 3540 = 13844.$

$$\therefore\ r_{\text{Corrected}} = \frac{n\Sigma XY - \Sigma X \Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma x)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{10\times13844 - 235\times594}{\sqrt{10\times5639 - (235)^2}\sqrt{10\times35588 - (594)^2}}$$

$$= \frac{138440 - 139590}{34.132\times55.1724}$$

$$= \frac{-1150}{34.132\times55.1724} = -0.610$$

Therefore, the required corrected correlation coefficient is – 0.610.

Example 21

From the following bi-variate distribution, compute two regression coefficients, coefficient of variation, coefficient of correlation and estimate the value of Y when value of X is 45.

Y \ X	10 – 20	20 – 30	30 – 40	40 – 50
10 – 20	20	26	–	–
20 – 30	8	14	37	–
30 – 40	–	4	18	3
40 – 50	–	–	4	6

Solution

Let, A = 25, B = 35 and $u = \frac{X-25}{10}$ and

$v = \frac{Y-35}{10}$ $\frac{Y-35}{10}$

Class	Y	v \ u	-1	0	1	2	f	fv	fv^2	fuv
		X	15	25	35	45				
10-20	15	-2	[40] 20	[0] 26	–	–	46	-92	184	40
20-30	25	-1	[8] 8	[0] 14	[-37] 37	–	59	-59	59	-29
30-40	35	0	–	[0] 4	[0] 18	[0] 3	25	0	0	0
40-50	45	1	–	–	[4] 4	[12] 6	10	10	10	16
		f	28	44	59	9	140	-141	253	27
		fu	-28	0	59	18	49			
		fu^2	28	0	59	36	123			
		fuv	48	0	33	12	27			

Here, N = Σf = 140

Σf u = 49, Σf v = - 141

$\Sigma f u^2 = 123$, $\Sigma f v^2 = 253$, Σfuv = 27

i. We have correlation coefficient for bi-variate distribution is,

$$r = \frac{N\Sigma fuv - \Sigma fu\,\Sigma fv}{\sqrt{N\Sigma fu^2 - (\Sigma fu)^2}\sqrt{N\Sigma fv^2 - (\Sigma fv)^2}}$$

$$= \frac{140 \times 27 - 49 \times (-141)}{\sqrt{140 \times 123 - (49)^2}\sqrt{140 \times 253 - (-141)^2}}$$

$$= \frac{3780 + 6909}{\sqrt{17220 - 2481}\sqrt{35420 - 19881}} = \frac{10689}{121.404 \times 124.655}$$

= 0.706

ii. For coefficient of variation, first we calculate the means and standard deviation of X and Y.

Mean of X, $\bar{X} = A + \frac{\Sigma fu}{N} \times h = 25 + \frac{49}{140} \times 10$

$= 25 + 3.5 = 28.5$

Mean of Y, $\bar{Y} = B + \frac{\Sigma fv}{N} \times k = 35 - \frac{141}{140} \times 10$

$= 35 - 10.071 = 24.929$

S.D. of X, $\sigma_x = \sqrt{\frac{\Sigma fu^2}{N} - \left(\frac{\Sigma fu}{N}\right)^2} \times h = \sqrt{\frac{123}{140} - \left(\frac{49}{140}\right)^2} \times 10$

$= \sqrt{0.8785 - 0.1225} \times 10 = 8.695$

S.D. of Y, $\sigma_y = \sqrt{\frac{\Sigma fv^2}{N} - \left(\frac{\Sigma fv}{N}\right)^2} \times k = \sqrt{\frac{253}{140} - \left(\frac{-141}{140}\right)^2} \times 10$

$= \sqrt{1.80714 - 1.01434} \times 10 = 8.904.$

$\therefore$ Coefficient of variation of $X = \frac{\sigma_x}{\bar{X}} \times 100\ \%$

$= \frac{8.695}{28.5} \times 100\ \% = 30.51\%$

Coefficient of variation of $Y = \frac{\sigma_Y}{\bar{Y}} \times 100\ \%$

$= \frac{8.904}{24.929} \times 100\ \% = 35.72\%$

iii. The regression coefficient of Y on X is

$$b_{yx} = r\frac{\sigma_Y}{\sigma_X} = 0.706 \times \frac{8.904}{8.695} = 0.723$$

And, the regression coefficient of X on Y is

$$b_{xy} = r\frac{\sigma_X}{\sigma_Y} = 0.706 \times \frac{8.695}{8.904} = 0.689$$

iv. To estimate the value of Y when value of X is given, we have to find the regression line of Y on X as,

$$Y - \bar{Y} = b_{yx}(X - \bar{X})$$

$$Y - 24.929 = 0.723(X - 28.5)$$

$$Y = 24.929 + 0.723X - 20.6055$$

$$\hat{Y} = 0.723X + 4.3235$$

Which, is required regression line of Y on X

Then, value of Y when, $X = 45$ is

$$\hat{Y} = 0.723 \times 45 + 4.3235$$

$$\therefore \hat{Y} = 36.858$$

Example 22

In two sets of variables X and Y with 50 observations each the following data were observed.

$$\bar{X} = 10,\ \bar{Y} = 6,\ \sigma_x = 3,\ \sigma_x = 2$$

Coefficient of correlation between X and Y is 0.3. However on subsequent verification it was found that one pair (X = 10, Y = 6) was inaccurate and hence waived out. With the remaining 49 pairs of values, how is the original value of correlation coefficient affected?

Solution

We are given, $n = 50,\ \bar{X} = 10,\ \bar{Y} = 6,\ \sigma_x = 3,\ \sigma_x = 2,\ r_{xy} = 0.3$

We have,

$$\bar{X} = \frac{\Sigma X}{n} \qquad \bar{Y} = \frac{\Sigma Y}{n}$$

$$10 = \frac{\Sigma X}{50} \qquad 6 = \frac{\Sigma Y}{50}$$

$$\Sigma X = 500 \qquad \Sigma Y = 300$$

$$\sigma_x^2 = \frac{1}{n}\Sigma(X - \bar{X})^2 \qquad \sigma_Y^2 = \frac{1}{n}\Sigma(Y - \bar{Y})^2$$

$$\sigma_x^2 = \frac{\Sigma X^2}{n} - (\bar{X})^2 \qquad \sigma_Y^2 = \frac{\Sigma Y^2}{n} - (\bar{Y})^2$$

$$n\sigma_X^2 = \Sigma X^2 - n\bar{X}^2 \qquad n\sigma_Y^2 = \Sigma Y^2 - n\bar{Y}^2$$

$$\Sigma X^2 = n(\sigma_X^2 + \bar{X}^2) \qquad \Sigma Y^2 = n(\sigma_Y^2 + \bar{Y}^2)$$

$$= 50(9 + 100) \qquad = 50(4 + 36)$$

$$= 5450 \qquad = 2000$$

Also, the value of Karl Pearson's correlation coefficient is

$$r = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$$

$$\therefore r\sigma_X \sigma_Y = Cov(X, Y) = \frac{\Sigma XY}{n} - \bar{X}\bar{Y}$$

$$0.3 \times 3 \times 2 = \frac{\Sigma XY}{50} - 10 \times 6$$

$$\therefore \Sigma XY = 50 \times 61.8 = 3090$$

One pair of observations (X = 10, Y = 6) is wrong. Omitting this pair of observations we have,

$$n = 50 - 1 = 49$$

Now, the corresponding correct values are

$$\Sigma X = 500 - 10 = 490 \qquad \Sigma Y = 300 - 6 = 294$$

$$\Sigma X^2 = 5450 - 10^2 = 5350 \qquad \Sigma Y^2 = 2000 - 6^2 = 1964$$

$$\Sigma XY = 3090 - 10 \times 6 = 3030$$

Now putting the corrected values of ΣX, ΣY, ΣX^2, ΣY^2 and ΣXY in the following formula we get corrected correlation coefficient

$$r = \frac{n\Sigma XY - \Sigma X\,\Sigma Y}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{49 \times 3030 - 490 \times 294}{\sqrt{49 \times 5350 - (490)2}\sqrt{49 \times 1964 - (294)2}}$$

$$= \frac{148470 - 144060}{\sqrt{262150 - 240100}\sqrt{96236 - 86436}} = \frac{4410}{\sqrt{22050 \times 9800}}$$

$$= \frac{4410}{14700} = 0.3$$

Therefore, the corrected correlation coefficient is 0.3. Thus, in this case, the original value of correlation coefficient is not affected.

Example 23

A computer while calculating correlation coefficient between two variables X and Y from 25 pairs of observations obtained the following results.

$$n = 25,\ \Sigma X = 125,\ \Sigma X^2 = 650,\ \Sigma Y = 100,\ \Sigma Y^2 = 460,\ \Sigma XY = 508$$

It was however discovered at the time of checking that two pairs of observations were not correctly copied. They were taken as (6, 14) and (8, 6) while the correct values were (8, 12) and (6, 8). Prove that the correct value of the correlation coefficient should be $\frac{2}{3}$.

Solution

We have to add the correct values and subtract the wrong values as in all sum values. The corresponding corrected sum values are

Correct $\Sigma X = 125 - 6 - 8 + 8 + 6 = 125$

Corrected $\Sigma Y = 100 - 14 - 6 + 12 + 8 = 100$

Corrected $\Sigma X^2 = 650 - 6^2 - 8^2 + 8^2 + 6^2 = 650$

Corrected $\Sigma Y^2 = 460 - 14^2 - 6^2 + 12^2 + 8^2 = 436$

Corrected $\Sigma XY = 508 - 6 \times 14 - 8 \times 6 + 8 \times 12 + 6 \times 8 = 520$

Corrected value of r is given by

$$\text{Corrected } r = \frac{n\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{n\Sigma X^2 - (\Sigma X)^2}\sqrt{n\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{25 \times 520 - 125 \times 100}{\sqrt{25 \times 650 - (125)^2}\sqrt{25 \times 436 - (100)^2}}$$

$$= \frac{13000 - 12500}{\sqrt{16250 - 15625}\sqrt{10900 - 10000}}$$

$$= \frac{500}{\sqrt{625 \times 900}} = \frac{500}{25 \times 30} = \frac{500}{750}$$

$$= \frac{2}{3} \text{ Thus verified}$$

Example 24

A student calculates the value of r as 0.7, when the value of n is 5 and he concludes that r is highly significant. Does he correct? Calculate the limits for population correlation coefficient. If the calculated value of PE (r) = 0.085 for r = 0.7 find the value of n.

Solution

We have, r = 0.7, n = 5

$$PE(r) = 0.6745\frac{1 - r^2}{\sqrt{n}} = 0.6745 \times \frac{1 - (0.7)^2}{\sqrt{5}} = 0.154$$

And, $6\,PE(r) = 6 \times 0.154 = 0.924$

Hence, this shows that r is not greater than 6 PE.

Thus, we can not make any decision about the significance of correlation coefficient. It is seen that his conclusion becomes wrong.

Limits for population correlation coefficients are

$r \pm PE(r) = 0.7 \pm 0.154$

$\therefore$ Upper limit of $r = 0.7 + 0.15 = 0.854$

Lower limit of $r = 0.7 - 0.154 = 0.546$

Now, if PE(r) = 0.085, r = 0.7, n = ?

We have, $PE(r) = 0.6745\frac{1 - r^2}{\sqrt{n}}$

$$0.085 = 0.6745 \times \frac{1 - (0.7)^2}{\sqrt{n}}$$

$$0.085\sqrt{n} = 0.344$$

$$\sqrt{n} = \frac{0.344}{0.085}$$

$$\sqrt{n} = 4.047$$

n = 16 (approximately)

Example 25

For the following information

	X	Y
Arithmetic mean(in Rs)	6	8
Standard deviation (in Rs)	5	40/3

Correlation coefficient between X and Y is 8/15.

Find
a. The regression coefficient of Y on X and X on Y
b. The two regression equations
c. The most likely value of Y when X = 100 rupees.

Solution

We have,

$\bar{X} = 6,\ \bar{Y} = 8,\ \sigma_x = 5,\ \sigma_y = 40/3,\ r = 8/15$

a. Regression coefficient of Y on X is

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{8}{15} \times \frac{40/3}{5} = 1.422$$

Similarly, regression coefficient of Y or Y is

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{8}{15} \times \frac{5}{40/3} = 0.2$$

b. The regression equation of Y on X is

$Y - \bar{Y} = b_{yx}(X - \bar{X})$

$Y - 8 = 1.422(X - 6)$

$\hat{Y} = 1.422X - 0.532$

Similarly, the regression equation of X on Y is

$X - \bar{X} = b_{xy}(Y - \bar{Y})$

$X - 6 = 0.2(Y - 8)$

$\hat{X} = 0.2Y + 4.4$

c. $\hat{Y} = ?$ When X = 100

$\hat{Y} = 1.422 \times 100 - 0.532$

$= 142.2 - 0.532$

Example 23

For the following information

	X	Y
Arithmetic mean(in Rs)	6	8
Standard deviation (in Rs)	5	40/3

Correlation coefficient between X and Y is 8/15.

Find a. The regression coefficient of Y on X and X on Y

b. The two regression equations

c. The most likely value of Y when X = 100 rupees.

Solution

We have,

$\bar{X} = 6,\ \bar{Y} = 8,\ \sigma_x = 5,\ \sigma_y = 40/3,\ r = 8/15$

a. Regression coefficient of Y on X is

$$b_{yx} = r\frac{\sigma_y}{\sigma_x} = \frac{8}{15} \times \frac{40/3}{5} = 1.422$$

Similarly, regression coefficient of Y or Y is

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} = \frac{8}{15} \times \frac{5}{40/3} = 0.2$$

b. The regression equation of Y on X is

$Y - \bar{Y} = byx\,(X - \bar{X})$

$Y - 8 = 1.422\,(X - 6)$

$\hat{Y} = 1.422\,X - 0.532$

Similarly, the regression equation of X on Y is

$X - \bar{X} = b_{xy}\,(Y - \bar{Y})$

$X - 6 = 0.2\,(Y - 8)$

$\hat{X} = 0.2Y + 4.4$

c. $\hat{Y} = ?$ When $X = 100$

$\hat{Y} = 1.422 \times 100 - 0.532$

$= 142.2 - 0.532$

$= 141.67$

Thus, the most likely value of Y is Rs 141.67.

Exercise – 7

1. What do you mean by correlation? Mention its tyes.
2. What are different methods of finding correlation between two variables? Explain briefly.
3. Explain the concept of simple multiple and partial correlation coefficient.
4. Define Karl Pearson's correlation coefficient and interpret the result of its coefficient.
5. Define Spearman's correlation coefficient and interpret the result of its coefficient.
6. Define the term 'regression'. Discuss tow regression lines.
7. Mention the properties of regression coefficients.
8. Draw a scatter diagram from the following data.

Height(inch)	62	72	70	60	67	70	64	65	60	70
weight(lbs)	50	65	63	52	56	60	59	58	54	65

Also indicate whether correlation is positive or negative.

9. If the covariance between X and Y variable is 18 and the variance of X and Y are 25 and 81 respectively. Find the coefficient of correlation between them,
10. Calculate the correlation coefficient between X and Y series from the following data.

	X series	Y series
No. of observations:	16	16
standard deviation:	3.01	3.03

$\Sigma (X - \bar{X})(Y - \bar{Y}) = 122$

11. For 10 observations on Height (X) and Weight (Y), the following data were obtained (in approximate units) $\Sigma X = 130$, $\Sigma Y = 220$, $\Sigma X^2 = 2290$, $\Sigma Y^2 = 5510$ and $\Sigma XY = 3467$, Compute the coefficient of correlation.
12. Calculate the coefficient of correlation using product moment formula from the data of price and supply given below:

Price(Rs.)	160	162	165	161	163	164	166
supply	63	62	64	63	62	66	68

13. The following table gives the age and blood pressure of 10 patients.

Age	56	42	36	47	49	42	60	72	63	55
Pressure	147	125	118	128	145	140	155	160	149	150

Compute the coefficient of correlation assuming 49 and 140 as the assumed means of age and pressure respectively.

14. Calculate Karl Pearson's coefficient of correlation between expenditure on advertising (Rs. '000) and sells (Lakh Rs.) from the data given below:

Ad. Expenses	39	65	62	90	82	75	25	98	36	78
Sales	47	53	58	86	62	68	60	91	51	84

15. From the following table calculate the coefficient of correlation by Karl Pearson's method.

X	6	2	10	4	8
Y	9	11	?	8	7

Arithmetic means of X and Y series are 6 and 8 respectively.

16. Calculate the Karl Pearson's coefficient of correlation from the following data:

Sum of deviation of X = 5
Sum of deviation of Y = 4
Sum of squares of deviation of X =40
Sum of squares of deviation of Y =50
Sum of product of deviation of X and Y = 32 and
Number of pairs of observation = 10

17. Calculate the coefficient of correlation between the age of students and pass percentage given below:

Age(year)	% pass	Age(year)	% pass
13-14	39	18-19	39
14-15	40	19-20	48
15-16	43	20-21	49
16-17	44	21-22	54
17-18	36		

18. Family income and percentage spent of food in case of 100 families gave the following bi-variate frequency distribution. Find if there is any relation.

Food exp.	Family income				
In (%)	200-300	300-400	400-500	500-600	600-700
10	-	-	-	3	7
15	-	4	9	4	3
20	7	6	12	5	-
25	3	10	19	8	-

19. Following data represent the bi-variate frequency distribution of 25 students getting marks in Statistics and Applied mathematic. Find the coefficient of correlation between the marks.

Marks in	Marks in statistics			
Math	30 - 40	40 - 50	50 - 60	60 - 70
25 - 35	3	1	1	-
35 - 45	2	6	1	2
45 - 55	1	2	2	1
55 - 65	-	1	1	1

20. From the data given below. Find the coefficient of correlation between the driver's age and the number of accidents made by him.

No. of	Drivers age				
accidents	25 - 30	30 - 35	35 - 40	40 - 45	45 - 50
0	-	3	3	7	8
1	-	-	9	4	1
2	3	5	10	3	-
3	4	9	6	-	-
4	12	7	3	1	-

21. In order to find the correlation coefficient between two variables X and Y from 12 pairs of observations, the following calculations were made, $\Sigma X = 30$, $\Sigma Y = 5$, $\Sigma X^2 = 670$, $\Sigma Y^2 = 285$ and $\Sigma XY = 334$
On subsequent verification, it was found that the pair (X = 11, Y = 4) was copied wrongly, the correct value being (X = 10, Y = 14). Find the correct value of correlation coefficient.

22. For a sample of 25 observations, the correlation coefficient is found to be 0.7. Find the limits within which correlation coefficient lies for population by using the probable error of correlation coefficient.

23. If the correlation coefficient is found to be 0.6 for a pair of 64 observations, find the probable error of correlation coefficient r and determine the limits of population correlation coefficient.

24. In a beauty contest two judges rank the 10 competitors in the following order

Competitors	1	2	3	4	5	6	7	8	9	10
Rank by A	6	4	3	2	1	7	9	8	10	5
Rank by B	4	1	6	7	5	8	10	9	3	2

Is there any concordance between the two judges ?

25. Ranks provided by two judges for twelve competitors in painting competition are shown below.

Competitor	1	2	3	4	5	6	7	8	9	10	11	12
Judge I	5	2	3	4	1	6	8	7	10	9	12	11
Judge II	4	5	2	1	6	7	10	9	11	12	3	8

Find Spearman's rank correlation coefficient.

26. Ten industries of Nepal have been raked as follows according to profits earned in 2007 and working capital for the year.

Industry	1	2	3	4	5	6	7	8	9	10
R1 (Profit)	8	6	7	2	3	5	10	1	9	4
R2 (working capital)	10	8	7	2	5	4	6	1	9	3

Calculate the Spearman's rank correlation coefficient

27. Ten competitors in an I.Q. test are raked by three judges in the following ordered.

Competitors	1	2	3	4	5	6	7	8	9	10
Rank by I	1	6	5	10	3	2	4	9	7	8
Rank by II	3	5	8	4	7	10	2	1	6	9
Rank by III	6	4	9	8	1	2	3	10	5	7

Use the method of rank correlation to gauge which pair of judge has the nearest approach to common linking in IQ.

28. Calculate the coefficient of rank correlation between demand and supply of the cement in '000 metric tones.

Demand	15	25	18	20	16	21	10	12
Supply	8	10	12	6	9	14	18	16

29. Calculate Spearman's coefficient of rank correlation for the following data.

X	115	22	148	251	83	47	325	92	70	164
Y	84	385	200	110	292	152	86	120	321	144

30. A company took an examination of eight applicants for an engineering post. From the marks obtained by the applicant in the mathematics and Statistics papers, compute the rank correlation coefficient.

Applicant	1	2	3	4	5	6	7	8
Mathematics	15	20	28	12	40	60	20	80
Statistics	40	30	50	30	20	10	30	60

31. Quotations of index number of equity share prices of a certain joint stock company and of prices of preference shares are given below.

Years	1999	2000	2001	2002	2003	2004	2005
Preference Shares	732	858	789	758	772	812	838
Equity Shares	978	992	988	983	983	967	971

Use the method of rank correlation to determine the relationship between equity share and preference share prices.

32. Calculate Spearman's rank correlation coefficient between the age of person and blood pressure.

Age	56	42	36	47	49	42	60	72	63	55
Pressure	145	125	118	128	145	145	155	160	149	150

33. Obtain Spearman's rank correlation coefficient for the following data.

X	68	64	75	50	64	80	75	40	55	64
Y	62	58	68	45	81	60	68	48	50	70

34. A sample of 12 fathers and their eldest son gave the following data about their heights in inches.

Father	65	63	67	64	68	62	70	66	68	67	69	71
Son	68	66	68	65	69	66	68	65	71	67	68	70

Calculate the correlation coefficient by rank method.

35. The coefficient of rank correlation of the marks obtained by 10 students in probability and theory of structure was found to be 0.8. It was later discovered that the difference in ranks in two subject obtained by one of the student was wrongly taken as 7 instead of 9. Find the correct coefficient of rank correlation.

36. Fit the regression line of Y on X from the following data.

X	57	58	59	59	60	61	62	64
Y	67	68	65	68	72	72	69	71

37. Find the regression equations from the following data.

X	11	12	13	14	15
Y	11	13	15	17	19

38. From the data given below, estimate the most likely height of a brother whose sister's height is 50 cm.

	Brother	Sister
Mean height	170 Cm	75 Cm
Standard deviation of height	5.22	3.85

The coefficient of correlation between the heights of brothers and sisters is 0.60.

39. Find the most likely price in market A corresponding to the Price of Rs. 75 at market B from the following data
Average Price in Market A = Rs. 67
Average Price in Market B = Rs. 65
Coefficient of Variation in Market A = Rs. 5.22
Coefficient of Variation in Market B = Rs. 2.85
The coefficient of correlation between them = 0.82

40. Estimate the loss in Production in a day when the number of workers in strike is 18,000 from the following information.
Mean number of workers on strike = 800
Mean loss of daily production in '000 Rs = 35
Standard deviation of number of workers on strike = 100
Standard deviation of daily production in '000 Rs = 2
The coefficient of correlation between number of workers on strike and daily production was == 0.80

41. The equation of two regression lines between two variables are expressed as $3X - 4Y + 30 = 0$ and $5Y - 2X + 8 = 0$.
   i. Identify which of the two can be called regression equation of Y on X and X on Y.
   ii. Find the mean of X and Y and correlation coefficient.

42. In a partially destroyed record of the following data available,
Variance of variable X = 25
Two regression equations are:
$5X - Y - 22 = 0$ and $64X - 45Y - 24 = 0$ Find the
   i. Mean value of variable X and Y
   ii. Coefficient of correlation between variables X and Y
   iii. Standard deviation of Variable Y.

43. The following table gives the ages and blood pressure of 10 women

Age	56	42	36	47	49	42	60	72	63	55
Pressure	14	12	11	12	14	14	15	16	14	15

Estimate the blood pressure of a women whose age is 45 year.

44. From the data given below,

Marks in Research	25	28	35	32	31	36	29	38	34
Marks in Probability	43	46	49	41	36	32	31	30	33

Find,

i. Two regression equations

ii. The coefficient of correlation between the marks in probability and research methodology.

iii. The most likely marks in Probability when a mark in research is 30.

45. From the following data between age of husbands and wives. Calculate the two regression equations and find the husbands age when wife's age is 20.

Wife's age	18	20	22	23	27	28	30
Husband's age	23	25	27	30	32	31	35

Answers:

8. Positive
9. r = 0.40
10. r = 0.836
11. r = 0.957
12. r = 0.725
13. r = 0.892, High
14. r = 0.7804
15. r = – 0.92
16. r = 0.704
17. r = 0.7225
18. r = – 0.438
19. r = 0.394
20. r = – 0.699
21. $r_{Corrected} = 0.77$
22. Lower Limit = 0.631 and Upper Limit = 0.769
23. PE = 0.054, Lower Limit = 546 and Upper Limit = 0.654
24. Yes, R= 0.25
25. R = 0.46
26. R = 0.8232
27. R12 = – 0.212, R13 = 0.636, R23 = – 0.297, 1st and 3rd
28. R = – 0.405
29. R = – 0.721
30. R = 0
31. R = 0.125
32. R = 0.8606
33. R = 0.545
34. R = 0.722
35. $R_c = 0.606$
36. $\hat{Y} = 29 + 0.67\ X$
37. X = 5.5 + 0.5y And Y = – 11 + 2x
38. 155 cm
39. Rs. 78.46
40. $\hat{Y} = 0.016\ X + 22.2$, Rs. 310200
41. i. 3X – 4Y + 30 = 0 is Y on X and 5Y - 2X + 8 = 0 is X on Y
ii Mean of X = 2 and mean of Y = 9 and r = 0.5477
42. i. 6 and 8 ii. 0.533 iii. 13.33
43. $\hat{Y} = 83.756 + 1.11X$ and Blood pressure = 133.708
44. i. $\hat{X} = 40.88 – 0.2337\ Y$ and $\hat{Y} = 59.146 – 0.664X$
ii. r = - 0.394 iii. 39.23 year
45. Husband's age = 25.34

Chapter 8

Theory of Probability

Origin of Probability

An Italian mathematician, Galileo (1564-1642) was the first man to attempt at a quantitative measure of probability while dealing with some problems related to the theory of dice in gambling.

Two French mathematicians, B. Pascal (1623-62) and P. Fermat (1601-65) were laid the first foundation of the mathematical theory of probability. Other main contributors are : T. Bayes, P.S Laplace (1749 – 1827) who after extensive research, over a number of years finally published 'Theory of inverse probability'. In addition to these, other outstanding contributors are Levy, Mises and R.A. Fisher.

Russian mathematician also have made very valuable contribution to modern theory of probability. Chief contributors, to maintain only a few of them are Chebyshev (1821 – 1894) who founded the Russian school of Statisticians, A. Markoff (1856– 1922), Liapoundoff (central limit theorem), A. Khintchine (law of large numbers) and A. Kolmogorov who axiomised the calculus of probability.

The 'probability' is the 'chance' or 'possibility' of occurrence or non-occurrence of any event under study. The words 'possibility', 'chance', 'probable', 'odds' etc are used in the sense of "uncertainty". Some examples where these terms are used are, it is not sure that I will attain the class. The chances of winning the game of the two teams A and B are equal. It is possible that he may not be able to join the B.E. program. etc.

A numerical measure of uncertainty of an event is known as probability with values from 0 to 1. In engineering the use of probability is very essential in every field where there is risk bearing such as calculation of flood distribution, estimate of earthquake recurrence time, to calculate the average number of telephone calls in a certain switching point and many more.

Definition of various terms

Before defining probability we should understand the following important terms related to probability.

Experiment

An action or operation which produces results or outcomes is called an experiment. In other word experiment is the process of observing a phenomenon that has variation in its outcomes. When we perform experiments in science and engineering repeatedly under identical condition, we get almost the same results.

experiments in science and engineering repeatedly under identical condition, we get almost the same results.

7.2.2 Random Experiment

If in each trial of an experiment conducted under homogenous condition the outcomes is not unique but may be any one of the possible outcomes. Then such an experiment is called a random experiment, in a random experiment, the outcomes of each trial depends on a chance which can't be predicted with certainty.

In other words an experiment is said to be random experiment if it satisfies the following three conditions the first condition is that all outcomes of an experiment is known in advance, the second condition is the outcomes of each trial depend on a chance which can't be predicted with certainty and the third condition is that the experiment could be conducted under identical and homogenous condition.

For example, if we toss a fair coin we may get either head or tail. If we through an unbiased die which has six faces marked with 1, 2, 3, 4, 5, 6, then any one of the faces may come up, it is not possible to say which face will come up.

7.2.3 Outcomes

The results of a random experiment are called outcomes.

7.2.4 Trial and Event

Performing a random experiment is called a trial and the collections of possible outcomes are termed as events. For example; if a coin is tossed repeatedly, the result is not unique, we may get any of the two faces head or tail. Thus tossing of a coin is a random experiment or trial and getting of a head or tail is an event. Similarly, throwing of a die is a trial and getting any one of the faces 1, 2, 3, 4, 5, 6 is an event. In the same experiment the possible events could also be stated as 'odd number of points', 'even number points', 'getting a point greater than 3' and so one.

7.2.5 Sample Space

The set of all possible outcomes of a random experiment is called sample space. The possible outcomes are called sample points. The sample space is usually denoted by the letter 'S'.

For example: In a tosses of a single fair coin the sample points are head(H) and Tail (T) and sample space is S = {H, T}

i. If two unbiased coins are tosses simultaneously , the sample space is S = { HH, HT, TH, TT}

ii. If three unbiased coins are tosses simultaneously , the sample space is S = { HHH, HHT, HTH, THH, TTH, THT, HTT, TTT}

iii. In throw of single fair die the sample space is S = { 1, 2, 3, 4, 5, 6}

iv. In throw of two fair dice the sample space is

$$S = \begin{Bmatrix} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{Bmatrix}$$

7.3 TYPES OF EVENTS

7.3.1 Simple or Compound Events

An event having only one sample point is called simple or elementary event. Otherwise it is known as compound or composite event. Thus in tossing of a single die the event of getting '5' is a simple event but the event of getting an odd number is a compound event.

7.3.2 Mutually Exclusive Events

Events are said to be mutually exclusive or disjoint if the occurrence of one precludes the occurrence of other i.e. one and only one of them can occur in a single trial. For example in tossing of a fair coin, the events head and tail are mutually exclusive events since if the coin fall with head upside tail can't turn up and vice-versa.

There is one crucial question to ask in deciding whether events are really mutually exclusive or not. If two or more events can occur together at a time they are not mutually exclusive. For example, if we draw either a king or a diamond from a pack of 52 cards then event king and diamond can occur together because we could draw the king of diamond, the king and diamond are not mutually exclusive.

7.3.3 Exhaustive Events

The total number of possible outcomes of a random experiment is known as the exhaustive number of event. For example, in tossing of a fair coin exhaustive cases are head and tail. In tosses of two fair coin at a time the exhaustive cases are (HH, HT, TH and TT)

7.3.4 Favourable Events

The number of cases Favourable to an event in a trial is the number of outcomes which entail the happening of the event. For example in drawing a card from a pack of cards the number of cases Favourable to drawing of a king is 4, for drawing of diamond is13 and for drawing a black card are 26.

7.3.5 Equally Likely Events

Outcomes of a trial are said to be equally likely if none of them is expected to occur in preference to other i.e. happening of all event has equal and independent chance. For example in throwing an unbiased die, all the six faces are equally likely to come; in a random toss of a fair coin head and tail are equally likely.

7.3.6 Independent and Dependent Events

Two or more events are said to be independent if the occurrence or non-occurrence of one event doesn't affect the occurrence or non-occurrence of the other events. Otherwise said to be dependence. For example, if we toss 2 coins together, the occurrence of head or tail on first coin doesn't affect the occurrence of head or tail on second coin. let E_1 be the event of occurrence of head or tail on first coin and E_2 be the event of occurrence of head or tail on second coin , so E_1 and E_2 are two independent events.

If we draw a card from a pack of well-shuffled cards and replace it before drawing the second card, the result of the second draw is independent of the first draw, but, however, if the first card drawn is not replaced then the second draw is dependent on the first draw.

7.4 BASIC PRINCIPLE OF COUNTING

If an event can occur in 'n' different way and corresponding to each of these 'n' ways a second event can occur in 'm 'different way. Then the two events together can occur in 'm x n' different ways, this rule can be extended to any number of events.

Example 1

There are 12 buses running between two places A and B , In how many ways can a man go from A to B and return by a different buses.

Solution

as the man go by any one of the 12 ways. Since he can't return by the same bus, he can return by an 11diffrernt ways
Therefore the total number of ways in which he can go to B and return
= 12 x 11 = 132 ways.

Example 2

Three students enter in the Class, whether there are 6 vacant seats. In how many ways can they seat themselves?

Solution

Clearly, the first student can occupy any of the 6 seats so; there are 6 ways in which the first student can sit himself. Now , the second student can occupy any of the remaining 5 seats , so , he can sited in 5 ways , similarly , the third student can occupy a seat in 4 ways

Hence, by the fundamental theorem of counting the required number of ways = 6 x 5 x 4 = 120 ways.

7.4.1 Factorial Notation

Let 'n' be a natural number. Then the continued product of natural number from 1 to n is called factorial n and it is denoted by n! .Also we define 0! = 1.and n! = 1 x 2 x 3 x.....(n-2) x (n-1) x n.

Here n! = n (n-1)!

7.4.2 Permutation and Combination

A permutation of a given set of thing is an arrangement of some or all of them in some order. When 3 letter p, q& r are arranged by taking two at a time there will be 6 permutation as pq, qr, rp, qp, rq, pr.
The process of selecting things from set is called combination, any different groups are formed and the order of the thing is not considered. For example combination of 3 letters p, q and r taking two at a time is pq, qr and rp.
The permutation of 'n' different things taking 'r' at a time is denoted by ${}^nP_r$

$$\text{or } P(n, r) = \frac{n!}{(n - r)!}$$

Similarly, the combination of 'n' different things taking 'r' at a time is denoted by ${}^nC_r$ or $C(n, r) = \frac{n!}{(n - r)!\, r!}$

Example 3

How many numbers of 5 different digits can be formed with digit 1, 2, 3, 4, 5, 6?

Solution

Here n = 6 and r = 5

$${}^nP_r = {}^6P_5 = \frac{6!}{(6 - 5)!} = \frac{6!}{1!} = 720 \text{ ways.}$$

Example 4

In how many ways can a committee of 3 persons be chosen out of 7?

Solution

Here n = 7, r = 3 and ${}^7C_3$ = ?

Hence required numbers of ways = ${}^nC_r = \frac{n!}{(n-r)!\ r!}$

$\therefore {}^7C_3 = \frac{7!}{3! \times (7-3)!} = \frac{7!}{3! \times 4!} = 35$ ways

7.5 DEFINITION OF PROBABILITY

The probability can be defined as a measure of a likelihood that a particular event will occur; it is a numerical measure of chance factor which value lies between 0 and 1.

The following is the broad classification of the different schools of thought in probability.

i. Classical or mathematical or Priori approach.
ii. Relative frequency or statistical or Empirical approach.
iii. Axiomatic Approach
iv. Subjective approach.

7.5.1 Classical or Mathematical or Priori Approach

If S is a sample space having 'n' points which arise out of all possible or exhaustive , equally likely and mutually exclusive outcomes of a random experiment and 'm' points in a S are Favourable to an event 'A', then the probability of the event A is given as

$$P(A) = \frac{\text{Favourable Cases}}{\text{Total exhaustive number of Cases}} = \frac{m}{n};\ 0 \le m \le n.$$

This definition was given by James Bernoulli, Who was the first person to obtain a quantitative measure of uncertainty.

Properties of Classical probability of an event A

i. It is clear from the definition of probability that n and m are positive integer , thus , we have

$0 \le m \le n.$

$0 \le \frac{m}{n} \le 1$

$\therefore 0 \le p(A) \le 1$, hence the probability of an event A is a number which is always lies between 0 and 1.

ii. If P(A) = 0, then A is called an impossible or null event and
If P(A) = 1, then A is called a certain or sure event.

iii. Probability is a pure number, it has no unit.

iv. Let $\bar{A}$ be event of non- occurrence of an event A then,

$$P(\bar{A}) = \frac{\text{Number of favourable cases for non-occurence of A}}{\text{Total cases}}$$

$$= \frac{n-m}{n} = 1 - \frac{m}{n} = 1 - P(A)$$

$\therefore P(A) + P(\bar{A} = 1$, ie total probability is always equal to unity.

v. We can compute the probability of an event A, i.e. P (A) by logical reasoning without conducting any experiment, since, the probability, P (A) can be computed prior to obtaining any experimental data, and it is also termed as a priori or mathematical probability.

Limitations of Classical Definition of Probability

Classical probability approach can't be adopted if any of the following situations exists.

i. If it is not possible to enumerate all possible outcomes. For example, the sample space of all even numbers.
ii. If the outcomes are not independent.
iii. If each and every outcomes are not equally likely.

7.5.2 Statistical or Empirical or Relative Frequency Approach

The definition of statistical probability goes after the name of *Richard* Ven Mises. Many limitations of classical definition could be removed by introducing the relative frequency approaches. It is defined as "If an experiment is performed repeatedly under essentially homogenous and identical conditions, if out of a large number of trials. Only 'n' are conducted under essentially homogenous and identical condition desultory into n outcomes , out of n outcomes m are Favourable to an event A , then the probability of A is defined as

$$P(A) = \lim_{n \to \infty} \left(\frac{m}{n}\right)$$

Since; in the relative frequency approach, the probability is obtained objectively by repetitive empirical observations, it is also known as 'Empirical probability'.

This definition removes some of the deficiencies of classical probability. But n is always, finite in physical experiments. Hence it is not definite that such a limit always exists.

7.5.3 Subjective Approach to Probability

The subjective probability is purely individualistic in nature in this approach; the probabilities of occurrence of an event are assigned by individuals and are based on their personal judgment, wisdom, intuition and expertise. These probabilities are called the subjective probability and

represent the degree of belief and the confidence one has in the occurrence of the respective event. Such probability can be calculated in all the cases though the probability assigned by different persons to the same event may vary from person to person.

Example 5

One of the cards of an ordinary deck of 52 cards is lost. What is the probability that a card lost is spade?

Solution

Here total number of possible cases when one card is lost (n) = 52 and Favourable number of cases for spade (m) = 13

$\therefore$ Probability that a lost card is spade

$$P(A) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{13}{52} = \frac{1}{4}$$

Example 6

An urn contains 4 white, and 6 black balls. What is the probability of drawing a black ball?

Solution

Here total number of exhaustive cases when one ball is drawn is (n) = 4+6 = 10, Also the event of interest is drawing a black ball, therefore Favourable cases for black ball (m) = 6

$$\therefore \text{Required Probability} = P(A) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{6}{10}$$

Example 7

In 150 tosses of a coin, 84 heads were observed. What is the probability of getting a tail in the next tosses?

Solution

Since 84 heads were observed in 150 tosses the remaining (150-84) = 66 outcomes must be tail, The probability of getting a tail in the next toss is

[$\because$ here N is large]

$$\therefore \text{Required Probability is } P(A) = \lim_{n\to\infty} \frac{m}{n} = \frac{66}{150} = 0.44$$

Example 8

What is the chance that a leap year selected at random will contains 53 Saturday.

Solution

In a leap year (366 days), there are 52 complete weeks and 2 days over. The following are the possible combinations for these two 'over' days:-

"(Saturday, Sunday), (Sunday, Monday), (Monday, Tuesday), (Tuesday, Wednesday), (Wednesday, Thursday), (Thursday, Friday), (Friday, Saturday)".

In order that a leap year selected at random should contain 53 Saturdays , one of the two 'over' days must be Saturday, since out of above 7 possibilities, 2 viz (i) Saturday, Sunday and (ii) Friday, Saturday are Favourable to this event

$$\therefore \text{Required Probability is } P(A) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} \quad \frac{m}{n} = \frac{2}{7}$$

Example 9

An ordinary fair die is rolled at random. Find the Probability of getting (i) a 4 (ii) Not 4 (iii) greater than 3 (iv) 2, 4, or 6

Solution

The sample space when a single die is rolled is S {1, 2, 3, 4, 5, 6}, therefore total number of possible outcomes (n) = 6

i. Total number of favorable cases of getting 4, (m) = 1

$$\therefore \text{required Probability } P(4) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{1}{6}$$

ii. Total number of favorable cases not getting 4 (m) = 5.

$$\text{Probability is } P(\text{not } 4) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{5}{6}$$

iii. Total number of favorable cases getting a number greater than 3, (m) = 3

$\therefore$ Required Probability = P (greater than 3)

$$= \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{3}{6} = \frac{1}{2}$$

iv. Total number of favorable cases for getting 2, 4, or, 6 (m) = 3

$$\therefore \text{Required probability} = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{3}{6} = \frac{1}{2}$$

Example 10

Three unbiased coins are tossed simultaneously .Find the probability of getting (i) exactly 3 head (ii) at least one head, (iii) one Head , (iv) at most one Head, (v) exactly 2 Head.

Solution

Here sample space when three coins are tossed simultaneously is given as;

S = {HHH, HHT, HTH, THH, HTT, THT, TTH and TTT}

$\therefore$ Total number of possible outcomes (n) = 8

i. Let A be the event of getting exactly 3 Head , Then A ={HHH},

$\therefore$ Total number of favorable cases getting exactly 3 Head (m) = 1

$\therefore$ Required Probability $P(A) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{1}{8}$

ii. Let B be the event of getting at least one head then

B = { HHH, HHT, HTH, THH, HTT, THT, TTH}

$\therefore$ Total number of favorable cases getting at least one head (m) = 7.

$\therefore$ Required Probability is $P(B) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{7}{8}$

iii. Let C be the event of getting one head then C ={TTH, THT, HTT}

$\therefore$ Total number of favorable cases getting one head (m) = 3

$\therefore$ Required Probability is $P(C) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{3}{8}$

iv. Let D be the event of getting at most one head then

D = { TTH, THT, HTT & TTT}

$\therefore$ Total number of favorable cases getting at most one head (m) = 4

$\therefore$ Required Probability is $P(D) = \frac{\text{Favourable Cases}}{\text{Total Number of Cases}} = \frac{m}{n} = \frac{2}{8}$

Example 11

Three electric bulbs are chosen at random from 15 bulbs of which 5 are defective. Find the probability that (i) None is defective, (ii) Exactly one is defective (iii) At least one is defective.

Solution

Three bulbs can be chosen from 15 bulbs in ${}^{15}C_3$ ways = 455 ways.

i. Again the number of selecting 3 non defective bulbs from 10 non defective bulbs = ${}^{10}C_3$ = 120 ways

$\therefore$ The probability that none is defective = $\frac{\text{Favourable Cases}}{\text{Total Cases}}$

$= \frac{120}{455} = \frac{24}{91}$

ii. The number of ways of selecting one defective from 5 defective bulbs and 2 non-defective from 10 non-defective bulbs = ${}^5C_1 \times {}^{10}C_2 = 225$

$\therefore$ The probability that exactly one defective = $\frac{\text{Favourable Cases}}{\text{Total Cases}}$

$= \frac{225}{455}$

$= \frac{45}{91}$

iii. The probability of at least one defective = 1 – probability of no defective

$= 1 - \frac{24}{91}$

$= \frac{67}{91}$

7.5.4 Axiomatic Probability

The modern theory of probability is based on the axiomatic approach introduced by the Russian Mathematician A.N. Kolmogorov in 1930's and it is defined as , For a given sample space (S) of a random experiment , the probability of occurrence of any event A is defined as a set function P(A) satisfying the following axioms."

i. Probability of an event A defined as P(A) is real and non-negative function i.e. $P(A) \geq 0$ (Axioms of positiveness).

ii. Probability of Sample space is unity i.e. P(S) =1 (Axiom of certainty)

iii. If $A_1, A_2 \ldots A_n \ldots$ is any finite or infinite sequence of disjoint event of sample space S then

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

7.6 THEOREMS OF PROBABILITY

We have two theorems of probability by which the calculation of probabilities can becomes easily and be facilitated to a great extent, the theorems of probabilities are

i. Additive theorem of probability

ii. Multiplicative theorem of probability

7.6.1 Additive Theorem of Probability

There are two cases arise for addition theorems or rules of probability, these are (i) when the cases are mutually exclusive and (ii) when the cases are not mutually exclusive

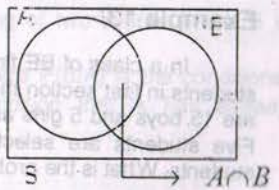
Case I: - Addition theorem of Probability for not mutually exclusive events.

Theorem A

The probability of occurrence of at least one of the two events A and B is given by $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Proof

Let us suppose that a random experiment results in a sample space S with N sample points, then by definition



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

Where n(A U B) is the number of occurrences favorable to the event (A U B) . From the adjoining diagram we get.

$$P(A \cup B) = \frac{[\{n(A) - n(A \cap B)\} + n(A \cap B) + \{n(B) - n(A \cap B)\}]}{N}$$

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

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

Case II: Addition theorem of probability for mutually exclusive events: If the events A and B are mutually disjoint ie $(A \cap B) = \phi$ then

$$P(A \cap B) = \frac{n(A \cap B)}{N} = \frac{n(\phi)}{N} = 0$$

$\therefore P(A \cup B) = P(A) + P(B)$.

Therefore probability of happening of any one of the two mutually exclusive events is equal to the sum of their individual probabilities.

This theorem can be extended to n mutually disjoint events as $P(A_1 \cup A_2 \cup \ldots\ldots\ldots\ldots \cup A_n) = P(A_1) + P(A_2) + \ldots\ldots\ldots + P(A_n)$

Example 12

If a card is drawn form a pack of cards, What is the probability of drawing a queen or king from pack of 52 cards ?

Solution

Let A be the event of drawing a queen and B be the event of drawing a king , then P(A) = the probability of drawing a queen $= \frac{m}{n} = \frac{4}{52} = \frac{1}{13}$

And P(B) = the probability of drawing a King $= \frac{m}{n} = \frac{4}{52} = \frac{1}{13}$.

Since the two events are mutually exclusive, the probability of drawing either a queen or king is given by

$$P(A \text{ or } B) = P(A \cup B) = P(A) + P(B) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

Example 13

In a class of BE first year. There are three sections each including 20 students in first section there are 10 boys are 10 girls, in second section there are 15 boys and 5 girls and in the third section there are 12 boys and 8 girls. Five students are selected from each group to form a committee of 15 students. What is the probability that all the 15 students selected are girls ?

Solution

Let A, B and C are the events that 5 students selected from section first, second and third are girls.

i. 5 students can be selected from 20 students in $^{20}C_5$ ways

$\therefore$ Total exhaustive cases (n) = $^{20}C_5 = 15504$

The Five girls students can be selected from 10 girls in $^{10}C_5$ ways

$\therefore$ Favourable cases of 5 girls students drawn form section first $(m_1) = {}^{10}C_5 = 252$

Therefore, Probability that the 5 girls students selected from section first $P(A) = \frac{m_1}{n} = \frac{252}{15504}$

ii. The five girls students can be selected from 5 girls in $^5C_5$ ways

$\therefore$ Favourable cases of 5 girls students drawn form section second $(m_2) = {}^5C_5 = 1$

Therefore, Probability that the 5 girls' students selected from section second $P(B) = \frac{m_2}{n} = \frac{1}{15504}$

iii. The five girls students can be selected from 8 girls in $^8C_5$ ways

$\therefore$ Favourable cases of 5 girls students drawn form section third $(m_3) = {}^8C_5 = 56$

Therefore, Probability that the 5 girls' students selected from section third $P(C) = \frac{m_3}{n} = \frac{56}{15504}$

Therefore, all the 15 students selected are girls is given as

$$P(A \cup B \cup C) = P(A) + P(B) + P(C)$$

$$= \frac{252}{15504} + \frac{1}{15504} + \frac{56}{15504}$$

$$= 0.01977$$

7.6.2 Multiplication Law of Probability

There are two cases (i) when events are independent and (ii) when the events are dependent.

Theorem B

The probability of simultaneous happening of two events A and B is given by $P(A \cap B) = P(A)\, P(B/A)$; $P(A) \neq 0$

$P(A \cap B) = P(B)\, P(A/B)$; $P(B) \neq 0$, where P(B/A) is the conditional probability of happening of B under the condition that A has already happened.

Proof

Let A and B be the events associated with the sample space S of a random experiment with exhaustive number of outcomes N ie n(S) = N, then

by the definition $P(A \cap B) = \frac{n(A \cap B)}{N}$. For the conditional event A/B, the Favourable outcomes must be out of the sample point B. In other words for the event A/B, the sample space is B and then $P(A/B) = \frac{n(A \cap B)}{n(B)}$,Similarly , we have $P(B/A) = \frac{n(B \cap A)}{n(A)}$

$$\therefore P(A \cap B) = \frac{n(A \cap B)}{N} = \frac{n(A)}{N} \times \frac{n(A \cap B)}{n(A)} = P(A) \times P(B/A)$$

$$\text{Similarly, } P(A \cap B) = \frac{n(A \cap B)}{N} = \frac{n(B)}{N} \times \frac{n(B \cap A)}{n(B)} = P(B) \times P(A/B)$$

For independent event

If A and B are independent, so that the probability of occurrence of or non–occurrence of A is not affected by the occurrence or non-occurrence of B, then we have P (A/B) = P (A) and P (B/A) = P (B)

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

Example 14

Each of the two Persons A and B throw a die, what is the Probability that A get 4 and B gets a multiplication of 3?

Solution

Let E_1 and E_2 are the event of getting 4 and a multiplication of 3 thrown a die by two Persons A and B respectively. Then $E_1=\{4\}$ and $E_2=\{3, 6\}$

$\therefore$ P (E_1) = the probability that A get 4 = $\frac{1}{6}$

And P (E_2) = the probability that B get a multiple of 3 = $\frac{2}{6}$.

Since these two event E_1& E_2 are independent then the required probability $P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B) = \frac{1}{6} \times \frac{2}{6} = \frac{2}{36} = \frac{1}{18}$

Example 15

A speaks truth 70% of cases and B in 85% of cases in what percentages of cases are they likely to contradict and don't contradict each other in stating the some facts

Solution

P (A speak truth) = 70% = 0.7 and P (A don't speak truth) = 0.3

P (B speak truth) = 85% = 0.857 and P (B don't speak truth) = 0.15

i. The probability that both are contradict to each others

$$P(A \text{ and } \bar{B} \text{ or } B \text{ and } \bar{A}) = P(A \cap \bar{B}) + P(B \cap \bar{A})$$

$$= P(A)\,P(\bar{B}) + P(B)\,P(\bar{A})$$

$$= 0.7 \times 0.15 + 0.85 \times 0.3$$

$$= 0.36$$

Thus, they contradict in 36% of cases

ii. The probability that both don't contradict to each other is

$$P(A \text{ and } B \text{ or } \bar{A} \text{ and } \bar{B}) = P(A \cap B) + P(\bar{A} \cap \bar{B})$$

$$= P(A)\,P(B) + P(\bar{A})\,P(\bar{B})$$

$$= 0.7 \times 0.85 + 0.3 \times 0.15$$

$$= 0.64$$

Thus, they don't contradict in 64% of the cases

7.6.3 Probability of Happening of at Least One Event

If there are n independent events and the probability of their occurrence are $p_1\ p_2 \ldots\ldots\ldots p_n$ then the probability of non of them happening is P(non of them) = $(1 - p_1)(1 - p_2)\ldots\ldots\ldots(1 - p_n)$. Hence, the probability of occurrence of at least one of them is;

P (at least one of them) = 1- P(non of them)

$$= 1 - [(1 - p_1)(1 - p_2) \ldots (1 - p_n)]$$

Example 16

A drawer contains 50 bolts and 150 nuts. Half of the bolts and half of the nuts are rusted. If one items is chosen at random. What is the probability that it is rusted or a bolt ?

Solution

Let, A be the event of drawing a rusted item and B be the event of drawing a bolt. Now, total number of bolts is 50 and number of rusted bolts is 25, total number of bolts is 150 and number of rusted nuts are 75.

So, total of items is 200 and total number of rusted items is 100. The events A and B are not mutually exclusive events

$$P(A) = P(\text{rusted items}) = \frac{100}{200} = \frac{1}{2}$$

$$P(B) = P(\text{Drawing bolts}) = \frac{50}{200} = \frac{1}{4}$$

$$P(A \cap B) = P(\text{Drawing rusted bolts}) = \frac{25}{200} = \frac{1}{8}$$

$$\therefore P(A \cup B) = P(\text{Drawing rusted or a bolt}) = P(A) + P(B) - P(A \cap B)$$

$$= \frac{1}{2} + \frac{1}{4} - \frac{1}{8} = \frac{5}{8}$$

Example 17

Rice and Milton are asked to solve a problem, the probability of Rice solving it is 2/3 and that of Milton solving it is ¾. Find the probability that at least one of them will solve it?

Solution

Let A be the event of solving problem by Rice and B be the event of solving problems by Milton. Then $P(A) = \frac{2}{3}$ and $P(B) = \frac{3}{4}$.

$\therefore P(\bar{A})$ = Probability that Rice can't solve Problem = $1 - P(A) = \frac{1}{3}$

And $P(\bar{B})$ = Probability that Milton can't solve Problem = $1 - P(B) = \frac{1}{4}$

$\therefore$ Required probability that at least one of them will solve it is

P(at least one of them) = 1– P(non of them solve it)

$\therefore P(\text{Non of them}) = P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = \frac{1}{3} \times \frac{1}{4} = \frac{1}{12}$

$\therefore P(\text{At least one of them solve it}) = 1 - \frac{1}{12} = \frac{11}{12}$.

Example 18

The probabilities are 0.7, 0.3 and 0.2 that a potential investor will purchase shares in Banking sector and non-banking sector or both. Find the probability that the investor will purchase shares in

i. At least one of the sectors.

ii. Neither of the sectors.

Solution

Let A be the event that the investor purchase Banking share and B be the event that the investor purchase non-banking share.

Then P (A) = 0.7and P (B) = 0.3 and $P(A \cap B) = 0.2$.

$\therefore$ P(Investor purchase at least one of the shares)

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.7 + 0.3 - 0.2 = 0.8$

$P(\bar{A})$ = Probability that the investor will not purchase banking share

$= 1 - P(A) = 1 - 0.7 = 0.3$

$P(\bar{B})$ = Probability that the investor will not purchase non-banking share $= 1 - P(B) = 1 - 0.3 = 0.7$

$\therefore$ Required probability that at neither of the share purchase by the investor $= P(\bar{A} \cap \bar{B}) = P(\bar{A}) \times P(\bar{B}) = 0.7 \times 0.3 = 0.2$

7.7 CONDITIONAL PROBABILITY

Many times the information is available that a event has occurred and one is required to find out the probability of occurrence of another event B, utilizing the information about A , such a probability is known as conditional probability and is denoted by P(B/A) i.e. the probability of the event B given A. For example, suppose we know that a accident will happen (A) then one is interested to know the probability of death of patient (B) i.e. we want to calculate P(B/A) ,

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

If A and B are independent, then

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

7.8 INDEPENDENT EVENT

Definition: An event A is said to be statistically independent of another event B if the conditional probability of A given B i.e. P (A/B) is equal to the unconditional probability of B i.e. If P(A/B) = P(A) provided that $P(B) \neq 0$. Similarly an event B is said to be statistically independent of event A if

P(B/A) = P(B)

Therefore, for independent events the probability of occurenceo of two events simultaneously is defined as

$P(A \text{ and } B) = P(A \cap B) = P(A)\,P(B)$

7.8.1 Theorems of INDEPENDENT

Theorem: If two events A and B are independence then prove that

i. A and $\bar{B}$ are independent

ii. $\bar{A}$ and B are independent

iii. $\bar{A}$ and $\bar{B}$ are independent

Proof

If two events A and B are independent then we have

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

i. Taking $P(A \cap \bar{B}) = P(A) - P(A \cap B)$

$= P(A) - P(A)\,P(B)$

$= P(A)\{1 - P(B)\}$

$= P(A)\,P(\bar{B})$

Which, implies that the events A and $\bar{B}$ are independent

ii. Taking $P(\bar{A} \cap B) = P(B) - P(A \cap B)$

$= P(B) - P(A)\,P(B)$

$= P(B)\{1 - P(A)\}$

$= P(B)\,P(\bar{A})$

Which, implies that the events $\bar{A}$ and B are independent

iii. Taking $P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B})$

$= 1 - P(A \cup B)$

$= 1 - \{P(A) + P(B) - P(A \cap B)\}$

$= 1 - P(A) - P(B) + P(A)\,P(B)$

$= \{1 - P(A)\} - P(B)\{1 - P(A)\}$

$= P(\bar{A})\,P(\bar{B})$

Which, implies that the events $\bar{A}$ and $\bar{B}$ are independent

Example 19

Two Urns contains 3 white, 7 red, 15 black and 10 white, 6 red, 9 black balls respectively. One ball is drawn from each Urn. Find the probability that both the balls are of the same colour.

Solution

Let W, R and B denote respectively the event of a white, red and black ball being drawn from the Urn so.

$P(W_1) = 3/25$; $P(W_2) = 10/25$

$P(R_1) = 7/25$; $P(R_2) = 6/25$

$P(B_1) = 15/25$; $P(B_2) = 9/25$

Now P(both balls are white) $= P(W_1 \cap W_2)$

$= P(W_1).P(W_2) = 3/25 \times 10/25 = 30/625$

[Since W_1 and W_2 are independent]

Similarly, P(both balls are Red) $= P(R_1 \cap R_2)$

$= P(R_1).P(R_2) = 7/25 \times 6/25 = 42/625$

[Since R_1 and R_2 are independent]

And P(both balls are black) $= P(B_1 \cap B_2) = P(B_1).P(B_2)$

$= 15/25 \times 9/25 = 135/625$

[Since B_1 and B_2 are independent]

Hence the required probability is

P (both balls are of same colour)

$P\{(W_1 \cap W_2) \cup [R_1 \cap R_2) \cup (B_1 \cap B_2)\}$

$= P[(W_1 \cap W_2) + P[R_1 \cap R_2) + P(B_1 \cap B_2)$

[since $(W_1 \cap W_2)$, $(R_1 \cap R_2)$ & $(B_1 \cap B_2)$ are mutually exclusive]

$= 30/625 + 42/625 + 135/625 = 207/625$

Example 20

The following information was obtained concerning 1000 employees of an industrial concern.

Department	Gender		Total
	Male	Female	
Manufacturing	280	220	500
Production control	175	125	300
Quality Control	115	85	200
Total	570	430	1000

If an employee is selected at random, what is the probability that

i. Employee selected is male given that belongs to production control department.

ii. Employee chosen is Female from manufacturing department.

Solution

Let M denotes the event that employee chosen is male.

Let P denotes the event that employee chosen from production control department.

Let F denote the event that employee chosen is Female.

Let D denote the event that employee chosen from manufacturing department.

i. $P(M/P) = \frac{P(M \cap P)}{P(P)} = \frac{175}{1000}$

$= \frac{175}{300} = 0.583$

$\therefore$ P(Male/Production Control) = 0.583

ii. $P(F \cap D) = \frac{220}{1000} = 0.22$

Example 21

The probability that a management trainee will remain with the company is 0.6 and the probability that the trainee will earn more than Rs. 5,000 per year is 0.2. The probability that an employee is a management trainee who remained with the company or who earned more than Rs. 500 per year is 0.5. What is the probability that an employee earned more than 5,000 per year given that he remained with the company ?

Solution

Let R = event of a management trainee will remain with the company.

M = event of a trainee will earn more than Rs 5,000 per year.

Therefore,

$P(R) = 0.6$ $P(M) = 0.2$

$P(R \cup M) = 0.5$ $P(M/R) = ?$

We know,

$P(R \cup M) = P(R) + P(M) - P(R \cap M)$

$0.5 = 0.6 + 0.2 - P(R \cap M)$

$\therefore P(R \cap M) = 0.3$

$\therefore$ required probability

$$P(M/R) = \frac{P(R \cap M)}{P(R)} = \frac{0.3}{0.6} = 0.5$$

Example 22

In a survey of 1000 customers about the quality and timely, deliverance of "The momo world" Marketed by a company following information prevailed.

		Timely deliverance		Total
		Yes	No	
Quality Satisfaction	Yes	500	225	725
	No	200	75	275
	Total	700	300	1000

i. What is the probability that a customer satisfied with quality but timely not delivered ?

ii. What is the probability that a customer timely delivered but not satisfied with quality ?

iii. What is the probability that a customer timely delivered and satisfied with quality ?

Solution

Let T = the event of selecting the customers timely delivered.

S = The event of selecting the customers satisfied with quality.

i. $$P(S/TN) = \frac{P(S \cap TN)}{P(TN)} = \frac{\frac{225}{1000}}{\frac{300}{1000}} = 0.75$$

ii. $$P(T/NS) = \frac{P(T \cap NS)}{P(NS)} = \frac{\frac{200}{1000}}{\frac{275}{1000}} = \frac{200}{275} = 0.75$$

iii. $$P(T \cap S) = \frac{500}{1000} = 0.5$$

Example 23

In an examination 70% of students have passed in Hydraulics, 30% of the students have passed in Hydrology and 20% have passed in both. A student is selected at random. What is the probability that the student has passed in Hydraulics of it in known that he has passed in Hydrology.

Solution

Let A and B be the events that a student selected at random pass in Hydraulics and pass in Hydrology respectively.

Then $P(A) = 0.7$ $P(B) = 0.3$ $P(A \cap B) = 0.2$

The probability that student selected has passed in Hydraulics given that he has passed in Hydrology is

$$P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0.2}{0.3} = \frac{2}{3}$$

7.8.2 The Rule of Total Probability

The general multiplication rule leads to an alternative rule called the rule of total probability. An event A can occur either when an event B occurs or when it doesn't occur. That is, A can be written as the disjoint union of two events $A \cap B$ and $A \cap \bar{B}$

Consequently, $P(A) = P(A \cap B) + P(A \cap \bar{B})$

$$= P(B)\,P(A/B) + P(\bar{B})P(A/\bar{B}),$$

This is obtained by the general multiplication rule applied to each of the two terms involving the probabilities of intersections.

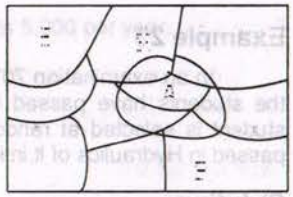
7.9 BAYES'S THEOREM

Bayes' Probability is also known as inverse probability. The idea of inverse probability was given by Sir Thomas Bayes in 1763 which was reprinted in Biometrika in the year 1958. The problem of inverse probability arises when we have an outcome and we want to know the probability of its belonging to specific trial or population out of many alternative trials or population.

Statement: If $E_1, E_2, \ldots\ldots E_n$ are n mutually exclusive and exhaustive event of sample space "S" such that $\bigcup_{i=1}^{n} E_i = S$ and $P(E_i) > 0$ for i = 1, 2 .....n And if A is any arbitrary event which is associated with Ei's such that $P(A) > 0$. Then

the probability that A has occurred and given by the event Ei is given as P(Ei/A) i.e.

$$P(E_i/A) = \frac{P(A/E_i)\,P(E_i)}{\sum_{i=1}^{n} P(A/E_i)\,P(E_i)}$$



Proof

The events $E_1, E_2 \ldots\ldots E_n$ are mutually exclusive events and A is any arbitrary event which is associated with E_i's and $P(A)>0$ & $P(E_i)>0$. Then from the conditional probability of two events A and E_i

$P(A \cap E_i) = P(A) \times P(E_i/A) = P(E_i) \times P(A/E_i)$

Or $P(E_i/A) = \frac{P(Ei) \times P(A/E_i)}{P(A)}$ ... (i)

The event A is defines as

$A = (A \cap E_1) \cup (A \cap E_2) \cup \ldots. \cup (A \cap E_n)$,

Taking probability on both sides we get

$P(A) = P(A \cap E_1) + P(A \cap E_2) + \ldots + P(A \cap E_n)$,

Since each E_i's are mutually exclusive and thus

$(A \cap E_1), (A \cap E_2) \ldots (A \cap E_n)$ are also mutually exclusive.

$\therefore P(A) = P(E_1) \times P(A/E_1) + P(E_2) \times P(A/E_2) + \ldots + P(E_n) \times P(A/E_n)$

$$P(A) = \sum_{i=1}^{n} P(E_i) \times P(A/E_i) \quad \ldots (ii)$$

From equation (ii) equation (i) becomes

$$P(Ei/A) = \frac{P(A/E_i)\,P(E_i)}{\sum_{i=1}^{n} P(A/E_i)\,P(E_i)}$$

For 3 mutually exclusive and exhaustive events,

$$P(E_1/A) = \frac{P(E_1)\,P(A/E_1)}{P(E_1)\,P(A/E_1) + P(E_2)\,P(A/E_2) + P(E_3)\,P(A/E_3)}$$

$$P(E_2/A) = \frac{P(E_2)\,P(A/E_2)}{P(E_1)\,P(A/E_1) + P(E_2)\,P(A/E_2) + P(E_3)\,P(A/E_3)}$$

$$P(E_3/A) = \frac{P(E_3)\,P(A/E_3)}{P(E_1)\,P(A/E_1) + P(E_2)\,P(A/E_2) + P(E_3)\,P(A/E_3)}$$

Example 24

In Kathmandu city, 25% of all cars emits excessive amount of pollutions, If the probability is 0.99 that a car emitting excessive amount of pollutants will fail the city's vehicular emission test and the probability is 0.17 that a car not emitting excessive amount of pollutions will not nevertheless fail the test. What is the probability that a car which fails the test actually emits excessive amount of pollutants.

Solution

Let E_1 be the events that cars emits excessive amount of pollution and E_2 be the events that cars not exits excessive amount of pollution. Then according to question $P(E1) = 25\% = 0.25$and $P(E_2) = 0.75$

And A is the event that car fail the city's vehicular emission test. Then $P(A/E_1) = 0.99$ and $P(A/E_2) = 0.17$. The car which fail the test actually emits excessive amount of pollutant i.e. $P(E_1/A) = ?$

Then form Bayes's theorem

$$\therefore P(E1/A) = \frac{P(A/E_1)\,P(E_1)}{P(A/E_1)\,P(E_1) + P(A/E_2)\,P(E_2)}$$

$$= \frac{0.25 \times 0.99}{0.25 \times 0.99 + 0.17 \times 0.75} = 0.66$$

Example 25

In a college, there will be three candidates for the position of Principal. Mr. Chettri, Mr. Thapa and Dr. Singh whose chance of getting the appointment are in the proportion 4:2:3 respectively. The probability that Mr. Chettri would start electronics department in the college is 0.3. The probabilities of Mr. Thapa and Dr. Singh doing the same are respectively 0.5 & 0.8 (i) what is the probability that there will be electronics department in the college? (ii) If there is electronics department in next year, what is the probability that Dr. Singh is the Principal?

Solution

Let us defined the following events, A be the event that to start of electronics department, E_1 is that Mr. Chettri is selected as Principal, E_2 is that Mr. Thapa is selected as Principal, E_3 is that Dr. Singh is selected as Principal, then we are given $P(E1) = 4/9$, $P(E_2) = 2/9$ and $P(E_3) = 3/9$. $P(A/E_1) = 0.3$, $P(A/E_2) = 0.5$, $P(A/E_3) = 0.8$

i. The required Probability that there will be electronics department in the college in next year is given by

$P(A) = P(A \cap E_1) + P(A \cap E_2) + P(A \cap E_3)$

$= P(E_1) \times P(A/E_1) + P(E_2) + \times P(A/E_2) + P(E_3) \times P(A/E_3)$

$= 4/9 \times 0.3 + 2/9 \times 0.5 + 3/9 \times 0.8$

$= 23/45.$

ii. The required probability that there is electronics department in next year that Dr. Singh is the Principal is given by Bayes's theorem

$$P(E_3/A) = \frac{P(E_3) \times P(A/E_3)}{P(A)}$$

$$= \frac{3/9 \times 0.8}{23/45} = \frac{12}{23}$$

Example 26

In a test, an examinee either guesses or copies or knows the answer to multiple choice questions with four choices. Only one answer being correct, the probability that he makes a guess is $\frac{1}{3}$ and the probability that he copies the answer is $\frac{1}{6}$. The probability that his answer is correct, viven that he coies it is $\frac{1}{8}$. Find the probability that he know the answer to the question, given that he correctly answers it.

Solution

Let, A: the event of answering by guess work.

B: the event of answering by coping.

C: the event of answering by knowing.

D: the event of answering correctly.

Then $P(A) = \frac{1}{3}$, $P(B) = \frac{1}{6}$

As the question is answered by either guessing or copying or knowing

$$P(C) = 1 - P(A) - P(B) = 1 - \frac{1}{3} - \frac{1}{6} = \frac{1}{2}$$

Now, P(D/A) = Probability of answering correctly by guessing $= \frac{1}{4}$

[Since there are 4 choice of only one of them is correct]

P(D/B) = probability of answering correctly by copying $= \frac{1}{8}$, given

P(D/C) = probability of answering correctly by knowing = 1

[Since if he know, he certainly answer correctly]

We have to find the probability that he know when he answered correctly i.e. P(C/D) by using Bayes's theorem as

$$P(C/D) = \frac{P(C) \times P(D/C)}{P(A) \times P(D/A) + P(B) \times P(D/B) + P(C) \times P(D/C)}$$

$$= \frac{0.5 \times 1}{\frac{1}{3} \times \frac{1}{4} + \frac{1}{6} \times \frac{1}{8} + \frac{1}{2} \times 1} = \frac{24}{29}$$

Example 27

There are three machines A, B and C producing 1000, 2000 and 3000 articles per hour respectively. These machines are known to be producing 1%, 2% and 3% defectives respectively. one article is selected at random from an hour production of the three machines and found to be defective. What is the probability that the article is produced from (a) Machine A (b) Machine B (c) Machine C.

Solution,

Let E_1 be the event of drawing an article produced by the machine A
E_2 be the event of drawing an article produced by the machine B.
E_3 be the event of drawing an article produced by the machine C. and
D be the event of drawing a defective article produced by either machine A or B or C.

n = 1000 + 2000 + 3000 = 6000, Total number of articles.

Then,

$$P(E_1) = \frac{1000}{6000} = \frac{1}{6}$$

$$P(E_2) = \frac{2000}{6000} = \frac{1}{3}$$

$$P(E_3) = \frac{3000}{6000} = \frac{1}{2}$$

$$P(D/E_1) = 1\% = \frac{1}{100}$$

$$P(D/E_2) = 2\% = \frac{2}{100}$$

$$P(D/E_3) = 3\% = \frac{3}{100}$$

By using Bayes's Theorem

$$P(E_1/D) = \frac{P(E_1)\ P(D/E_1)}{P(E_1)\ P(D/E_1) + P(E_2)\ P(D/E_2) + P(E_3)\ P(D/E_3)}$$

$$= \frac{\frac{1}{6} \times \frac{1}{100}}{\frac{1}{6} \times \frac{1}{100} + \frac{2}{6} \times \frac{2}{100} + \frac{3}{6} \times \frac{3}{100}}$$

$$= \frac{\frac{1}{600}}{\frac{14}{600}} = \frac{1}{14}$$

$$P(E_2/D) = \frac{P(E_2)\ P(D/E_2)}{P(E_1)\ P(D/E_1) + P(E_2)\ P(D/E_2) + P(E_3)\ P(D/E_3)}$$

$$= \frac{\frac{2}{6} \times \frac{2}{100}}{\frac{14}{600}} = \frac{4}{14}$$

$$P(E_3/D) = \frac{P(E_3)\ P(D/E_3)}{P(E_1)\ P(D/E_1) + P(E_2)\ P(D/E_2) + P(E_3)\ P(D/E_3)}$$

$$= \frac{\frac{3}{6} \times \frac{3}{100}}{\frac{14}{600}} = \frac{9}{14}$$

Example 28

Two urns contain 1 white 6 red and 4 white, 3 red balls one of the urn is selected at random and a ball is drawn from it. Find

i. The probability of drawing a white ball.

ii. The probability of drawing the ball from first urn of the ball drawn is white.

Solution

Probability of random selection of urn = $P(E_1) = P(E_2) = \frac{1}{2}$

Probability of white ball from First urn = $P(A/E_1) = \frac{1}{7}$

Probability of white ball from a 2nd urn = $P(A/E_2) = \frac{4}{7}$

i. Probability of drawing a white ball = P(A)

$$= P(E_1)\ P(A/E_1) + P(E_2)\ P(A/E_2)$$

$$= \frac{1}{2} \times \frac{1}{7} + \frac{1}{2} \times \frac{4}{7}$$

$$= \frac{5}{14}$$

ii. The probability of drawing a white ball from first urn is

$$P(E_1/A) = \frac{P(E_1) \times P(A/E_2)}{P(E_1) \times P(A/E_1) + P(E_2) \times P(A/E_2)}$$

$$= \frac{\frac{1}{2} \times \frac{1}{7}}{\frac{5}{14}} = \frac{\frac{1}{14}}{\frac{5}{14}} = \frac{1}{5}$$

Example 29

In a bolt Factory, machines A, B and C manufacture 60%, 25% and 15% respectively. OF the total of their output 1%, 2% and 1% are defective bolts. A bolt is drawn at random from the total production and found to be defective. From which machines the defective bolt is expect to have been manufactured ?

Let E_1, E_2 and E_3 bolt events of choosing order A, B and C respectively,

Then,

P(E1) = 60% = 0.6

P(E2) = 25% = 0.25

P(E3) = 15% = 0.15

Let A be the events producing defective items, then

$P(A/E_1) = 1\% = \frac{1}{100}$

$P(A/E_2) = 2\% = \frac{2}{100}$

$P(A/E_3) = 1\% = \frac{1}{100}$

To find which machine produce more defective bolt we need to find out the value of P(E1/A) , P(E2/A) and P(E3/A)

$$P(E_1/A) = \frac{P(E_1)\ P(A/E_1)}{P(E_1)\ P(A/E_1) + P(E_2)\ P(A/E_2) + P(E_3)\ P(A/E_3)}$$

$$= \frac{\frac{60}{100} \times \frac{1}{100}}{\frac{60}{100} \times \frac{1}{100} + \frac{25}{100} \times \frac{2}{100} + \frac{15}{100} \times \frac{1}{100}}$$

$$= \frac{\frac{60}{10000}}{\frac{60 + 25 + 15}{10000}} = \frac{60}{100} = \frac{3}{5} = 0.6$$

Exercise - 8

1. What do you understand by mutually exclusive and independent events?
2. Elucidate the approaches to probability and explain the classical approach to probability.
3. What do you understand by axiomatic definition of probability? What are the axioms of probability?
4. State and prove the addition theorem of probability.
5. What is a conditional probability and independent event?
6. State and proof the Bayes's theorem of probability.
7. A restaurant has 6 different desserts. Find the number of ways a customer can choose 2 of the desserts.
8. A store has 8 different mystery books. Find the number of ways a customer can buy 3 of the books.
9. A box contains 6 blue socks and 4 white socks. Find the number of way when two socks can be drawn from the box. (i) There are no restrictions. (ii) they are different colour, (iii) they are to be the same colour
10. A class contains 9 boys and 3 girls. Find number of ways a teacher can select a committee of 4.
11. A woman has 11 close friends. Find the number of ways she can invite 5 of them to dinner.
12. Two unbiased dice are thrown. Find the probability that (i) the first dice shown6, (ii) the total of the numbers on the dice is 8, (iii) the total of the numbers on the dice is greater than 8, (iv) the total of the numbers on the dice is 13 and greater, (v) the total of the numbers on the dice is any number from 2 to 12.
13. Four cards are drawn from a pack of 52 cards. Find the probability that. (i) They are a king, a queen, a Jack and a ace, (ii) Two are Kings and two are Queens, (iii) Two are black and two are red, (iv) There are two cards of hearts and two cards of diamonds.
14. A committee of 4 students is to be appointed from 3 students of the civil department, 4 students of the computer department 2 students of the Electronics and communication department and 1 from architecture department. Find the probability of forming the committee in the following manner. (i) There must be one from each department, (ii) it should have at least one from the computer department, (iii) The architecture must be in the committee,

15. An urn contains 6 white, 4 red and 9 black balls, if 3 balls are drawn at random , find the probability that
   i. Two of the balls drawn are white
   ii. One is of each colour
   iii. None is red,
   iv At least one is white.
16. A person is known to hit the target 3 out of 4, where as another person is known hit it in 2 out of 3 shot. Find the probability of the target hit at all when they both try.
17. There are two machines in a factory. The probability of machine X works 24 hours without failure is 0.8 and the probability of machine Y working 24 hours without failure is 0.5. Find the probability that
   i. Both machines will work for 24 hors without failure
   ii. Only one machine will work for 24 hours without failure.
18. The odds are 7:5 against A, a person who is now 40 years old living till he is 80 years and the odds are 2:3 in favour of B who is now 30 years of age living till he is 70 years. Find the chance that at least one of these two persons will be Alive 40 years hence.
19. In a group of 100 engineers 10 are bachelors and other are married, what is the probability of selecting a sample of 20 engineers out of which 5 are bachelors, if the sample is selected at random
20. The chance that A can solve a certain problems is 2/3 the chance that B can solve it is ¾. If they both try, find he probability that (i) A solves it but B can't , (ii) B solve it but A can't , (iii) Both of them can't solve it, (iv) at least one of them will solve it (v) only one of them solve it
21. A construction company is bidding for two contract A and B, the probability that the company will get contract A is 3/5, the probability that the company will get contract B is 1/3 and the probability that the company will get both the contract is 1/8. What is the probability that company will get contract A or B?
22. A class consists of 100 students out of these 25 are girls and 75 are boys, 10 of them are rich, remaining poor. Out of 100 students only 20 are fair complexioned. What is the probability of selection a fair complexioned rich girl?
23. There are 3 bags, each containing 5 white balls and 3 black balls. Also there are 2 bags, each containing 2 white balls and 4 black balls. A bag is selected at random and a ball is drawn at random is white. Find the probability that this white ball is drawn from a bag of the first group.
24. A manufacturing firm produces steel pipe in the plants with daily volume of 500, 1000 and 2000 units respectively. According to past

experience it is known that the fractions of defective outputs produced by the three plants are respectively 0.005, 0.008 and 0.010. If a pipe is selected from a day's total production and found to be defective. From which plant the defective pipe is expected to have been produced?

25. There are three machines A, B and C producing 1000, 2000 and 3000 articles per hours respectively, these machines are known to be producing 1%, 2% and 3% defective respectively. One article is selected at random from an hour production of the three machines and found to be defective. What is the probability that the article is produced from
(i) machine A, (ii) machine B, (iii) machine C.

26. In a certain university of Nepal the percentage of Himali, Hilly and Terai region students is 20, 40 and 40 respectively. If 90% of Himali, 50% of hilly and 80% of Terai students are smokers. Findthe probability that a randomly selected smoker student is a hilly student.

27. The chance that doctor 'A' will diagnose a disease X correctly is 60%. The chance that a patient will die by his treatment after correct diagnosis is 40%, and the chance of death by wrong diagnosis is 70%. A patient of doctor A, who had disease X, died. What is the probability that his disease was diagnosed correctly

28. In a certain college of Kathmandu valley the geographical distribution of students is as follows 40 % come from the Central development region, 30% come from Eastern region, 20% come from western region, 8% from Mid-western region and 2% from Far western region. The following proportion of students have black hair, 80% of Central , 70% of Eastern , 60% of Western , 65% of the mid western and 75% of the Far western. Find the probability that a student who has black hair comes from the central region.

29. There are ten Urns of which each of three contains 1 white and 9 black balls, each of the other three contains 9 white and 1 black balls and of the remaining four each contains 5 white and 5 black balls. One of the urns is chosen at random and a ball taken from it is white. What is the probability that an urn containing 1 white and 9 black balls ware selected.

30. In a factory producing portable radio, there are three machines producing 1000, 2000, and 3000 radio per days receptively. These machines produce 1%, 2% and 1% defectives radios respectively. One radio is selected at random from a day's production of the three machines and found to be defective. What is the probability that this radio is produced from first machine ?

31. A mail order firm sends 60% of its orders by a parcel service that delivers late 2% of the time. It sends the other 40% by a combination bus service and local transportation that delivers late 7% of the time.

Find the probability that an order will e delivered late. Also, using Bayes theorem to find the probability that an order delivered late was sent by the parcel service.

32. A firm has 80% of its service calls made by a contractor, and 10% of these calls result in customer complaints. The other 20% of the service calls are made by their own employees, and these calls have a 5% complaint rate, Find the probability of a complaint. Also, using Bayes theorem to find the probability that a complaint was from a customer whose service was provided by the contactor.

33. A box containing 5000 IC chips, of which 1000 are manufactured by company A and the rest by company B. Ten percent of chips made by company A and 5 percent of the chips by company B are defective. If a randomly chosen chip are found to be defective, find the probability that it comes from company A.

34. A car manufacturing factory has two plants. Plant A manufactures 70% o cars and plant B manufactures 30%. At plant A, 80% of cars are rated of standard quality and at plant B, 90% of cars are rated of standard quality. A car is picked up at random and is found to be of standard quality, what is the probability that it has come from (i) plant A and (i) plant B.

35. Rack A contains 2 white and 3 red books and rack B contains 4 white and 3 red books. One book is chosen at random from one of the rack and is found to be the red. Find the probability that it was drawn from rack B.

36. An insurance company insured 2000 Civil engineers, 4000 Electrical engineers and 6000 Mechanical engineers. The probability of an accident involving Civil engineer, Electrical engineer and Mechanical engineer during their jobs is 0.01, 0.03 and 0.15 respectively. One of the insured engineers meets with an accident. What is the probability that he is civil engineer ?

37. In a certain college, 4% of the boys and 1% of the girls are taller than 1.72 meter. Furthermore, 60% of the students are girls. Now if a student is selected at random and is taller than 1.72 meter. What is the probability that the student is a girl ?

38. The probability that a contractor will get a plumbing contract is 2/3 and the probability that he will not get an electric contract is 5/9. If the probability of getting at least one contract is 4/5. What is the probability that he will get both the contracts ?

What is the probability that a family has two cars and an income over Rs 35000 a year

ANSWERS:

7. 15 ways **8.** 56 ways

9. 45, 24 and 21 **10.** 495

11. 462 **12.** 1/6, 5/36, 5/18, 0, 1

13. $\left[\frac{56}{{}^{52}C_4}, \frac{{}^{4}C_2\,{}^{4}C_2}{{}^{52}C_4}, \frac{{}^{26}C_2\,{}^{26}C_2}{{}^{52}C_4}, \frac{{}^{13}C_2\,{}^{13}C_2}{{}^{52}C_4}\right]$ **14.** 4/35, 13/14, 2/5

15. $\left[\frac{{}^{6}C_2\,{}^{13}C_1}{{}^{19}C_3}, \frac{{}^{6}C_1\,{}^{4}C_1\,{}^{9}C_1}{{}^{19}C_3}, 1 - \frac{{}^{13}C_3}{{}^{19}C_3}\right]$ **16.** $\frac{11}{12}$ **17.** i. 0.4 , ii. 0.6

18. $\frac{13}{20}$ **19.** 0.022

20. $\frac{1}{6}, \frac{1}{4}, \frac{1}{12}, \frac{11}{12}, \frac{5}{12}$ **21.** $\frac{97}{120}$ **22.** $\frac{1}{200}$

23. 0.29, 0.18 **24.** $\frac{3}{11}$ **25.** $\frac{14}{45}$

26. $\frac{45}{61}$ **27.** From third plant **28.** $\frac{1}{14}, \frac{2}{7}, \frac{9}{14}$

29. $\frac{2}{7}$ **30.** $\frac{6}{13}$ **31.** 0.4423

32. $\frac{3}{50}$ **33.** $\frac{1}{8}$

34. P(Late) = 0.040 and P(Parcel/late) = 0.30

35. P(Complaint) = .018 and P(Contractor/complaint) = 0.4444

36. $\frac{1}{3}$ **37.** i. $\frac{56}{83}$ ii. $\frac{27}{83}$ **38.** 0.5833

39. $\frac{1}{52}$ **40.** i. 0.0245, ii. 0.551

41. i. 0.3623, ii. 0.4058, iii. 0.6378

42. i. 0.3478, ii. 0.3846, iii. 0.2857, iv. 0.459 and v. 0.255

43. 0.45

CHAPTER 8

8. Random Variables and Distribution Functions

8.1 RANDOM VARIABLE

A rule that assigns a real number to each outcome of a random experiment than the set of this real number is called a random variable (r v). Thus random variable associates numerical values with each elementary outcomes of an experiment. It can take any one of the various possible values each with definite probability.

For example, in a throw of a die if X denotes the number obtained then X is a random variable which can take any one of the values 1, 2, 3, 4, 5, 6, each with equal probabilities $\frac{1}{6}$.

A random variable is usually denoted by any of the Capital Latin letters X, Y, Z, U, V... etc and particular values which the random variable takes are denoted by the corresponding small letters.

Let us consider a random experiment of three tosses of a coin then the sample space S consist of 2^3 =8 sample points given as

S ={HHH, HHT,HTH,THH, HTT,THT, TTH, TTT}.

Let us consider the variable X, which is the number of heads obtained, then X is a random variable which can take any one of the values 0, 1, 2, 3.

Outcomes	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
Values of X	3	2	2	2	1	1	1	0

8.1.1 Types of Random Variables

There are two types of random variable namely.

a. Discrete random variable
b. Continuous random variable

8.1.2 Discrete Random Variable

A discrete random variable takes a finite number of values or an infinite number of values that can be arranged in a sequence that can be counted with positive integers 1, 2 .... And so forth. Further if there is a natural gap between two values of the variable then the variable is discrete random variable. For example, if we toss a coin, the variable can take only two values 0 and 1 assigned to tail and head respectively.

i.e. $X = \begin{cases} 0 & \text{if x is Tail} \\ 1 & \text{if x is head} \end{cases}$

In a rolling of a die, only six values of the variable x, i.e. 1, 2, 3, 4, 5 and 6 are possible, hence the variable x is discrete, here the variable

X = { x: x is 1, 2, 3, 4, 5 and 6}

The other example are i) The number of people in queue in a Cinema hall for ticketing, ii) The number of children in a family, iii) number of run made by a Cricketer in a match etc.

8.1.3 Continuous Random Variable

A random variable X, which can take all possible values in Real line R or between certain limits is said to be a continuous random variable. For a continuous random variable the probability at a point is zero. For example the weight of middle aged people in Kathmandu lying between 40 kg and 150 kg is a continuous random variable.

Symbolically X = {x: 40 ≤ x ≤150}, the other example of continuous random variable are measurement of weight, height, temperature, electric voltage, rainfall, hardness of steel, distance covered by a vechile etc.

Properties of Random Variable

The important properties of random variable are as follows

i. If X is a random variable and a and b are any two constants, then aX+b is also a random variable
ii. If X is a random variable, then X^2 is also a random variable
iii. If X is a random variable then 1/ X is also a random variable
iv. If X and Y are random variables defined on the same sample space S then, X+Y, X-Y, aX, bY or aX+bY are also random variables with a and b are non negative constants
v. If $X_1, X_2 \ldots\ldots X_n$ are n independent random variables then $U_n = \min(X_1, X_2 \ldots\ldots X_n)$ and $V_n = \max(X_1, X_2 \ldots\ldots X_n)$ are also random variables

8.2 PROBABILITY MASS FUNCTION (P.M.F.)

If X is a one –dimensional discrete random variable taking at most a countable numbers $x_1, x_2 \ldots x_n$ with each value of the variable X , we associate a number

$P(x_i) = P(X = x_i)$; i = 1, 2, ... , n.

This is known as the probability of variate values x_i and satisfying the following two conditions

i. Each $P(x_i) = P(X = x_i) \geq 0$ i.e. $P(x_i)$ are all non-negative and
ii. $\Sigma P(x_i) = 1$ i.e. the total probability is unity.

Then $P(x_i) = P(X = x_i)$ is called probability mass function of random variable X. The set of ordered pairs $\{x_i, P(x_i)\}$ i = 1, 2, ...,n is called the probability distribution of the random variable X.

8.3 PROBABILITY DENSITY FUNCTION (P.D.F)

In case of continuous random variable, we do not talk of probability at a particular point, which is always zero, i.e. P (X = xi) = 0, but we always talk of probability in an interval.

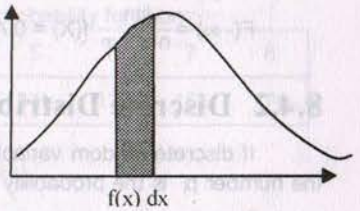
If X is continuous random variable and f(x) is a continuous function of X. Then f(x)dx gives the probability of the event that random variable X lies in the interval $\left(x - \frac{dx}{2} \text{ and } x + \frac{dx}{2}\right)$

i.e. $f(x)\,dx = P\left(x - \frac{dx}{2} \leq X \leq x + \frac{dx}{2}\right)$.

Then f(x) is called probability density function (p.d.f.). It is also known as frequency function because it also gives the proportion of units lying in the interval $x - \frac{dx}{2}$ and $x + \frac{dx}{2}$. If X has the range [a, b], i.e. $a \leq X \leq b$ then the probability density function of random variable X must satisfy the following two properties

i. $f(x) \geq 0$ for every $x \in [a, b]$
ii. $\int_a^b f(x)\,dx = 1$,

it implies that the total area under the frequency curve is always unity.



Example 1

If probability functions of a variable X is defined as

$f(x) = \frac{3}{4}x(2 - x)$, $0 \leq x \leq 2$, , test whether f(x) is p.d.f. or not.

Solution

If given probability function is probability density function then we must have $\int_a^b f(x)\,dx = 1$

$$\text{L.H.S.} = \int_0^2 \frac{3}{4}x(2 - x)\,dx = \frac{3}{4}\left[\frac{2x^2}{2} - \frac{x^3}{3}\right]_0^2 = \frac{3}{4} \times \frac{4}{3} = 1$$

Hence the given function f(x) is probability density function

8.4 DISTRIBUTION FUNCTION

A function $F_X(x)$ of a random variable X for a real value x giving the probability of an event $(X \leq x)$ is called a cumulative distribution function (cdf)

or simply distribution function, symbolically, $F_X(x) = P(X \le x)$, obviously X lies in the interval [x]

8.4.1 Properties of Distribution Function

Some importance properties of distribution functions are.

i. If a and b are two constant values such that a < b and F is the distribution function then $P(a < X \le b) = F(b) - F(a)$

Proof : The events 'a < X ≤ b' and 'X ≤ a' are disjoint and their union is the event 'X ≤ b'. Hence by addition theorem of probability

$P(a < X \le b) + P(X \le a) = P(X \le b)$

$P(a < X \le b) = P(X \le b) - P(X \le a)$

$P(a < X \le b) = F(b) - F(a)$

ii. If F(x) is the distribution function of a mono-variate X, then $0 \le F(x) \le 1$

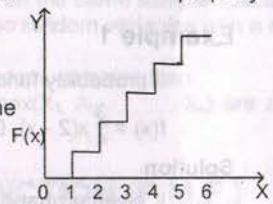
iii. If F(x) is the distribution function of mono-variate X , then

$F(-\infty) = \lim_{n \to -\infty} f(X) = 0$ And $F(\infty) = \lim_{n \to \infty} f(X) = 1$

8.4.2 Discrete Distribution Function

If discrete random variable X takes sample points $x_1, x_2 \ldots\ldots x_n$ and the number p_i is the probability function satisfies the property $p_i \ge 0$ and

$\sum_{i=1}^{\infty} p_i = 1$ such that $F(x) = \sum_{i:X \le x} p_i$ is called the discrete distribution function of random variable X.



For example; If x_i is just the integer i , so that

$P(X=i) = P_i = i$; i = 1, 2.....n, then F(x) is a

Step function having jump p_i at i and being Constant between each pair of integers.

8.4.3 Continuous Distribution Function

If X is a continuous random variable with the pdf f(x) then the function F(x) defined as $F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx$ is called the distribution function (df) or sometimes the cumulative distribution function (cdf) of the random variable X.

Properties

i. $0 \le F(x) \le 1$; $-\infty < x < \infty$

ii. $F'(x) = \frac{dF(x)}{dx} = f(x) \ge 0$, f(x) being pdf

iii. $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$ and $F(\infty) = \lim_{x \to \infty} = F(x) = 1$

iv. $P(a \le X \le b) = \int_a^b f(x)\,dx$

$= \int_{-\infty}^{b} f(x)\,dx - \int_{-\infty}^{a} f(x)dx$

$= P(X \le b) - P(X \le a) = F(b) - F(a)$

Example2

A random variable X has the following probability function

x	0	1	2	3	4	5	6	7	8
P(x)	k	3k	5k	7k	9k	11k	13k	15k	17k

i. Determine the values of k.

ii. Find P(X < 3) , $P(X \ge 3)$ and P(0< X <5).

iii. Find the distribution function of X.

iv. Find the smallest values of x for which $P(X \le x) > \frac{1}{2}$ and

Solution

i. if given probability function p(x) is probability mass functin then total probability must be unity that is $\sum_{i=0}^{\infty} p(x) = 1$

or, 81k = 1,

Therefore the value of constant $k = \frac{1}{81}$

ii. $P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2)$

$= \frac{1}{81} + \frac{3}{81} + \frac{5}{81} = \frac{1}{9}$

Now, $P(X \ge 3) = 1 - P(X<3)$

$= 1 - \frac{1}{9} = \frac{8}{9}$

And $P(0 < X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$

$= \frac{3}{81} + \frac{5}{81} + \frac{7}{81} + \frac{9}{81}$

$$= \frac{24}{81}$$

iii. The distribution function F(x) of random variable X is given in the adjoining Table

X	P(x)	F(x)= P (X ≤ x)
0	0.012	0.012
1	0.037	0.049
2	0.062	0.111
3	0.086	0.198
4	0.111	0.309
5	0.136	0.444
6	0.160	0.605
7	0.185	0.790
8	0.210	1.000

iv. By trial, we get from table such that the smallest values of x for which $P(X \leq x) > \frac{1}{2}$ is; x = 6

8.4 MATHEMATICAL EXPECTATION

Once we have constructed the probability distribution of a random variable, we often want to compute the mean or expected value of the random variable. The expected value of a discrete random variable is a weighted average of all possible values of the random variable, where the weights are the probabilities associated with the corresponding values.

If X is a random variable which can assume any one of the values $x_1, x_2, \ldots x_n$ with respective probabilities $P(x_1), P(x_2) \ldots P(x_n)$. Then the mathematical expectation of r v X usually called the expected value of X and denoted by E(X) and it is defined as

$$E(X) = x_1 P(x_1) + x_2 P(x_2) + \ldots\ldots\ldots + x_n P(x_n) = \Sigma x P(x)$$

Where $\sum_{i=1}^{n} p(x) = 1$ i.e. total probability is unity

Similarly, if X is continuous random variable with probability density function f(x) then mathematical expectation of r.v X is defined as

$$E(X) = \int_a^b xf(x)\,dx \text{ ; where } f(x) \geq 0 \text{ for every } x \in [a, b] \text{ and } \int_a^b f(x)\,dx = 1$$

8.4.1 Physical meaning of Mathematical Expectation

Let us consider the following frequency distribution of the discrete random variable X as follwos

X :	x_1	x_2	...	x_n
F :	f_1	f_2	...	f_n

Then the arithmetic mean of the discrete random variable is given by

$$\bar{x} = \frac{\Sigma fx}{N} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots f_nx_n}{N}$$

$$\bar{x} = \frac{f_1}{N} x_1 + \frac{f_2}{N} x_2 + \ldots\ldots\ldots\ldots + \frac{f_n}{N} x_n$$

We observed that, out of total of N cases f_i cases are favourable to each of x_i events.

Therefore $P(X = x_i) = \frac{f_i}{N} = P(x_i)$, i = 1, 2.....n.

Which implies that, $\frac{f_1}{N} = P(x_1)$, $\frac{f_2}{N} = P(x_2)$ ..... $\frac{f_n}{N} = P(x_n)$

Therefore, $\bar{x} = x_1P(x_1) + x_2P(x_2) + \ldots\ldots + x_nP(x_n) = \Sigma x\, P(x) = E(X)$.

Hence, mathematical expectation of a random variable is nothing but its simply arithmetic mean of random variable.

8.4.2 Properties of Mathematical Expectation

i. If 'k' is a constant then E(k) = k

$$E(k) = \sum_{i=1}^{n} k\, p(x_i) = k \sum_{i=1}^{n} p(x_i) = k \qquad [\text{since } \sum_{i=1}^{n} p(xi) = 1]$$

ii. The mean of random variable Y defined as Y = a x + b with mean of random variable X as μ is = a E(X) + b = a μ + b

'a' and 'b' are two constants then E (aX +b) = a E(X) + b

E (aX + b) = E(aX) + E (b)

= a E(X) + b

= a E(X) + b

iii. The mathematical expectation of the sum of random variables is equal to the sum of their expectation, provided all the expiation exists.

Symbolically, Let X, Y Z ... T are random variables then

E (X + Y + Z + ... +T) = E (X) + E (Y) + E (Z) + ... + E (T)

iv. The mathematical expectation of product of number of independent random variables is equal to the product of their expectation. Symbolically, if X, Y, Z .... T are n independent random variables then

$E(X\ Y\ Z\ \ldots\ T) = E(X)\,E(Y)\,E(Z)\ \ldots\ E(T)$

Example 3

An unbiased dice is thrown at random, what is the expectation of the number on it ?

Solution

Let X denotes the number on the dice. Then X is a random variable which can take any one of the values each with equal probability $\frac{1}{6}$ as

X	1	2	3	4	5	6
P(x)	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Therefore, mathematical expectation of number on dice is,

$$E(X) = \Sigma\, xP(x) = 1\times\frac{1}{6} + 2\times\frac{1}{6} + 3\times\frac{1}{6} + 4\times\frac{1}{6} + 5\times\frac{1}{6} + 6\times\frac{1}{6}$$

$$= \frac{21}{6} = 3.5$$

Therefore, expected vale of number on a dice is 3.5

8.4.3 Expected Value of Function of a Random Variable

Consider a random variable X with probability density function f(x). If g(x) is a function of random variable than by definition it is also a random variable. Then, expected value of function of a random variable g(x) i.e.E{g(x)} is defined as:

$$E[g(x)] = \begin{cases} \int_a^b g(x)\, f(x)\, dx, & \text{for continuous random variable} \\ \sum_x g(x)\, P(x), & \text{for discrete random variable} \end{cases}$$

8.5 VARIANCE OF RANDOM VARIABLE

The square of standard deviation (σ) is known as Variance. If X is a random variable then the variance is denoted as $Var(X) = \sigma^2$ and defined as

$$Var(X) = E\{X - E(X)\}^2$$
$$= E\{X^2 - 2X\,E(X) + [E(X)]^2$$
$$= E(X^2) - 2E(X)\,E(X) + [E(X)]^2 = E(X^2) - \{E(X)\}^2$$

Where $E(X^2) = \sum_{i=1}^{n} x_i^2\, p(x_i)$, If X is discrete random variable and

$= \int_a^b x^2 f(x)\, dx$, If X is continuous random variable

8.5.1 Properties of Variance of Random Variable

i. If k is a constant then variance of constant is always zero i.e. V(k) =0

ii. If 'a' and 'b' are two constants, then the variance of random variable Y defined as Y = a X + b is defined as

$Var(Y) = V(aX + b) = V(aX) + V(b) = a^2 V(X)$

iii. If 'a' and 'b' are two constants and X and Y are two random variables then the variance of random variable Z defined as Z = a X ± bY is defined as

$Var(Z) = V(aX \pm bY)$

$= a^2\ V(X) + b^2\ V(Y) \pm 2ab\ Cov(X,Y)$

$= a^2\ V(X) + b^2\ V(Y)$

[if X and Y are independent then Cov(X,Y) = 0]

Example 4

A random variable X has the following probability function

x	0	1	2	3	4	5	6	7	8
P(x)	k	3k	5k	7k	9k	11k	13k	15k	17k

i. Determine the values of k.

ii. Find the Mean Value and Variance of Random Variable X

Solution

i. From Example Number 1 the value of constant is $k = \frac{1}{81}$

ii. The mean value and Variance of Random variable X is calculated as follows

X	P(x)	X P (x)	X^2 P (x)
0	$\frac{1}{81}$	0.000	0.000
1	$\frac{3}{81}$	0.037	0.037
2	$\frac{5}{81}$	0.123	0.247

3	$\frac{7}{81}$	0.259	0.778
4	$\frac{9}{81}$	0.444	1.778
5	$\frac{11}{81}$	0.679	3.395
6	$\frac{13}{81}$	0.963	5.778
7	$\frac{15}{81}$	1.296	9.074
8	$\frac{17}{81}$	1.679	13.432
Total	**1**	5.481	34.519

Therefore, Mean value of random variable X is

$E(X) = \Sigma X P(x) = 5.481$

And for the variance of random variable first we calculate $E(X^2)$ as

$E(X^2) = \Sigma x^2 P(x) = 34.519$

Now, the variance of random variable X is

$\text{Variance}(X) = E(X^2) - \{E(X)\}^2$

$= 34.519 - (5.481)^2$

$= 34.519 - 30.046$

$= 4.472$

Example 5

A college professor never finishes his lecturer before the bell rings to end of the period and always finishes his lecturer within one minute after the bell rings. Let x is the time that elapse between the bell and the end of the lecturer, and suppose the probability density function of x is as

$$f(x) = \begin{cases} k x^2, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

i. Find the value of constant k

ii. What is the probability that the lecturer ends within 2 seconds of bell rings?

iii. What is the probability that the lecturer continuous beyond the bell for between 20 to 40 seconds

Solution

i. If the given function is probability density function then we must have

$$\int_a^b f(x)\,dx = 1$$

$$\text{Or, } \int_0^1 kx^2 dx = 1$$

$$\text{Or, } k\frac{1}{3} = 1$$

$$\text{Or, } k = 3$$

ii. For Probability that the lecturer ends within 20 seconds upper limits of $X = \frac{20}{60} = \frac{1}{3}$ seconds

$$P(X \le \tfrac{1}{3}) = \int_0^{0.33} f(x)\,dx = \int_0^{0.33} 3x^2\,dx = 0.037$$

ii. For Probability that the lecturer continuous beyond the bell for between 20 to 40 seconds upper limits of $X = \frac{40}{60} = \frac{2}{3}$ seconds

$$P(\tfrac{1}{3} \le X \le \tfrac{2}{3}) = \int_{0.33}^{0.66} f(x)\,dx = \int_{0.33}^{0.66} 3x^2\,dx = 0.2593$$

8.6 MEDIAN OF RANDOM VARIABLE

Median is the point which divides the whole data in to two equal parts. It is a positional average, which is defined for continuous random variable with probability density function f(x) as

$$\text{Median} = \int_a^M f(x)\,dx = \int_M^b f(x)\,dx = \frac{1}{2}$$

8.7 QUARTILE AND DECILES OF RANDOM VARIABLE

Quartiles divide the data into four equal parts therefore there are three quartile namely Q_1 Q_2 and Q_3 but value of Q_2 is equal to Median so values of and Q_1 and Q_3 are calculated as

$$Q_1 = \int_a^{Q_1} f(x)\,dx = \frac{1}{4} \text{ and } Q_3 = \int_a^{Q_3} f(x)\,dx = \frac{3}{4}$$

Also there are nine deciles which divide the whole data into ten equal parts which are calculated as follows

$$D_i = \int_a^{D_i} f(x)\,dx = \frac{i}{10}\ ;\ \text{where } i = 1, 2 \ldots 9$$

8.8 MODE OF RANDOM VARIABLE

Mode is the value of variable X which maximizes the probability density function f(x). Therefore, value of mode is given by the relationship

$f'(x) = 0$ and $f''(x) < 0$ provided it lies in [a, b]

Example 6

In three tosses of a coin, Let X, be the number of heads. Tabulate the possible outcomes with the corresponding value of X. By simply counting & derive the probability distribution of X and hence calculate the expected value and variance of X.

Solution

Let H represents a head, T a tail and X, the random variable denoting the number of heads. Then the following table shows the value of random variable with sample points.

S. No.	outcome	No. of heads
1	HHH	3
2	HHT	2
3	HTH	2
4	THH	2
5	HTT	1
6	THT	1
7	TTH	1
8	TTT	0

The random variable X takes the value 0, 1, 2, 3. Since from the above table, we find that the numbers of cases favorable to the coming of 0, 1, 2 & 3 heads are 1, 3, 3 and 1 respectively.

Therefore, $P(x = 0) = \frac{1}{8}$, $P(X = 1) = \frac{3}{8}$, $P(X = 2) = \frac{3}{8}$, $P(X = 2) = \frac{1}{8}$

Then, probability distribution of X can be summarized as follows.

X	0	1	2	3
P(x)	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{8}$

Then, Expected value of random variable X is

$$E(X) = \Sigma x P(x) = 0 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} + 3 \times \frac{1}{8} = \frac{12}{8} = 1.5$$

We have $Var(X) = E(X^2) - [E(X)]^2$

Now, $E(X^2) = \Sigma x^2 P(x) = 0^2 \times \frac{1}{8} + 1^2 \times \frac{3}{8} + 2^2 \times \frac{3}{8} + 3^2 \times \frac{1}{8}$

$$= 0 + \frac{3}{8} + \frac{12}{8} + \frac{9}{8} = 3$$

Hence, $Var(X) = E(X^2) - [E(X)]^2$

$= 3 - (1.5)^2 = 0.75.$

Thus, expected value of random variable X is 1.5 and its variance is 0.75

Example 7

A continuous random variable X has the Probability density function defined as

$$f(x) = \begin{cases} k\,x(1-x)^2, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

i. Find the value of constant K.

ii. Also Find the mean and Variance of Random Variable X

Solution.

i. If the given function is probability density function then we must have

$$\int_a^b f(x)\,dx = 1$$

$$\text{Or, } \int_0^1 k\,x(1-x)^2 dx = 1$$

$$\text{Or, } k\frac{1}{12} = 1$$

$$\text{Or, } k = 12$$

Thus the required probability density function is

$$f(x) = \begin{cases} 12\,x(1-x)^2, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}$$

ii. The mean of the random variable is calculate as

$$E(X) = \int_a^b x\,f(x)\,dx = \int_0^1 12\,x^2(1-x)^2 dx = \frac{12}{30}$$

$$\text{Now, For variance } E(X^2) = \int_a^b x^2 f(x)\,dx = \int_0^1 12\,x^3(1-x)^2 dx = \frac{12}{60}$$

Thus the value of variance is

$Variance(X) = E(X^2) - \{E(X)\}^2$

$= 0.20 - (0.4)^2$

$= 0.20 - 0.16$

$= 0.04$

Exercise - 9

1. Define Probability mass function and probability density function.
2. Define random variable, what are the types of random variable?
3. What do you understand by discrete distribution function and continuous distribution function?
4. Define mathematical expectation of a random variable. Show that mathematical expectation of a random variable is an arithmetic mean of a random variable.
5. If a discrete random variable has the probability function as

X	0	1	2	3	4	5	6	7	8
P(x)	k	2k	3k	5k	5k	4k	3k	2k	k

(i) Determine the values of k. ii.

(ii) Find $P(X < 4)$, $P(X \geq 4)$, $P(0 < X < 5)$.

(iii) Find the smallest values of x for which $P(X \leq x) > \frac{1}{4}$ and

(iv) Find the distribution function of X.

6. A random variable X has the following probability function. Find the expectation and variance of the random variable.

X	-1	0	1	2
P(x)	$\frac{1}{3}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$

7. If the probability density functions of a variable X is defined as

$$f(x) = \begin{cases} c\,x\,(2-x) & ; 0 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

Find the value of constant c. Also, fine the mean and variance of random variable X.

8. If a random variable X forms the following function.

X	3	2	1	0	-1	-2	-3
P(x)	0.1	0.2	3k	k	2k	0	0.1

Find the value of k. Also find the mean and variance of X.

9. Let X be a random variable with $f(x) = \begin{cases} ke^{-2x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$

Find the value of K and its mean and variance

10. A random variable X is defined as the sum of faces when a fair dice is thrown twice. Find the expected value of X and its variance.
11. A r.v X has the following probability function as;

X	-2	-1	0	1	2	3
P(x)	0.1	k	0.2	2k	0.3	k

Find the value of k, and calculate the mean and variance of r.v.

12. Let X, Y, Z be three random variables taking values 1, 2, 3, 4, 5 & 6 with corresponding probability distribution as;

r.v	1	2	3	4	5	6
P(x)	0.03	0.5	0.04	0.06	0.1	0.2
P(y)	0.02	0.4	0.3	0.05	0.03	0.2
P(Z)	0.04	0.6	0.4	0.6	0.1	0

Whether P(x), P(y) and P (z) are probability mass function and why?

13. State with reason, if the following probability distribution are admissible or not

i.	X	0	1	2	
	P(x)	0.3	0.2	0.5	
ii.	X	-1	0	1	
	P(x)	0.4	0.4	0.3	
iii.	X	0	1	2	3
	P(x)	0.2	0.3	0.3	0.1
iv.	X	-2	-1	0	1	2
	P(x)	0.3	0.4	-0.2	0.2	0.3

14. A random variable X has the following probability function

X	0	1	2	3	4	5	6	7
P(x)	0	k	2k	2k	3k	k^2	$2k^2$	$7k^2+k$

Find (i) value of k, (ii) evaluate $P(X \geq 6)$, (iii) $P(X < 6)$
(iv) $P(0 < X < 5)$ and (v) Find mean and variance of X.

15. A bakery has the following schedule of daily demand for cakes. Find the expected number of cakes demanded per day.

No. of cake	0	1	2	3	4	5	6	7	8	9
Probability	0.02	0.07	0.09	0.12	0.20	0.20	0.18	0.10	0.01	0.01

16. Obtain the probability distribution of 'number of heads' in three tosses of a coin. Hence find the mean and variance of the distribution
17. The monthly demand for transistor is known to have following probability distribution.

Demand	1	2	3	4	5	6
P(x)	0.1	0.15	0.2	0.25	0.18	0.12

Determine the expected demand for transistor.

18. For the following given the probability function, Find the mean and variance of random variable X.

X	0	1	2	3
P(x)	0.1	0.3	0.5	0.1

19. Suppose a fair coin is tossed 6 times and heads is called a success then the binomial distribution is given as

X	0	1	2	3	4	5	6
P(x)	1/64	6/64	15/64	20/64	15/64	6/64	1/64

Find the mean number of success and its variance.

20. A random variable X is distributed at random between the values 0 and 1 so that its pdf is
$f(x) = k x^3 (4 - x)^2$
Where, k is a constant. Find the value of k, using this value of k find mean and variance of distribution.

21. A variable X is distributed at random between the value 0 and 4 and its pdf is given by
$f(x) = k x^3 (4 - x)^2$,
Find the value of k, the mean and variance of the distribution

22. A random variable X has pdf as
$f(x) = k x (2 - x); \leq x \leq 2$
0, otherwise
Find the value of k, mean and variance of random variable.

23. Find the expected sales ('000Rs) of Toyota car in Kathmandu city in a week from the following information.

Day	Sun	Mon	Tue	Wed	Thu	Fri
Sales	90	60	90	50	75	55
Probability	0.25	0.18	0.12	0.05	0.2	0.2

Answers:

5. (i) $\frac{1}{26}$, (ii) $\frac{11}{26}, \frac{15}{26}, \frac{15}{26}$, (ii) 3 **6.** Mean = 0.5, var = 1.583

7. $\frac{3}{4}$, 1, $\frac{1}{5}$ **8.** $k = \frac{1}{10}$, 0.5, and 2.85

9. $k = \frac{1}{2}$, Mean $= \frac{1}{2}$, var (x) $= \frac{1}{4}$ **10.** E(x) = y and Var (x) = 31.89

11. k = 0.1, Mean = 0.8, var(x) = 2.16 **12.** Y

13. (i) Yes, since $\Sigma P(x) = 1$, (ii) No. since $\Sigma P(x) > 1$
(iii) No, since $\Sigma P(x) < 1$, (iv) No, P(x) never negative

14. (i) k = 0.10 (ii) $P(x \geq 6) = \frac{4}{5}$, (iii) $P(x < 6) = \frac{1}{5}$, $P(o < x < 5) = \frac{19}{100}$

15. 4.36 **16.** Mean $\frac{3}{2}$ and variance $\frac{3}{4}$

17. 3.62 **18.** 1.6, 0.64

19. Mean = 3, variance = 1.5 **20.** $K = \frac{30}{77}, \frac{422}{539}$ and 0.0295.

21. 15/1024, 16/7, 2,$\sqrt{6}$ /7 **22.** $\frac{3}{4}$, 1, 1/5 **23.** Rs. 72600

Chapter 10

Joint probability Distribution

Bi-variate Random Variable

In many situations, a scientist (or a statistician) has to measure more than one factor or character at a time on the same item or some adjustant items. These measurements may have different units of measurements and moreover the range or the domain of the variables may also be different. If two variables are studied simultaneously in respect of their distribution, they are known as bi-variate random variable. Two or more variables studied concurrently with regard to their distributions are known in general multivariate studies. The variables may be discrete or continuous or the mixture of the two. But in this chapter we have dealt with the situations when both the variables are either discrete or continuous. Just like uni-variate variables, the two variables have also to be studied, which behave after differently than uni-variate because of the impact of one variable over the same sample space? In this situation the variables can be jointly discrete or jointly continuous. Here, we are dealing with two dimensional space or planes.

Definition

"Let X and Y be two random variables defined on the same sample space S then the function (X,Y) that assigns a points in R^2, is called a two dimensional random variables Let $A = \{a < X \leq b\}$, $B = \{c < Y \leq d\}$ be two events of sample space S. then the event

$\{a < X \leq b, c < Y \leq d\} = \{a < X \leq b\} \cap \{c < Y \leq d\} = A \cap B$
$\therefore P[\{a < X \leq b\} \cap \{c < Y \leq d\}] = P[A \cap B]$

Two dimensional discrete random variable

A r.v (X, Y) is said to be a two dimensional discrete r.v If it can take only countable number of points (x, y) in a two-dimensional space. The r.v (X, Y) is also said to be joint discrete r.v. For example, Let us consider an experiment of tossing a coin three times and take the variable X as the number of Heads in first tossing and Y the number of tails in three tosses. In this problem our bi-variate r.v is the ordered pair (X, Y). Also both the variables X and Y are discrete, the possible outcomes are

S {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}. The pairs of variate values (x, y) are (x, y) : {(1, 0) , (1,1), (1,1), (1,2) , (0,1) , (0,2), (0,2), (0,3)}.

Two dimensional continuous random variable

A r.v (X, Y) is said to be a two dimensional continuous r.v if it can take all the values between $-\infty$ to ∞ in a two dimensional space. The r. v (X, Y) is also said to be joint continuous r.v. For example, we may measure the weight and the hardness of a rock, the volume and pressure of a gas, the

8. The marginal probability mass functions are:

X	0	1	2	3
P(x)	1/32	3/32	9/32	19/32

y	0	1
P(y)	14/32	18/32

9. The conditional distribution of Y when X = x is as

Y \ X	0	1	2
0	0	1/6	2/9
1	1/3	2/6	3/9
2	2/3	3/6	4/9

10. $f(x) = \frac{3}{4}\frac{(2x+3)}{(1+x)^4}$; $0 < x < \infty$, $f(y) \frac{3}{4}\frac{(3+2y)}{(1+y)^4}$, $0 < y < \infty$

$f(Y/X) = 6\frac{(1+x+y)}{(1+y)^4(3+2x)}$; $0 < y < \infty$

11. $c = 1/4$, $f(x) = \frac{1}{4}(4x+2)$, $f(y) = \frac{1}{4}(1+y)$ (iii) dependent

12. Marginal probability density function of X and Y is

x:	1	2	3
p(x):	$\frac{1}{3}$	$\frac{11}{27}$	$\frac{7}{27}$

and

y	1	2	3
P(y)	$\frac{11}{27}$	$\frac{7}{27}$	$\frac{1}{3}$

And the conditional distribution of X given Y = 2 is

x:	1	2	3
p(X/y=2)	1/7	3/7	3/7

13. i $k = \frac{1}{8}$ ii. $f(x) = \begin{cases} \frac{1}{4}x^3; 0 < x < 2, \\ 0, \text{otherwise} \end{cases}$ $f(y) = \begin{cases} \frac{8}{3} - 2y; -2 < y < 2 \\ 0, \text{otherwise} \end{cases}$

14. i $k = 8$ ii. $\frac{1}{32}$ (iii) $f(x) = \begin{cases} 4(1-x^2); 0 < x < 1 \\ 0, \text{otherwise} \end{cases}$

$f(y) = \begin{cases} 4y^3; 0 < y < 1 \\ 0, \text{otherwise} \end{cases}$ $f(x/y) = \frac{2x}{y^2}$, $0 < x < y$, $0 < y < 1$

$f(y/x) = \frac{2y}{(1-x^2)}$, $x < y < 1$, $0 < x < 1$

iv. X and Y are dependent

CHAPTER 10

10. Theoretical Probability distributions

10.1 INTRODUCTION

The experimental or observed or empirical frequency distributions studied earlier are based on actual sample observations, which are drawn from a population. The main features of sample frequency distribution are averages, dispersions. Skewness, kurtosis, correlation etc but their knowledge is not sufficient if complete information about the distribution is necessary. If we take a large number of observations of an actual distribution a more scientific way of drawing conclusions about the population characteristic are through the study of theoretical distributions. In the theoretical probability distribution, the value of the variable may be distributed according to some definite probability law which can be expressed mathematically.

A theoretical probability distribution is a systematic arrangement of probabilities associated with mutually exclusive and collectively exhaustive elementary outcomes where as a frequency distribution is listing of the observed frequencies of all the outcomes of an experiment. Probability distributions are related to frequency distribution. It is likely that if the number of experiment is increased sufficiently, the empirical distribution may approach very closely to theoretical probability distribution. The main goals of theoretical distribution are to provide a logical basis for making decisions and predictions on the basis of limited information on theoretical considerations.

In statistics, we come across many distributions either experimentally or theoretically. We have two types of theoretical probability distributions.

a. Discrete theoretical probability distributions.

b. Continuous theoretical probability distributions.

10.2 DISCRETE THEORETICAL PROBABILITY DISTRIBUTIONS

Following are the commonly used discrete theoretical probability distributions.

i. Uniform distribution
ii. Bernoulli distribution
iii. Binomial distribution
iv. Poisson distribution
v. Negative Binomial distribution

vi. Geometric distribution
vii. Hyper-Geometric distribution

10.3 DISCRETE UNIFORM DISTRIBUTION

A discrete random variable X is said to have a discrete uniform distribution over the range [1, n] if its probability mass function is expressed as follows.

$$P(X=x) = P(x) = \begin{cases} \frac{1}{n} \text{ for } x = 1,2, \ldots\ldots\ldots n \\ 0, \text{ Otherwise} \end{cases}$$

Here n is known as the parameter of the distribution and lies in the set of all positive integers. For example, the outcomes of rolling a fair die or drawing cards successively from a well shuffled deck of cards follows uniform distribution, in this distribution, probability at any point remains the same.

Mean and Variance of uniform distribution

$$\text{Mean} = E(X) = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n(n+1)}{2n} = \frac{(n+1)}{2} \text{ and}$$

$$E(X^2) = E(X^2) = \frac{1}{n}\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6n} = \frac{(n+1)(2n+1)}{6}$$

$$\text{Therefore } Var(X) = E(X^2) - [E(X)]^2 = \frac{n(n+1)(2n+1)}{12}$$

The moment generation function of uniformly distributed. rv X is

$$M_x(t) = E(e^{tx}) = \frac{1}{n}\sum_{i=1}^{n} e^{tx} = \frac{e^t(1-e^{nt})}{n(1-e^t)}$$

10.4 BERNOULLI DISTRIBUTION

A random variable X, marked by only two values of 0 & 1 or -1 & 1 with probability of occurrence p and probability of non-occurrence q = 1-p respectively . Where P(X=1) = p and P(X = 0) = q is called a Bernoulli variate and its distribution is called Bernoulli distribution

If we flip a coin, it will show either head or tail on the upper face, the sex of an expected baby will either be male or female (excluding expectations), a manufactured item will either be defective or non-defective are some of the examples of Bernoulli distribution.

Definition: A discrete random variable X which assumes only two values 0 and 1 is said to have a Bernoulli distribution with parameter p if its probability mass function is given by

$$P(X=x) = \begin{cases} p^x q^{1-x} \text{ for } x = 0 \& 1 \\ 0, \text{ otherwise} \end{cases} \text{ where } 0 \le p \le 1$$

Thus The probability mass function of the Bernoulli distribution can be written as $P(X=x) = \begin{cases} p \text{ if } x = 1 \\ q, \text{ if } X = 0 \end{cases}$

10.4.1 Mean and Variance of Bernoulli distribution

$$\text{Mean} = E(X) = \sum_{x=0,1} XP(x) = 0 \times q + 1 \times p = p \text{ and}$$

$$E(X^2) = \sum_{x=0,1} X^2 P(x) = 0^2 \times q + 1^2 \times p = p$$

Therefore Variance of X , $Var(X) = E(X^2) - [E(X)]^2 = p^2 - p = pq$.

Moment generating function of Bernoulli distributed of random variable X is

$$M_x(t) = E(e^{tx}) = \sum_{x=0,1} e^{tx} P(x) = (q + pe^t)$$

10.5 BINOMIAL DISTRIBUTION

Binomial distribution was discovered by James Bernoulli (1654-1705) in the year 1700 and was published posthumously in 1713, eight years after his death.

Let a random experiment be performed repeatedly, each repeat being called trial and let the occurrence of an event in a trial be called a success and it's non –occurrence a failure. This distribution can be used under the following conditions.

i. The random experiment is performed repeatedly a finite and fixed number of times. In other words number of Bernoulli trials is finite and fixed. For example if a coin is tossed 12 time then n = 12.

Ii Each Bernoulli trial has only two possible outcomes success or failure, head or tail, yes or no. for example, if a coin is tossed then either head or tail is received

iii. All the trials are statistically independent. By this we mean that the results of a particular trial are not going to affect the result of any other trial. For example if a coin is tossed, the trials would be independent.

iv. The probability of success 'p' remains same for each trial and probability of failure q = 1 –p.

v. The sum of probabilities of success and failure is unit.

vi. The probability of success vis-à-vis failure is not very low

vii. The random variable X denotes a number of successes in n-trials.

Any experiment which satisfies the above conditions is known as Binomial Experiment.

Flip a conin12 times, A manufacturing of pen drive machine produce 2% defective pen drives in a batch of 100, etc are the example of binomial trial.

10.5.1 Probability Mass Function of Binomial Distribution

The probability of x successes and consequently (n-x) failures in 'n' independent trials, in a specified order say FFSFSSSF.....SFS, where S represent success and F represents failure, is given by the compound probability theorem as

P (F F S F S S S F ... S F S)

= P (F) P(F) P(S) P(F) ... P(S) P(F) P(S)

= q q p q ... p q p. [Since trial are independent]

= p p p ... p q q q ... q q q

(x successes) (n-x failures)

$= p^x q^{n-x}$

But x success in 'n' trials can occur in ${}^nC_x$ ways and the probability for each of these ways is same (say) = $p^x q^{n-x}$.

Hence the probability of x success in n trials in any order is given by the addition theorem of probability by the expression ${}^nC_x p^x q^{n-x}$.

Here ${}^nC_x p^x q^{n-x}$ is the $(x+1)^{th}$ terms in the binomial expansion of $(p+q)^n$

Definition: A discrete random variable X is said to follows binomial distribution with two parameters n and p if it assumes only non-negative values and its probability mass function is given by

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

10.5.2 Properties of Binomial Distribution

The following are the important properties of Binomial distribution.

i. The two independent constants n and p in the binomial distribution are known as the parameters of the binomial distribution , sometimes 'n' is also known as the degree of the binomial distribution.

ii. Binomial distribution is a discrete distribution as X can take only the integral values viz 0, 1, 2, 3 .....n.

iii. Any random variable which follows binomial distribution is known as binomial variate.

iv. We shall use the notation $X \approx B(n, p)$ to denote that the random variable X follows binomial distribution with parameter n and p.

v. The P(x) defined in binomial distribution is probability mass function

because $\sum_{x=0}^{n} P(x) = \sum_{x=0}^{n} {}^nC_x p^x q^{n-x} = (q+p)^n = 1.$

vi **Mean and Variance of Binomial distribution**

Let X follows binomial distribution with parameters n and p then

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

$$\therefore \text{Mean} = E(X) = \Sigma x P(x) = \sum_{x=0}^{n} x P(x) = \sum_{x=0}^{n} x \, {}^nC_x p^x q^{n-x}$$

$$= \sum_{x=0}^{n} x \frac{n!}{(n-x)!x!} p^x q^{n-x} = \sum_{x=0}^{n} x \frac{n(n-1)!}{(n-x)!x(x-1)!} pxqn-x$$

$$= np \sum_{x=1}^{n} {}^{n-1}C_{x-1} p^{x-1} q^{n-x}$$

$= n p (q + p)^{n-1} = n p$

Thus, mean of binomial distribution is 'np'

Similarly, Variance of X = Var (X) = $E(X^2) - [E(X)]^2$

Then, $E(X^2) = \Sigma x^2 P(x) = \sum_{x=0}^{n} \{x(x-1) + x\} P(x)$

$$= \sum_{x=0}^{n} x(x-1) \, {}^nC_x p^x q^{n-x} + \sum_{x=0}^{n} xP(x)$$

$$= \sum_{x=0}^{n} \frac{n(n-1)(n-2)!\, p^x q^{n-x}}{(n-x)!\,(x-2)!} + np$$

$$= n(n-1)p^2 \sum_{x=2}^{n} {}^{n-2}C_{x-2}\, p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2(q+p)^{n-2} + np = n(n-1)p^2 + np$$

Therefore, $Var(X) = E(X^2) - [E(X)]^2$

$$= n(n-1)p^2 + np - \{np\}^2$$

$$= np(1-q) = npq$$

Therefore, the value of Mean and variance of binomial distribution is '**np**' and '**npq**'

vii. For the binomial distribution, variance is less than mean.

Since q is the probability of failure, for binomial distribution we always have

$0 < q < 1$, multiply by np on all sides we get

$0 < npq < np$

$0 <$ variance $<$ mean

Thus, for binomial distribution mean is greater than that of variance.

viii. If $p = q = \frac{1}{2}$, the binomial distribution is symmetrical.

ix. If $p < \frac{1}{2}$, the binomial distribution is positive skewed.

x. If $p > \frac{1}{2}$, the binomial distribution is negative skewed.

xi Binomial distribution is unimodal if 'np' is a whole number and the mean and mode are equal, each being 'np'

10.5.3 Recurrence Relation of Binomial Distribution

Let X follows binomial distribution with parameters n and p then

$P(X=x) = {}^nC_x\, p^x\, q^{n-x}$ and

$P(X = x+1) = {}^nC_{x+1}\, p^{x+1}\, q^{n-x-1}$

Hence $\frac{P(X=x+1)}{P(X=x)} = \frac{{}^nC_{x+1}\, p^{x+1}\, q^{n-x-1}}{{}^nC_x\, p^x\, q^{n-x}} = \frac{(n-x)p}{(x+1)q}$

$\therefore P(X = x+1) = \frac{(n-x)p}{(x+1)q}$

This relation is used to calculate the probabilities of binomial distribution when probability of X = 0 is known

If a Binomial trial is repeated N-times the theoretical frequency or expected frequency of the binomial distribution is given by

$e(x) = f_e = N\ p(X=x) = N \cdot {}^nCx\, p^x\, q^{n-x}$, $x = 0, 1, 2 \ldots n$

This is known as the fitting of Binomial distribution

10.5.4 Moment Generating Function of Binomial Distribution

Let $X \approx B(n, P)$ then probability mass function is given as

$$P(X=x) = \begin{cases} {}^nC_x\, p^x\, q^{n-x}, \text{for } x = 0,1,2, \ldots\ldots n \\ 0, \text{ otherwise} \end{cases} \quad 0 \le p \le 1,\ q = 1-p$$

Then the moment generating function is defined as

$$M_X(t) = E(e^{tx}) = \Sigma e^{tx} P(x) = \sum_{x=0}^{n} e^{tx}\, {}^nC_x\, p^x\, q^{n-x}$$

$$= \sum_{x=0}^{n} {}^nc_x (Pe^t)^x\, q^{n-x} = (q + pe^{tx})^n$$

Example 1

Whether the following is binomial distribution or not? The mean of distribution is 3 and its variance is 4?

Solution

If the given binomial distribution has parameters n and p, then we have given mean = np = 3 and variance = npq = 4

$\therefore$ q = 4/3 >1, which is impossible, since probability can't be exceed unity. Hence the given statement is not binomial distribution.

Example 2

Suppose we have 10 balls in a bowl, 3 of the balls are red and 7 of them are blue. Define success S as drawing a red ball. If we draw samples with replacement. Let's say n =20 i.e. if we draw 1 ball at a time with replacement 20 times. Find the probability of drawing 5 red ball and expected number of red balls.

Solution

Three balls are red out of 10 balls so probability of selecting a red ball is P(S) = 0.3 for every trial and n = 6

$X \sim B(20, 0.3)$ and

$$P(X = 5) = \binom{20}{5} 0.3^5 (1 - 0.3)^{20-5}$$

$$= 15504 \times (0.3)^5 (0.7)^{15} = 0.1789$$

$E(X) = 20(0.3) = 6,$

Variance $(x) = 20(0.3)(0.7) = 4.2$

Example 3

Nine coins are thrown simultaneously. Find the probability of getting at least six heads.

Solution

Let p- probabilty of getting a head in throw of a coin = ½.

And q – probability of not getiing a head = ½.

Let x be number of heads then the probability of getting x heads in a random throw of 9 coins is given by binomial law as

$$P(X=x) = {}^nC_x\, p^x q^{n-x} = {}^9C_x\, (1/2)^x (1/2)^{9-x} = {}^9C_x\, (1/2)^9,\ x = 0, 1, \ldots\ldots 9$$

Therefore, probability of getting at least six head is given by

$$P(X \geq 6) = P(6)+P(7)+P(8)+P(9) = (1/2)^9 \{{}^9C_6 + {}^9C_7 + {}^9C_8 + {}^9C_9\} = 0.254$$

Example 4

The incidence of occupational disease in a industry is such that the workmen have a 20% chance of suffering from it .what is the probabiltty that out of 8 workmen 5 or more will contract the disease?

Solution

Let suffering of a disease is success then p = 20% = 0.2, q = 0.8 and n = 8

Therefore, x-success from n trial is given by binomial distribution as

$$P(X=x) = {}^nC_x\, p^x q^{n-x} = {}^8C_x\, (0.2)^x (0.8)^{8-x} = ,\ x = 1, 1, \ldots 8$$

Then the probabilty that five or more workmen will contract the disease is given as follows

$$P(X \geq 5) = P(5) + P(6) + P(7) + P(8)$$

$$= {}^8C_5(0.2)^5(0.8)^{8-5} + {}^8C_6(0.2)^6(0.8)^{8-6} + {}^8C_7(0.2)^7(0.8)^{8-7} + C_8(0.2)^8(0.8)^{8-8}$$

$$= 0.0104$$

Example 5

What is the probability of guessing correctely at least seven of the 12 answers in the multi;e choice onjective test with four choices in each

questions ? Also calculate the mean number of guess correct answers and its variance

Solution

In multiple choice objective test with four choices in each quesitons, the probabilioty of guessing an answer correltely is givnen by

$$p = \frac{1}{4} \text{ and } q = \frac{3}{4}$$

By Binomial probability law, the probability of guessing x-answers in a 12 questions test is given as

$$P(X = x) = {}^nC_x\, p^x q^{n-x}\ ;\ x = 0, 1, \ldots 12$$

$$= {}^{12}C_x \left(\frac{1}{4}\right)^x \left(\frac{3}{4}\right)^{12-x}$$

i. P(At least seven correct answer) = $P(X \geq 7)$

$$= P(X = 7) + P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$

$$= {}^{12}C_7(0.25)^7(0.75)^{12-7} + {}^{12}C_8(0.25)^8(0.75)^{12-8}$$

$$+ {}^{12}C_9(0.25)^9(0.75)^{12-9} + {}^{12}C_{10}(0.25)^{10}(0.75)^{12-10}$$

$$+ {}^{12}C_{11}(0.25)^{11}(0.75)^{12-11} + {}^{12}C_{12}(0.25)^{12}(0.75)^{12-12}$$

$$= 0.0115 + 0.0024 + 0.0004 + 0.000035 + 0.0000021$$

$$= 0.01434$$

Therefore, (At least seven correct answer) = $P(X \geq 7) = 1.43\%$

ii. Also mean number of correct answer = np = 12 x 0.25 = 3 and

iii. Variance of correct answer = npq = 12 x 0.25 x 0.75 = 2.25

Example 6

Fit a binomial distribution to the following information.

X:	0	1	2	3	4
f:	5	10	20	30	15

Solution

With the usual, we have n = 4, N = Σf = 80

If P be the parameter of the binomial distribution, then

$\bar{x} = nP.$

$$\text{Now } \bar{x} = \frac{\Sigma fx}{N} = \frac{200}{80} = 5/2 = 2.5$$

$$\therefore \quad np = \bar{x}$$

$p = \frac{\bar{x}}{n} = \frac{2.5}{4} = \frac{5}{8}$ & $q = 1 - p = 3/8$

Let x be the number of successes in the experiment then, the expected frequency of getting x success is given by.

$f(x) = N_x P(x = x) = N^n c_x\, p^x\, q^{n-x}$, when x = 0, 2, ...

Here, N = 80, n = 4, p = 5/8 & q = 3/8

$\therefore\ f(x) = 80 \times {}^4c_r \left(\frac{5}{8}\right)^x \left(\frac{3}{8}\right)^{4-x}$, where x = 0, 2, 2, 3, 4

$\therefore$ Fitting of Binomial, distribution

x	f(x) = N X P(X = x)
0	f(0) = 1.588 ~2
1	f(1) = 10.5 ~ 11
2	f(2) = 26.36 ≅ 26
3	f(b) = 29.29 ~ 29
4	f(4) = 12.20 ~ 12

Hence, the fitted binomial distribution is as

x:	0	1	2	3	4
f:	2	11	28	29	12

Example 7

12 dice were thrown 4096 times and a throw of 6 was recorded as success, the observed frequencies ware as given below. Fit a binomial distribution.

Success:	0	1	2	3	4	5	6	7 and over
Frequency:	447	1145	1181	796	380	115	24	8

Solution

We know that if the dice are unbiased, then probability of getting a 6 with one die is 1/6 then.

$p = \frac{1}{6}$ and $q = 1 - p = \frac{5}{6}$

If getting a face six is success then x success from n trial is given by

$P(X = x) = {}^nc_n\, p^x\, q^{n-x}$

And expected frequencies are calculated as using the formula

$e(x) = N\,P(X = x) = N.\ {}^nc_x\, p^x\, q^{n-x}$

Here N = 4096, n = 12

$\therefore\ e(x) = 4096\ {}^{12}C_x \left(\frac{1}{6}\right)^x \left(\frac{5}{6}\right)^{12-x}$

$\therefore\ e(0) = 4096\ {}^{12}C_0\ (1/6)^0\ (5/6)^{12-0} = 459.4 = 459$

$\therefore$ Fitting of a binomial distribution is

X	e(x) = N P(X = x)
0	e(0) = 459.39 = 459
1	e(1) = 11102
2	e(2) = 1212
3	e(b) = 808
4	e(4) = 364
5	e(5) = 116
6	e(6) = 27
7 or more	e(7 or more) = 8

10.6 POISSON DISTRIBUTION

Poisson distribution was discovered by the French mathematician and Physicist Simeon Denis Poisson (1781-1840) in the year 1837. when the probability of success of an event is very small so that q = 1-p, the probability of failure is almost equal to unity and 'n' the number of trial is very large such that the average number of series is a finite number, the resulting distribution is known as Poisson distribution . It is also a discrete probability distribution.

This distribution deals with the evaluation of probabilities of rare events such as

i. Number of car accidents on road

ii. Number of earthquakes in a year

iii. Number of printing mistakes per page

iv. Number of blind children born in a city

v. The number of twin births in a city hospitals etc.

The Poisson distribution is good approximation of the Binomial distribution when 'n' is greater than or equal to 20 and 'p' is less or equal to 0.05. it is also noted that if $n \geq 100$, the approximation is generally excellent so long as $np \leq 10$.

10.6.1 Conditions of Poisson distribution

Poisson distribution is a limiting case of binomial distribution under the following conditions.

i. The discrete random variable X is the number of occurrence of an event in an interval of time.

ii. The occurrences of the events must be random.

iii. The occurrences of the events must be rare and independent of ech other.

iv. The occurrence must be uniformly distributed over the interval being used.

v. The number of trials 'n' is indefinitely large i.e. $n \to \infty$

ii. The constant probability of success 'p' for each trial is indefinitely small i.e. $p \to 0$

iii. The product of n and p i.e. $np = \lambda$ (say) is finite so that

$p = \frac{\lambda}{n}$ and $q = 1 - \frac{\lambda}{n}$ where λ is read as lambda.

10.6.2 Derivation of Probability Mass Function of Poisson Distribution

The probability of x-successes in a series of n independent trial is given by Binomial distribution as

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

Here we want to obtain the probability mass function as limiting cases as $n \to \infty$ and $p \to 0$ such that $np = \lambda$ $P = \frac{\lambda}{n}$ and $q = 1 - \frac{\lambda}{n}$. Hence

$$\lim_{n\to\infty} P(x) = \lim_{n\to\infty} \frac{n!}{(n-x)!x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty} \frac{n(n-1)(n-2)\ldots\ldots\{n-(x-1)\}(n-x)!}{(n-x)!x!} \left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^{n-x}$$

$$= \lim_{n\to\infty} \frac{n^x \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\ldots\ldots\ldots\left(1-\frac{x-1}{n}\right)\lambda^x \left(1-\frac{\lambda}{n}\right)^n}{x!\, n^x \left(1-\frac{\lambda}{n}\right)^x}$$

$$= \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, 2 \ldots\ldots\ldots\ldots \infty$$

$$\left\{\because \lim_{x\to\infty}\left(1+\frac{a}{x}\right)^x = e^a \text{ and } \lim_{x\to\infty}\left(1+\frac{a}{x}\right)^\alpha = 1, \text{ where '}\alpha\text{' is not function of } x\right\}$$

10.6.3 Definition of Poison Distribution

A discrete random variable X is said to follows a poison distribution with single parameters λ if it assumes only non-negative values and it's probability mass function is given by

$$P(x) = P(X = x) = \begin{cases} \frac{e^{-\lambda}\lambda^x}{x!}, \text{ for } x = 0,1,2, \ldots\ldots \infty \\ 0, \text{ otherwise} \end{cases} \quad \lambda \geq 0$$

10.6.4 Properties of Poisson distribution

Following are the important properties of Poisson distribution.

i. The Poisson distribution depends upon the value of λ, the average number of success per unit, λ is called the parameter of the Poisson distribution.

ii. Like binomial distribution Poisson distribution is a discrete probability distribution, it is concerned with occurrence that can take integral value 0, 1, 2 ... ∞.

iii. Any random variable which follows Poisson distribution is known as Poisson variate.

iv. We shall use the notation $X \sim P(\lambda)$ to denote that random variable X follows Poisson distribution with parameter λ.

v. The probability of success $P(X=x) = P(x)$ defined in the Poisson distribution is pmf because

$$\sum_{x=0}^{\infty} P(x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!} = e^{-\lambda}e^{\lambda} = 1$$

vi. **Mean and Variance of Poisson distribution**

Since $X \sim P(\lambda)$, then probability mass function of Poisson distribution is

$$P(x) = \frac{e^{-\lambda}\lambda^x}{x!},\ x = 0, 1, \ldots \infty$$

$$\text{Mean} = E(X) = \sum_{x=0}^{\infty} x\,P(x) = \sum_{x=0}^{\infty}\frac{xe^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}$$

$$= e^{-\lambda}\lambda e^{\lambda} = \lambda$$

Thus mean of Poisson distribution is λ

Similarly,

$$E(X^2) = \sum_{x=0}^{\infty} x^2 P(x) = \sum_{x=0}^{\infty} \{x(x-1) + x\} P(x)$$

$$= \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda} \lambda^x}{x!} + \sum_{x=0}^{\infty} xP(x)$$

$$= e^{-\lambda} \lambda^2 \sum_{x=0}^{\infty} \frac{x(x-1) \lambda^{x-2}}{x(x-1)(x-2)!} + \lambda$$

$$= e^{-\lambda} \lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = e^{-\lambda} \lambda^2 e^{\lambda} + \lambda = \lambda^2 + \lambda$$

Therefore, $V(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

Thus, for Poisson distribution the mean and variance is equal to λ

viii. Poisson distribution is independent of trial n.

ix. Poisson variates are independent to each others.

x. The Poisson distribution is a positively skewed distribution. As λ increase, the graph of Poisson distribution would get closer to a symmetrical continuous curve.

Xi Poisson distribution is unimodal if λ is not an integer and unique modal value is integral part of λ and it is bimodal if λ is a integral and models value being λ and $(\lambda - 1)$

10.6.5 Moment Generating Function of Poisson Distribution

Since $X \sim P(\lambda)$, then probability mass function of Poisson distribution is

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2 \ldots \infty$$

Then moment generating function of random variable X which follows Poisson distribution is defined as

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \frac{e^{-\lambda} \lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e t)^x}{x!} = e^{-\lambda} e^{\lambda e t} = e^{\lambda(e^t - 1)}$$

Therefore, Mean $= E(X) = \left.\frac{\partial M_x(t)}{\partial t}\right|_{t=0} = \lambda e^{\lambda(e^t-1)} e^t \Big|_{t=0} = \lambda$

And $E(X^2) = \left.\frac{\partial^2 M_x(t)}{\partial t^2}\right|_{t=0} = \left.\frac{\partial \lambda e^{\lambda(e^t-1)} e^t}{\partial t}\right|_{t=0}$

$$= \lambda\{\lambda e^{\lambda(e^t-1)} e^t . e^t\} + \lambda e^{\lambda(e^t-1)} e^t \Big|_{t=0} = \lambda^2 + \lambda$$

Therefore, $V(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$

10.6.6 Recurrence Relation of the Poisson Distribution

For a Poisson distribution with parameter λ we have,

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, \ldots \infty$$

And, $P(X = x+1) = \frac{e^{-\lambda} \lambda^{x+1}}{(x+1)!}, \quad x = 0, 1, \ldots \infty$

$$\therefore \frac{P(X = x+1)}{P(X = x)} = \frac{\lambda}{(x+1)}$$

Which implies $P(X = x+1) = \frac{\lambda}{(x+1)} P(X = x)$

Which is the required recurrence formula for Poisson distribution used to calculate the probability of Poisson distribution when probability of X =0 is known.

If an experiment, satisfying the requirement of Poisson distribution, is repeated N-times, then an expected frequency or theoretical frequency of getting x-successes is given by

$$e(x) = f_e = N\ P(X = x) = N \frac{e^{-\lambda} \lambda^x}{x!}; \quad x = 0, 1, \ldots \infty.$$

If λ is known we can easily calculate the different probabilities of variate values and its expected frequency, If λ is unknown we use the relation

Mean of Poisson data $= \bar{x} = \lambda$ to estimate the value of λ

10.7.7 Some Example of Poisson Distribution

Following are the instances where Poisson distribution may be successfully employed.

i. Number of deaths from a disease such as heart attack, HIV Aids, Cancer or due to snake bite in a year.

ii. Number of twin births in a hospital.

iii. Number of suicides reported in a particular city.

iv. The number of defective materials in a packet say, pins blades etc. manufactured by a good concern.

v. The emission of radioactive (alpha) particles.

vi. Number of accidents taking place per day on a busy road.

vii. Number of printing mistakes per page of the book.

viii. Number of telephone calls coming in a switch board in a particular minute.

ix. Number of plane accidents per week.

Example 8

A manufacturer of cotter pins knows that 5% of his product is defective. If he sells cotter pins in boxes of 100 and guarantees that no more than 10 pins will be defective. What is the approximate probability that a box will fail to meet the guaranteed quality?

Solution

We are given sample size n = 100, probability of a defective pin = 5% = 0.05. Since 'p' is small and n is large we may use Poisson distribution

$\therefore = \lambda$ = mean number of defective pins = n p = 100 x 0.05 = 5.

Let the random variable X denote the number of defective pins in a box of 100. Then by Poisson probability law, The probability of x defective in a box is given as

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-5}(5)^x}{x!}, \quad x = 0, 1 \ldots \infty$$

Therefore probability that a box will fail to meet the guarantees quality

$$P(X > 10) = 1 - \sum_{x=0}^{10} \frac{e^{-5}(5)^x}{x!} = 1 - P(X \le 10) = 1 - e^{-5} \sum_{x=0}^{10} \frac{5^x}{x!}$$

Example 9

A company has two machine of mixture of concrete, which it hires out day to day. The number of demands for a mixture on each day is distributed as a Poisson distribution with mean 1.5.

Calculate the proportion of days on which

i. Neither mixture is used, and

ii The proportion of days on which some demand is refused

Solution

Here the random variable X, which denotes the number of demands for a mixture on any day following Poisson distribution with mean $\lambda = 1.5$. The proportion of days on which there are x demands for a mixture is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-1.5}1.5^x}{x!} \quad x = 0, 1, 2 \ldots \infty$$

i. proportion of days on which neither mixture is used is given by

$P(X = 0) = e^{-1.5} = 0.223$

ii. Proportion of days on which some demands are refused is

$P(X > 2) = 1 - P(x \le 2)$

$= 1 - \{P(X = 0) + P(X = 1) + P(X = 2)\}$

$= 1 - e^{-1.5}\left(1 + \frac{1.5^1}{1!} + \frac{1.5^2}{2!}\right)$

$= 0.19115$

Example 10

An insurance company insures 4,000 people against loss of both eyes in a motorcycle accident. Based on previous data the rates were computed on the assumption that on the average 10 persons in 1,000,000 will have motorcycle accident each year that result in this type of injury. What is the probability that more than 3 of the insured will collect on their policy in a given year?

Solution

Number of insured persons (n) = 4000

Probability of loss of both eyes in a motorcycle accident = $\frac{10}{100000}$

Since p is very small and n is large, we may approximate the given distribution by Poisson distribution. Thus the parameter λ of the Poisson distribution is $\lambda = n\,p = 4000 \times 0.0001 = 0.4$

Let, the random variable X denotes number of motorcycle accident in the batch of 4000 people. Then by Poisson probability law

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = \frac{e^{-0.4}\,0.4}{x!}, \quad x = 0, 1, 2 \ldots \infty$$

Hence the required probability that more than 3 of the insured will collect on their policy is given by

$$P(X>3) = 1 - P(x \leq 3)$$

$$= 1 - \{P(x = 0) + P(x = 1) + P(X = 2) + P(X = 3)\}$$

$$= 1 - e^{-0.4}\left(1 + \frac{(0.4)}{1!} + \frac{(0.4)^2}{2!} + \frac{(0.4)^3}{3!}\right)$$

$$= 0.0008$$

Example 11

After correcting 100 pages of the proof of a book, the proof reader finds that there are on the average 2 error per 5 pages. How many pages would one expect to find with 0, 1, 2, 3 and 4 errors, in 1000 pages of the first print of the book?

Solution

Let the random variable X denoted the number of errors per pages, then the mean number of error per page is given by $\lambda = 2/5 = 0.4$

Using Poisson probability law, probability of x errors per page is given by,

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!} = , x = 0, 1, 2 \ldots \infty$$

Expected numbers of pages with x errors per page in a book of 1000 page are calculated as using the formula as

$$e(X = x) = N\,P(X = x) = 1000\,\frac{e^{-0.4}\,0.4^x}{x!} , x = 0, 1, 2 \ldots \infty$$

Calculation for expected Poisson frequencies

No of error per pages (x)	Probabilities P(x)	Expected Frequencies f(x)
0	P(0) = 0.6703	670.3 ~ 670
1	P(1) = 0.2681	268.12 ~ 268
2	P(2) = 0.0536	53.624 ~ 54
3	P(3) = 0.0071	7.129 ~ 7
4	P(4) = 0.0007	0.71 ~ 1

10.8 NEGATIVE BINOMIAL DISTRIBUTION

One if interested in the probability of r^{th} success after a number of failure say x- failures, then negative Binomial distribution will be used, it is also known as Waiting time distribution or Pascal distribution after the French mathematician **Blaise Pascal** (1623-62) because he contributed to develop the mathematical model.

In sampling procedure sample size is fixed, but in negative Binomial experiment sample size is not fixed but is determined as the sample size required to achieve x-failures r^{th} successes. This is known as inverse sampling. Negative Binomial distribution basically depends upon the binomial distribution but the last trial must be success in Negative binomial experiment. The negative binomial experiment consists of independent trials which are repeated until a desirable number of success are occurs, with the probability of success that remain constant at each trial. In fact a negative binomial distribution is the converse of a binomial distribution. In binomial distribution, the number of trials is fixed and the number of successes to occur is random variable but in case of negative binomial distribution the number of successes is fixed and the number of trial is a random variables.

Negative binomial distribution can be used under the following conditions.

i. The experiment consists of sequence of independent trials.

ii. Each trial has only two possible outcomes; a success and a failure.

iii. The probability of success is constant from trial to trial.

iv. The experiment continues until a total of 'r' success have been observed with last trial success

v. We count the number of trials before the r^{th} success occurs.

10.8.1 Derivation of Negative Binomial Distribution

Let the probability of success of an event A i.e. P(A) = p and the probability of failure of an event A^c i.e. $P(A^c) = q = 1 - p$. Assume that trials are independent and the probability of success 'p' in a trial remains constant from trial to trial.

Also, let us assume that x-failure preceding the r^{th} success in r+x independent trials. In this situation the last trial must be success with probability p. Here the probability of x- failures is q.q.q.q..... to x times = q^x and the probability of r^{th} successes = p p p p..... to r- times = p^r.

The r^{th} event to be success implies that there should be x-failure preceding the r^{th} success that is there must be (r -1) success out of (x + r -1) trial and x-failure can be chosen from (x+r-1) events. This can be selected in $x+r-1C_x$ ways.

Therefore the probability of (r-1) success among (x + r -1) trial is given by binomial distribution as

$$= {}^{x+r-1}C_{r-1}\, q^{(x+r-1)-(r-1)} = {}^{x+r-1}C_{r-1}\, p^{r-1}\, q^x$$

Since the last trial is success whose probability is p. By the multiplication theorem of probability, the probability of x-failure preceding the r^{th} success is given as

$$P(x = x) = {}^{x+r-1}C_{r-1}\, q^x p^{r-1}. p$$

$$P(x = x) = {}^{x+r-1}C_{r-1}\, p^r q^x$$

$$P(x = x) = {}^{x+r-1}C_x\, p^r q^x \qquad [\text{Since } {}^{x+r-1}C_{r-1} = {}^{x+r-1}C_x]$$

We also have,

$$\binom{x+r-1}{x} = \frac{(x+r-1)!}{(x+r-1-x)!\ x!}$$

$$= \frac{(x+r-1)(x+r-2)\ldots(r+1)\,r\,(r-1)!}{(r-1)!\ x!}$$

$$= \frac{(-1)^x(-r)(-r-1)\ldots.(-r-x+2)(-r-x+1)(-r-x)!}{(-r-x)!\ x!}$$

$$= (-1)^x \frac{(-r)!}{(-r-x)!\ x!}$$

$$= (-1)^x \binom{-r}{x}$$

$$\therefore\ P(X=x) = \binom{-r}{x} p^r (-q)^x \quad , x = 0, 1, 2\ldots\ldots\ldots\ldots,$$

This is the probability of x-failure preceding the rth success. Also which is the $(x+1)^{th}$ term in the expansion of $p^r(1-p)^{-r}$,a binomial expansion with a negative index, hence the distribution is known as negative binomial distribution.

10.8.2 Definition of Negative Binomial Distribution

A discrete random variable X is said to follow a negative Binomial distribution with parameter r and P if it assumes only non negative values and its probability mass function of x- failure preceding the r^{th} success is given by

$$P(x=x)=P(x) = \begin{cases} {}^{x+r-1}C_x P^r q^x \text{ for } x = 0,1,2\ldots \\ 0, \quad \text{otherwise} \end{cases}$$

$$\text{Or, } P(x) = \binom{-r}{x} p^r (-q)^x\ , x = 0, 1, 2\ldots$$

10.8.3 Properties of Negative Binomial Distribution

i. The negative binomial distribution depends upon 'r' and 'p' then 'r' and 'p' are called the parameter of the negative binomial distribution.

ii. Negative binomial distribution is a discrete probability distribution. It is concerned with occurrence that can take integral values as 0, 1, 2 ...

iii. Any random variable which follows negative binomial distribution is known as negative binomial variate.

iv. We shall use the notation X~NBD(r, p) to denote the random variable X follows negative binomial distribution with parameters 'r' and 'p'.

v. The P(X=x) defined in NBD is pmf because

$$\sum_x P(x) = \sum_x {}^{-r}C_x p^r (-q)^x = p^r \sum_x {}^{-r}C_x (-q)^x = p^r(1-q)^{-r} = 1 \text{ ie total probability is unit.}$$

vi. Mean and variance of negative binomial distribution are $\frac{rq}{p}$ and $\frac{rq}{p^2}$ respectively.

Let x follows negative binomial distribution then the probability of x-failure, preceding the r^{th} success is given by

$$P(X=x) = P(x) = {}^{x+r-1}C_x\, p^r q^x = \binom{-r}{x} p^r (-q)^x$$

$$\therefore \text{ Mean} = E(X) = \sum_{x=0}^{\infty} x\, P(x)$$

$$= \sum_{x=0}^{\infty} x \binom{-r}{x} p^r (-q)^x = \sum_{x=0}^{\infty} \frac{x(-r)!}{(-r-x)!\ x!} p^r (-q)^x$$

$$= p^r(-r)\sum_{x=1}^{\infty}\frac{(-r-1)!}{(-r-x)!\ (x-1)!}(-q)^x$$

$$= p^r(-q)(-r)\sum_{x=1}^{\infty}\binom{-r-1}{x-1}(-q)^{x-1} = rq\,p^r(1-q)^{-r-1} = \frac{rq}{p}$$

Now, variance of Negative Binomial distribution is
Variance(X) = $E(X^2) - [E(X)]^2$

Now, $E(X^2) = \sum_{x=0}^{\infty} x^2 P(x) = \sum_{x=0}^{\infty}\{x(x-1)+x\}P(x)$

$$= \sum_{x=0}^{\infty} x(x-1)\binom{-r}{x} p^r(-q)^x + \sum_{x=0}^{\infty} x\,P(x)$$

$$= \sum_{x=0}^{\infty}\frac{x(x-1)(-r)!}{(-r-x)!\ x!}p^r(-q)^x + \frac{rq}{p}$$

$$= p^r(-r)(-r-1)\sum_{x=2}^{\infty}\frac{(-r-2)!}{(-r-x)!\ (x-2)!}(-q)^{x-2}(-q)^2$$

$$= qr(r+1)p^r\sum_{x=2}^{\infty}\binom{-r-2}{x-2}(-q)^{x-2} + \frac{rq}{p}$$

$$= qr(r+1)p^r(1-q)^{-r-2} + \frac{rq}{p} = \frac{rq(r+1)}{p^2} + \frac{rq}{p}$$

$\therefore$ Var(X) = $E(X^2) - [E(X)]^2$

$$= \frac{rq(r+1)}{p^2} + \frac{rq}{p} - \left(\frac{rq}{p}\right)^2$$

$$= \frac{rq}{p^2}$$

vii. In Negative Binomial Distribution the number of success is fixed and number of trials is a random variable.

viii. Variance of Negative Binomial Distribution is greater than its mean i.e.

$$\frac{rq}{p^2} > \frac{rq}{p}$$

$\Rightarrow p < 1$, which is true.

10.8.4 Moment Generating Function of Negative Binomial Distribution

Let x follows negative binomial distribution then the probability of x-failure, preceding the r^{th} success is given by

$$P(X=x) = P(x) = {}^{x+r-1}C_x\, p^r q^x = \binom{-r}{x} p^r(-q)^x$$

The moment generating function of random variable X which follows NBD is given by

$$M_X(t) = E(e^{tX}) = \sum_{x=1}^{\infty} e^{tx} P(x) = \sum_x e^{tx}\,{}^{-r}C_x\, p^r(-q)^x$$

$$= pr\sum_x {}^{-r}C_x (qe^t)^x = p^r(1-qe^t)^{-r}$$

Alternatively, we can obtain the mean and variance by partial derivative of moment generating function with respect to 't' as 't' tends to 0 as

$$\text{Mean} = E(X) = \left.\frac{\partial M_x(t)}{\partial t}\right|_{t=0}$$

$$= p^r(-r)(1-qe^t)^{-r-1}(-qe^t)\Big|_{t=0} = rp^r qe^t(1-qe^t)^{-r-1}\Big|_{t=0}$$

$$= rp^r q(1-q)^{-r-1} = \frac{rq}{p}$$

$$E(X^2) = \left.\frac{\partial^2 M_x(t)}{\partial t^2}\right|_{t=\theta} = rp^r q\{e^t(-r-1)(1-qe^t)^{-r-2}(-qe^t) + e^t(1-qe^t)^{-r-1}\}\Big|_{t=0}$$

$$= rp^r q\{q(r+1)(1-q)^{-r-2} + (1-q)^{-r-1}\} = \frac{rq^2(r+1)}{p^2} + \frac{rq}{p}$$

$$= \frac{rq}{p^2}(q(r+1)+p) = \frac{rq}{p^2}(qr+1)$$

Therefore, $V(X) = E(X^2) - [E(X)]^2 = \frac{rq}{p^2}(qr+1) - \left(\frac{rq}{p}\right)^2 = \frac{rq}{p^2}$

Example 12

An item is produced in a large numbers. The machine is known to produce 5% defectives items. A quality control inspector is examining the items by taking them at random. What is the probability that at least 4 items are to be examining in order to get 2 defective?

Solution

If 2-defectives are to be obtained then it can be happen in 2 or more trials. The probability of defective (success) is p = 0.05 for every trial. It is a negative binomial situation with r = 2 success, and p = 0.05 and x replaced by (x-2) and the required probability is given by

$$P(X=4) + P(X=5) + \ldots\ldots$$

$$= \sum_{x=4}^{\infty} {}^{(x-2+2-1)}C_{x-2}\,(0.05)^2\,(0.95)^{x-2}$$

$$= \sum_{x=2}^{3} {}^{x-1}C_{x-2}\,(0.05)^2(0.95)^{x-2}$$

$$= 1 - \{(0.05)^2 + 2\,(0.05)^2\,(0.95)\} = 0.9928$$

Example 13: Banach's Matchbox problem

A statistician carries one match box each in his right and left pockets, suppose initially each box contains N match sticks, each time he wants a match stick, he selects a box at random, consider the moment when the statistician finds a box empty . Show that the probability that there is exactly r match sticks in one box when the other is found empty is $\binom{2N-r}{N-r}\left(\frac{1}{2}\right)^{2N-r}$

Solution

Lets denote two match boxes by A and B . Let the choice of box A be regarded as failure and that of box B is regarded as success

Since he selects the match box at random, Probability of selecting of B box = p = ½ and q = 1-p = ½, probability of selecting A box. The box B will be found empty if it selected $(N+1)^{st}$ time . At these stage the A box contains exactly r matches', if (N-r) matches have already been drawn from it.

Hence B box will be found empty at the moment when A box contains exactly r matches iff (N-r) failures precede the $(N+1)^{st}$ success. Thus in total (N+1+N-r) = (2N-r+1) Trials, the last one must be success and out of remaining (2N-r) trials, we should have (N-r) failures and N successes. Therefore Probability that B box is found empty while there are r matches in another box A is $\binom{2N-r}{N-r} p^{N+1} q^{N-r}$,

Then the probability that A box is found empty when there are r matches in box B is also $\binom{2N-r}{N-r} q^{N+1} p^{N-r}$,

Therefore required probability is

$$\binom{2N-r}{N-r} p^{N+1} q^{N-r} + \binom{2N-r}{N-r} q^{N+1} p^{N-r}$$

$$= 2\binom{2N-r}{N-r}\left(\frac{1}{2}\right)^{N+1}\left(\frac{1}{2}\right)^{N-r}$$

$$= 2\binom{2N-r}{N-r}\left(\frac{1}{2}\right)^{2N-r}\left(\frac{1}{2}\right)$$

$$= \binom{2N-r}{N-r}\left(\frac{1}{2}\right)^{2N-r}$$

10.9 HYPER-GEOMETRIC DISTRIBUTION

When the population is finite and sampling is drawn without replacement, so that the events are statistically dependent, although random, we obtain hyper geometric distribution. Consider a box with N balls, M of which are white and N-M are red, Suppose that we draw a sample of n balls at random (without replacement) from the box. Then the probability of getting x-white balls out of n balls is $\dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

Since x-white balls can be drawn from 'M' white balls in ${}^{M}C_x$ ways and out of the remaining (N-M) red balls, (n-x) balls can be chosen in ${}^{N-M}C_{n-x}$ ways, the total number of favorable case is ${}^{M}C_x \;{}^{N-M}C_{n-k}$. And total numbers of exhaustive cases are ${}^{N}C_n$. Therefore the probability of x-white ball out of M is given by

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

Binomial distribution is applicable in which the probability of success is same for all trials. However, if it varies from trial to trial, hyper-geometric distribution is more suitable.

Under the following condition we use hyper-geometric distribution.

i. the population or universe where the sample drawn consist of N finite individual, objects or elements.

ii. Each individual con be characterized as a success or failure and there are M successes in the whole population.

iii. A random sample of n individuals is selected without replacement in such a way that each subset of size n is equally likely to be chosen.

10.9.1 Definition of Hyper Geometric Distribution

A discrete random variable X is said to follow the hyper geometric distribution with parameters N, M and n if it assumes only non-negative values and its probability mass function is given by

$$P(X=x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad x = 0, 1, 2 \ldots \min(n, M); \quad M \le N; \quad n \le N$$

0, otherwise

Where N is a positive integer, M is a positive integer not exceeding N and n is a positive integer that is at most N.

10.9.2 Properties of Hyper-Geometric Distribution

The following are the important properties of hyper-geometric distribution

i. The three independent constants, N, M and n in the distribution are known as the parameters of hyper geometric distribution

ii. Hyper geometric distribution is a discrete distribution as X can take only the integer values viz, 0, 1, 2.....min (n, M).

iii. Any random variable which follows hyper geometric distribution is known as hyper geometric variate.

iv. We shall use the notation X~HG (k; N, M, n) to denote that the random variable X follows hyper geometric distribution with three parameters N, M and n.

v. The P(X = x) defined in the hyper geometric distribution is the pmf because $\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$

$$\sum_{x=0}^{n} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = 1 \quad \left(\because \sum_{k=0}^{n} \binom{M}{k}\binom{N-M}{n-k} = \binom{N}{n} \right)$$

vi. Mean and variance of the hyper geometric distribution is

$$\text{Mean} = E(X) = \sum_{x=0}^{n} x\, P(X=x)$$

$$= \sum_{x=0}^{n} x \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \sum_{x=0}^{n} x \frac{M!\binom{N-M}{n-x}}{x!\,(M-x)!\binom{N}{n}}$$

$$= \frac{M}{{}^{N}C_{n}} \sum_{x=0}^{n} \binom{M-1}{x-1}\binom{N-m}{n-x}$$

$$= \frac{M}{{}^{N}C_{n}} \sum_{x=1}^{n} \binom{M-1}{x-1}\binom{N-1-(M-1)}{n-1-(x-1)}$$

$$\therefore E(X) = \frac{M}{{}^{N}C_{n}}\binom{N-1}{n-1} = \frac{nM}{N}$$

Therefore, Mean of Hyper geometric distribution is $\frac{nM}{N}$

Similarly,

$$E(X(x-1)) = \sum_{x=0}^{n} x(x-1)\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \sum_{x=0}^{n} x(x-1)\frac{M!\binom{N-M}{n-x}}{x!\,(M-x)!\binom{N}{n}}$$

$$= \frac{M(M-1)}{{}^{N}C_{n}} \sum_{x=2}^{n} \binom{M-2}{k-2}\binom{N-M}{n-k}$$

$$= \frac{M(M-1)}{{}^{N}C_{n}}\binom{N-2}{n-2} = \frac{M(M-1)\,n(n-1)}{N(N-1)}$$

$$\sum_{x=0}^{n} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = 1 \left(\because \sum_{k=0}^{n} \binom{M}{k}\binom{N-M}{n-k} = \binom{N}{n} \right)$$

vi. Mean and variance of the hyper geometric distribution is

$$\text{Mean} = E(X) = \sum_{x=0}^{n} x\, P(X=x)$$

$$= \sum_{x=0}^{n} x \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \sum_{x=0}^{n} x \frac{M!\binom{N-M}{n-x}}{x!\,(M-x)!\binom{N}{n}}$$

$$= \frac{M}{{}^{N}C_{n}} \sum_{x=0}^{n} \binom{M-1}{x-1}\binom{N-m}{n-x}$$

$$= \frac{M}{{}^{N}C_{n}} \sum_{x=1}^{n} \binom{M-1}{x-1}\binom{N-1-(M-1)}{n-1-(x-1)}$$

$$\therefore E(X) = \frac{M}{{}^{N}C_{n}}\binom{N-1}{n-1} = \frac{nM}{N}$$

Therefore, Mean of Hyper geometric distribution is $\frac{nM}{N}$

Similarly,

$$E(X(x-1)) = \sum_{x=0}^{n} x(x-1)\frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \sum_{x=0}^{n} x(x-1)\frac{M!\binom{N-M}{n-x}}{x!\,(M-x)!\binom{N}{n}}$$

$$= \frac{M(M-1)}{{}^{N}C_{n}} \sum_{x=2}^{n} \binom{M-2}{k-2}\binom{N-M}{n-k}$$

$$= \frac{M(M-1)}{{}^{N}C_{n}}\binom{N-2}{n-2} = \frac{M(M-1)\,n(n-1)}{N(N-1)}$$

$$\therefore E(X^2) = E\{x(x-1) + E(x)\} = \frac{M(M-1)\,n(n-1)}{N(N-1)} + \frac{nM}{N}$$

$$\text{Hence, } V(X) = E(X^2) - [E(x)]^2 = \frac{M(M-1)\,n(n-1)}{N(N-1)} + \frac{nM}{N} - \left(\frac{nM}{N}\right)^2$$

$$\text{Variance} = \frac{nM(N-M)(N-n)}{N^2(N-1)}$$

vii. Hyper geometric distribution tends to binomial; distribution if $N \to \infty$ and $\frac{M}{N} \to p$.

Example 14

A lot consisting of 50 bulbs is inspected by taking at random 10 bulbs and testing them. If the number of defective bulbs is at most 1, the lot is accepted, otherwise it is rejected. If there are, in fact, 10 defective bulbs in the lot, the probability of accepting the lot is given by hyper geometric distribution

$$\text{as } P(X=1) + P(X=0) = \frac{\binom{10}{1}\binom{40}{9}}{\binom{50}{10}} + \frac{\binom{40}{10}}{\binom{50}{10}} = 0.34871$$

Example 15

In an international film festival, a panel of 11 judges is formed to judge the best film. At least two films F_A and F_B were considered to be the best where the opinion of the judges got individual. Six judges were in favour of F_A whereas five in the favour of F_B. A random sample of five judges was drawn from the panel. Calculate the probability that out of five judges three are in favour of film F_A.

Solution

In the given problem total number of judge in population N = 11

Number of favorable cases for film F_A (M) =6

Number of favorable cases for film F_B (N - M) = 5, random sample without replacement (n) = 8 and k = 3 , then by using hyper-geometric distribution

$$P(X=x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

$$= \frac{\binom{6}{3}\binom{5}{2}}{\binom{11}{5}}$$

$$= 0.433$$

Exercise 11

1. Define Binomial distribution and find its mean and variance.
2. Define Poisson distribution and Show that mean and variance of Poisson distribution is equal to λ.
3. Define Binomial distribution and write down the importance properties of Binomial distribution.
4. Define Poisson distribution and write down the importance properties of Poisson distribution
5. What do you understand by Negative Binomial distribution; find its mean and variance also states the importance properties of Negative binomial distribution.
6. What do you understand by hyper-geometric distribution; fine its mean and variance also state the importance properties of hyper-geometric distribution
7. A machine produces 10 percent defective items; ten items are selected at random. Find the probability of not more than two items being defective.
8. The chances of a number hitting the target and missing the target are 3:2. Calculate the probability that the target will be hit at least once in five sots.
9. The incidence of occupational disease in an industry is such that workman have a 10% chance of suffering from it. What is the probability that in a group of 7, five or more will suffer from it ?
10. Suppose 10 percent of the items produced by a factory are defective. Suppose 4 items chosen at random. Find the probability that

    i. 2 are defective ii. 3 are defective iii. None are defective.
11. Assume that on the average one telephone number out of 15 called between 4 pm to 6 pm on Saturday is busy. What is the probability that if 6 random selected telephone numbers are called

    i. Not more than 3 called will be busy ?

    ii. At least 3 of them will be busy ?
12. If mean of Binomial distribution is 3 and variance is $\frac{3}{2}$. Find the probability of obtaining at least 4 success ?
13. Five unbiased coins are tossed 640 times. Fit a Binomial distribution to these data

No. of heads	0	1	2	3	4	5
Frequency	16	98	192	208	104	22

14. In 800 families with 5 children each with the assumption that probability for boy and girls birth is equal. how many families would be expected to have

    i. 3 boys and 2 girls ii. 2 boys and 3 girls

    iii. No girls iv. At the most two girls
15. Calculate the expected frequencies from the following data by using Binomial probability law.

No. of success(X)	5	4	3	2	1	0
Frequency	190	500	900	960	500	150

16. If the sum of mean and variance of a Binomial distribution is 4.8 for five trials. Find the probability distribution of it ?
17. The probability that a bomb dropped from a fighter plane will strike the target is $\frac{1}{5}$. If six bombs are dropped. Find the probability that

    i. Exactly two will strike the target ii. At least two will strike the target ?
18. If 6 of 18 new building in a city violate the building code. What is the probability that a building inspector who randomly selects 4 of the new building , for inspection, will catch

    i. None of the new building that violate the building code

    ii. 1 of the new building that violate the building code

    iii. 2 of the new building that violate the building code

    iv. At least 3 of the new building that violate the building code
19. The mean of Poisson distribution is 4. Find the probability that the variate takes the value (i) less than or equal to 3 (ii) equal to 5
20. If 2% of electric bulbs manufactured by a certain company are defective, find the probability that in a sample of 200 bulbs.

    a. Less than 2 bulbs are defective.

    b. At least two bulbs are defective.
21. Suppose 220 misprints are distributed randomly through out a book of 200 pages. Find the probability that a given page contains (a) no misprint (b) 1 misprint (c) 2 misprint (b) 2 or more misprint
22. Suppose 2 percentages of the people on the average are left handed. Find the Probability of 3 or more left handed among 100 people.

23. The probability that the noise level of a wide-band amplifier will exceed 2 dB is 0.05. Find the probabilities that among 12 such amplifiers the noise level of (a) one will exceed 2dB. (b) At most two will exceed 2dB. (c) Two or more will exceed 2dB.

24. A certain screw making machine produces on an average 2 defective screws out of 100 and packs them in boxes of 500. Find the probability that a box contains 15 defective screws.

25. From the past experience it is known that in a certain intersection there are on the average 4 traffic accidents per week. Find the probability that, in a given week there will be

    i. Less than 2 accidents ii. Exactly 2 accidents

    iii. More than 2 accidents

26. If the probability of a bad reaction from a certain injection is 0.001. Determine the chance that out of 200 individuals. More than two will get a bad reaction

27. Suppose it is known that in a certain area of a large city the average number or rats per quarter block is five. Assuming that the number of rats follows Poisson distribution, find the probability that in a randomly selected quarter block

    i. There are exactly five rats. ii. There are fewer than five rats

    iii. There are rats from five to seven inclusive.

28. Fit a Poisson distribution to the following data given the number of cells per square for 400 squares

No. of cells	0	1	2	3	4	5	6	7	8	9	10
No. of squares	103	143	98	42	8	4	2	0	0	0	0

29. Data was collected over a period of 10 years. Showing numner of deaths from horse kicks in each of the 200 army corps. The distribution of death was as follows

No of death	0	1	2	3	4
frequency	109	65	22	3	1

Fit a Poisson distribution to the data and calculate expected theoretical frequencies ?

30. Fit a Poisson distribution to the following data.

x:	0	1	2	3	4
f:	120	60	15	4	1

31. A shipment of 20 tape recorders contains 5 that are defective. If 10 of them are randomly chosen for inspection, what is the probability that 2 of the 10 will be defective? By hyper geometric probability law.

32. Among the 500 employees of a company 340 are union members, while the others are not. If 13 of the employees are chosen by lot to serve on the committee which administers the pension fund. Find the probability that 8 of them will be union members while the others are not. By using the hyper-geometric distribution.

33. A college library has 15 copies of certain introductory Statistics text book of which eight are first edition and seven are second edition. The course teacher has requested that five books be put on two days reserve. If the copies are selected at random, what is the probability that the book selected are (X = 0, 1, 2, 3, 4, 5) second edition. If such experiment is repeated in 60 college library fin the hyper-geometric distribution.

34. A committee of 5 is to be selected from a group of 6 men and 9 women. If the selection is made randomly what is the probability that the committee consists of 3 men and 2 women.

35. An experiment succeeds twice as often as it fails. What is the probability that in next five trials there will be three successes ?

Answers:

7. 0.929, **8.** 0.989, **9.** 0.00018

10. **i.** 0.0486, **ii.** 0.0036, **iii.** 0.6561, **11.** **i.** 0.9997 **ii.** 0.005

12. $\frac{11}{32}$

13. Expected frequencies are

x	0	1	2	3	4	5
f(x)	20	100	200	200	100	20

14. **i.** 250 **ii.** 250 **iii.** 25 **iv.** 400

15. Expected frequencies are

x	0	1	2	3	4	5

f(x)	100	500	1000	1000	500	100

16. $p = \frac{1}{5}$, $q = \frac{4}{5}$ and $n = 5$ **17.** i. 0.246 ii. 0.345

18. **i.** 0.197 **ii.** 0.395 **iii.** 0.296 **iv.** 0.111

19. 0.4335, 1563, **20.** **a.** 0.0916, **b.** 0.9084,

21. 0.333, 0.366, 0.201, 0.301, **22.** 0.325,

23. a. 0.3413, b. 0.9805, c. 0.1183 **24.** 0.035

25. **i.** 0.0916 **ii.** 0.1466 **iii.** 0.7618

26. 0.32 **27.** **i.** 0.1755 **ii.** 0.616 **iii.** 0.426

28. Expected frequencies are.

X	0	1	2	3	4	5	6	7	8	9	10
f	107	141	93	41	13	4	1	0	0	0	0

29. Expected frequencies are.

X	0	1	2	3	4
e(f)	109	66	20	4	1

30. Expected frequencies are.

X	0	1	2	3	4
e(f)	118	62	17	3	1

31. 0.348 **32.** 0.199

33. 0.0186, 0.163, 0.391, 0.326, 0.0932, 0.007

Expected frequencies are.

X	0	1	2	3	4	5
e(f)	1	10	23	20	6	0

34. 0.2398 **35.** $\frac{80}{243}$

Continuous Theoretical Probability Distribution

For continuous random variables we have number of theoretical probability distributions some of important Continuous theoretical probability distributions are listed below. But here we discuss Uniform, Normal, Gamma and Chi-square distribution in brief.

i. Continuous Uniform of Rectangular distribution.
ii. Normal distribution.
iii. Gamma distribution.
iv. Chi-square distribution.
v. Beta distribution.
vi. Exponential distribution.
vii. Weibul distribution.
viii. Logistic distribution.
ix. Cauchy distribution.

UNIFORM DISTRIBUTION

Definition: A continuous random variable X is said to have a continuous rectangular or continuous Uniform distribution over an interval (a, b) i.e. $(-\infty < a < x < b < \infty)$ if its probability density function is given by

$$f(x) = f(x, a, b) = \begin{cases} \frac{1}{b-a}, & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases}$$

Properties of Uniform distribution

The following are the important properties of Rectangular distribution

i. The two independent constants 'a' and 'b' in the distribution are known as the parameters of the distribution. The distribution is called uniform distribution on (a, b) since it assumes a constant (Uniform) value for all x in (a, b)

ii. The distribution is also known as rectangular distribution, since the curve y = f(x) describes a rectangle over the x-axis and between the ordinates at x = a, and x = b.

iii. It is continuous distribution as x can take the any values between (a, b) i.e. $a < x < b$.

iv. Any random variable which follows uniform distribution is known as uniform variate.

ii. The distribution is also known as rectangular distribution, since the curve y = f(x) describes a rectangle over the x-axis and between the ordinates at x = a, and x = b.

iii. It is continuous distribution as x can take the any values between (a, b) i.e. $a < x < b$.

iv. Any random variable which follows uniform distribution is known as uniform variate.

v. A uniform or rectangular variate X on the interval (a, b) is written as X~U(a, b) or X~R(a, b)

vi. The f(x) defined is pdf because $\int_a^b f(x)\,dx = 1$

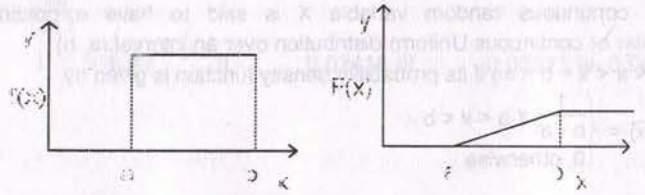
i.e. $\int_a^b \frac{1}{b-a}\,dx = \frac{1}{b-a}(b-a) = 1$

vii. The cumulative distribution function F(x) is given by

$$F(x) = \begin{cases} 0, & x \le a \\ \frac{x-a}{x-b}, & a < x < b \\ 1, & X \ge b \end{cases}$$

Since F(x) is not continuous at x = a and x = b, it is not differentiable at these points, Thus $\frac{dF(x)}{dx} = f(x) \neq 0$, exist everywhere.

viii. The graph of uniform pdf f(x) and the corresponding distribution function F(x) are given below.



ix. For a uniform variate X in (-a, a) the pdf is given by

$$f(x) = \begin{cases} \frac{1}{2a}, & \text{if } -a < x < a \\ 0, & \text{otherwise} \end{cases}$$

x. Mean and Variance of uniform distribution;

$$\text{Mean} = E(X) = \int_a^b xf(x)\,dx = \int_a^b x\frac{1}{b-a}\,\frac{(b^2 - a^2)}{2} = \frac{a+b}{2}$$

$$\text{And } E(X^2) = \int_a^b x^2 f(x)\,dx = \int_a^b x^2 \frac{1}{b-a}\,dx = \frac{a^2 + ab + b^2}{3}$$

$$\text{Therefore } V(X) = E(X^2) - [E(x)]^2 = \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \left(\frac{(b-a)^2}{12}\right)$$

xi. Median of uniform distribution is $\frac{a+b}{2}$, which is same as mean

xii. Mean deviation about mean $MD_{\bar{x}} = \frac{b-a}{2} = \int_a^b |x - \bar{x}|\, f(x)dx$

11.2.3 Moment Generating Function of Uniform Distribution

Moment generating function of uniform is given as

$$M_x(t) = E(e^{tX}) = \int_a^b e^{tX} f(x)\,dx = \int_a^b e^{tX}\frac{1}{b-a}\,dx = \frac{e^{tb} - e^{ta}}{b-a}$$

11.3 NORMAL DISTRIBUTION

The most important continuous probability distribution used in the entire fields of statistics is the Normal distribution. It was first discovered in 1733 by English Mathematician De-Moiver, who obtained this continuous distribution as a limiting case of the binomial distribution and applied it to problem arising in the game of chance, It was also known to Laplace, no later than 1774 but through a historical error it was credited to Gauss, who first made reference to it in the beginning of 19th century (1809) as the distribution of errors in Astronomy. Gauss used the normal curve to describe the theory of accidental errors of measurements involved in the calculation of orbits of heavenly bodies, throughout the eighteenth and nineteenth centuries, various efforts were made to establish the normal model as the underlying law ruling all continuous random variable, thus the name Normal. These efforts, however, failed because of false premises. The normal model has,

nevertheless, become the most important probability model in statistical analysis.

11.3.1Definition of Normal Distribution

A continuous random variable X is said to have a normal distribution with parameters μcalled mean and σ^2 called variance, if its pdf is given by the probability law as

$$f(x) = f(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} ; -\infty < x < \infty; \quad -\infty < \mu < \infty \ \& \ \sigma > 0$$

11.3.2 Important Characteristics (Properties) Of Normal Distribution and Normal Probability Curve

The normal probability curve with mean μ and variance σ^2 is given by the equation $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ and has the following properties.

i. The graph of probability f(x) of normal distribution is famous 'bell-shaped' curve. The top of the bell is directly above the mean μ and tails of curve of a normal distribution curve extended indefinitely on both sides of X = μ and never touch the x- axis. For large value of σ, the curve tends to flatten out and for small value of σ it has a sharp peak.

ii. Mean, median and mode of the distribution are coinciding.

iii. The two independent constants μ and σ in the distribution are known as the parameters of the distribution.

iv. Normal distribution is a continuous distribution as x can takes any value in the real line R i.e. $-\infty < x < \infty$

v. Any random variable which follows normal distribution is known as normal variate.

vi. We shall use the notation $x \sim N(\mu, \square^2)$ to denote that the random variable X follows normal distribution with mean μ and variance σ^2.

vii. The probability function f(x) defined in the distribution is pdf because

$$\int_{-\infty}^{\infty} f(x) = 1 \text{ i.e. } \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = 1$$

Let $\frac{(x-\mu)}{\sigma\sqrt{2}} = t, \quad \Rightarrow dx = \sqrt{2}\, dt\, \sigma$

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = \frac{\sqrt{2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t}\, dt = \frac{\sqrt{\pi}}{\sqrt{\pi}} = 1$$

viii. As x increase numerically, f(x) decrease rapidly the maximum probability occurring at the point X = μ and it is given by

$$P(x)|_{max} = \frac{1}{\sigma\sqrt{2\pi}}$$

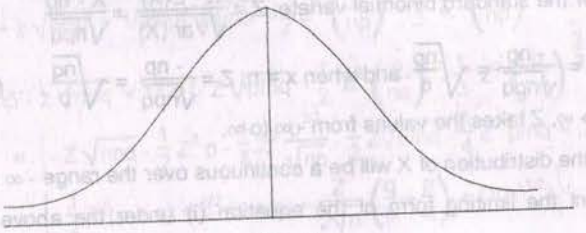
ix. The normal curve is symmetrical about the line X = μ , thus coefficient of skewness $\beta_1 = 0$ and $\lambda_1 = 0$. the normal distribution is masokurtic thus

$\beta_2 = 3$ and $\lambda_2 = 0$.

x. Since f(x) being the probability, can never be negative, no portion of the curve lies below the x –axis.

xi. Linear combination of independent normal variates is also a normal variate.

xii. X-axis is an asymptote to the normal curve.



Fig; normal probability curve

xiii. Mean deviation about mean = $\sqrt{\frac{2}{\pi}}\sigma = \frac{4}{5}\sigma$ (approx.)

xiv. Quartiles are $Q_1 = \mu - 0.6745\sigma$, $Q_3 = \mu + 0.6745$

xv. Quartile deviation Q.D. = $\frac{Q_2 - Q_1}{2} = \frac{2}{3}\sigma$ (approx).

xvi. Therefore QD: MD: SD = $\frac{2}{3}\sigma : \frac{4}{5}\sigma : \sigma$

Which implies QD:MD:SD = 10: 12: 15

11.3.3 Standard Normal Distribution

If X is a normal variate with parameters μ and σ^2 then random variable Z defined by $Z = \frac{x - \mu}{\sigma}$ is a Standard normal variate with mean zero (0) and variance unity (1) and the pdf of Standard normal distribution is given by

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < z < \infty$$

11.3.4 Normal Distribution as The Limiting Case Of Binomial Distribution

Normal distribution is another form of binomial distribution under the following conditions.

i. The number of trials n is indefinitely large i. e n $\square$ ∞ and

ii. Neither p nor q is very small.

The pmf of the binomial distribution with parameters n and p is given by $P(X=x) = \begin{cases} {}^nC_x p^x q^{n-x} \text{ for } x = 0,1,2.....n \\ 0, \text{ otherwise} \end{cases}$ .....(i)

Let us consider the standard binomial variate: $Z = \frac{X - E(X)}{\sqrt{Var(X)}} = \frac{X - np}{\sqrt{npq}}$

When $x = 0, Z = \frac{-np}{\sqrt{npq}} = \sqrt{\frac{np}{q}}$ and when $x = n, Z = \frac{-np}{\sqrt{npq}} = \sqrt{\frac{nq}{p}}$, ...(ii)

Thus when $n \to \infty$, Z takes the values from $-\infty$ to ∞.

Hence the distribution of X will be a continuous over the range $-\infty$ to ∞.

We want the limiting form of the equation (i) under the above two conditions

$$\therefore \lim_{n\to\infty} P(x) = \lim_{n\to\infty} \frac{n!}{(n-x)!} \frac{p^x q^{n-x}}{x!}$$ using Stirling's approximation to r! for large r as $\lim_{r\to\infty} r! = \sqrt{2\pi} e^{-r} r^{r+\frac{1}{2}}$

When $n \to \infty$ & consequently $x \to \infty$

$$\therefore \lim_{n\to\infty} P(x) = \frac{\sqrt{2\pi} e^{-n} n^{n+\frac{1}{2}} p^x q^{n-x}}{\sqrt{2\pi} e^{-x} x^{x+\frac{1}{2}} \sqrt{2\pi} e^{-(n-x)} (n-x)^{n-x+\frac{1}{2}}}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \frac{(np)^{x+\frac{1}{2}} (nq)^{n-x+\frac{1}{2}}}{x^{x+\frac{1}{2}} (n-x)^{n-x+\frac{1}{2}}}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{npq}} \left(\frac{np}{x}\right)^{x+\frac{1}{2}} \left(\frac{nq}{n-x}\right)^{n-x+\frac{1}{2}} \quad \text{.....} \quad \text{(iii)}$$

Also from equation (ii)

$$x = np + Z\sqrt{npq} \Rightarrow \frac{x}{np} = 1 + Z\sqrt{\frac{q}{np}} \text{ and } \frac{n-x}{nq} = 1 - Z\sqrt{\frac{p}{nq}}$$

Also $dz = \frac{1}{\sqrt{npq}} dx$.

Hence the probability differential of the distribution of Z, in the limits is;

$$dG(z) = g(z)\, dz = \lim_{n\to\infty} \left(\frac{1}{\sqrt{2\pi}} \frac{1}{N}\right) dz$$

Where $N = \left(\frac{x}{np}\right)^{x+\frac{1}{2}} \left(\frac{n-x}{nq}\right)^{n-x+\frac{1}{2}}$ which implies

$$\log(N) = (x + 1/2) \text{ Log } (x/nP) + (n - x + 1/2) \text{ Log } \{(n - x)/nq\}$$

$$= (np + z\sqrt{npq} + 1/2) \log \{(1+Z\sqrt{q/np} + \{np - z\sqrt{npq} + 1/2\} \log \{1-Z\sqrt{p/nq}\}$$

$$= (np + z\sqrt{npq} + 1/2) \{Z\sqrt{q/np} - \frac{1}{2}Z^2\left(\frac{q}{np}\right) + \frac{1}{3}Z^3\left(\frac{q}{np}\right)^{\frac{3}{2}} - .....\}$$

$$+ \{np - z\sqrt{npq} + 1/2\} \{-Z\sqrt{p/nq} - \frac{1}{2}Z^2\left(\frac{p}{nq}\right) - \frac{1}{3}Z^3\left(\frac{p}{nq}\right)^{\frac{3}{2}} - ... \}$$

$$= \left\{-Z\sqrt{npq} - \frac{1}{2}Z^2 p - \frac{1}{3}Z^3 \frac{p^{3/2}}{\sqrt{np}} - \frac{1}{2}Z\sqrt{p/nq} - \frac{1}{4}Z^2 p/nq + ...\right\}$$

$$= \left\{-\frac{1}{2}Z^2(p+q) + Z^2(p+q) + \frac{Z}{2\sqrt{n}}\left(\frac{q}{p} + \frac{p}{q}\right) + O(n^{-1/2})\right\}$$

$$= \left\{\frac{Z^2}{2} + O(n^{-1/2})\right\} = \left\{\frac{Z^2}{2}\right\} \text{ as } n\to\infty$$

$$\therefore \lim_{n\to\infty} \log(N) = \left\{\frac{Z^2}{2}\right\} \Rightarrow N = e^{\frac{z^2}{2}},$$

Substituting the value of logN in equation (iii)

We get $d\,G(z) = g(z)\, dz = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz$,

Hence, the probability function of Z is as follows,

$$g(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, -\infty < Z < \infty$$

This is the probability density function of normal distribution with mean zero and variance unity.

If X is normal variate with mean μ and standard deviation σ, then

$Z = \frac{X - \mu}{\sigma}$, is standard normal variate. The Jacobean of transformation of $Z = \frac{X - \mu}{\sigma}$ is $\frac{1}{\sigma}$

Hence the random variable X with mean $E(X) = \mu$ and variance $= \sigma^2$ is given by $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ $-\infty < x < \infty$

Remarks: Normal distribution can also be obtained as a limiting case of Poisson distribution with the parameter $\lambda \to \infty$

11.3.5 Continuity Correction to Improve the Approximation

The addition of 0.5 to the upper limits and subtraction of 0.5 to lower limits of the discrete random variable X when the continuous distribution is used as an approximation to the discrete distribution is called continuity correction and the factor 0.5 is called continuity correction factor.

Here in this book we consider Normal distribution as the limiting case of Binomial and Poisson distribution so the following are some of the examples of continuity correction.

i. $P(X = 20) = P(19.5 < X < 20.5)$

ii. $P(X \leq 20) = P(X < 20.5)$

iii. $P(X < 20) = P(X \leq 19.5)$

iv. $P(X > 20) = P(X \geq 20.5)$

v. $P(X \geq 20) = P(X > 19.5)$

vi. $P(20 < X < 25) = P(20.5 \leq X \leq 24.5)$

vii. $P(20 \leq X \leq 25) = P(19.5 < X < 25.5)$

11.3.6 Area Properties Of Normal Curve

The most important property of the normal probability curve is the area property. If the means and s.d of a normal distribution are μ and σ respectively then area under the normal curve between the ordinate at

$X = \mu - \sigma$ and $X = \mu + \sigma$ is 0.6826 that is
$P(\mu - \sigma < X < \mu + \sigma) = 0.6826$,
Similarly $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9544$
$P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9927$, which is almost unity.

The standard normal variate (SNV) defines as $Z = \frac{X - \mu}{\sigma}$ then we have

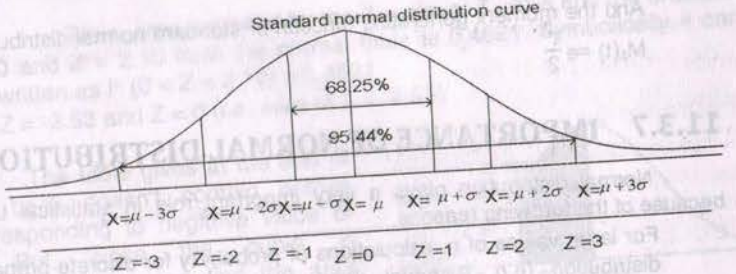
When $X = \mu - \sigma$, $Z = -1$
When $X = \mu + \sigma$, $Z = 1$
When $X = \mu - 2\sigma$, $Z = -2$
When $X = \mu + 2\sigma$, $Z = 2$
When $X = \mu - 3\sigma$, $Z = 3$, Hence area under the standard normal curve is

i. $P(-1 < Z < 1) = P(-1 < \frac{X - \mu}{\sigma} < 1) = P(\mu - \sigma < X < \mu + \sigma) = 0.6826$

ii. $P(-2 < Z < 2) = P(-2 < \frac{X - \mu}{\sigma} < 2) = P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9546$

iii. $P(-3 < Z < 3) = P(-3 < \frac{X - \mu}{\sigma} < 3) = P(\mu - 3 < X < \mu + 3\sigma) = 0.9927$,

These area are shown in the following normal probability curve



Since the location and shape of normal curve depends upon the values of its parameters (μ, σ), there are different bell-shaped curves with different means and standard deviations. In order to avoid the complicated task of preparing tables for each value of μ and σ a special form of normal distribution is used to convert this normal distribution in to one standardized form by making a very simple transformation, which is called standard normal distribution and corresponding curve as standard normal curve. The areas under the normal distribution denoted probability and can be obtained from the used of table prepared with reference to such standardized normal distribution. The slandered normal distribution table is given at the end of this book.

The table area under standard normal curve given at the end of book, which provides the area (Probability) between the two ordinate at Z = 0 and value of $Z = z_0$ (say) i.e. $P(0 < Z < Z_0)$.

If we have to find the area between x = 0 and x = 2.33 then from table $P(0 < Z < 2.33) = 0.4901$

Remarks: The total area under normal probability curve is unit i.e.

$$\int_{-\infty}^{\infty} f(x) = 1 \text{ i.e. } \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = 1$$

And $P(0 < Z < \infty) = P(-\infty < Z < 0) = 0.5$ [since the normal curve is symmetrical]

11.3.6 Moment Generating Function Of Normal Distribution

The moment generating function of normal probability distribution is defined as

$$M_x(t) = E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx = e^{\mu t \frac{t^2 - \sigma^2}{2}}$$

And the moment generating function of standard normal distribution is

$$M_z(t) = e^{\frac{t^2}{2}}$$

11.3.7 IMPORTANCE OF NORMAL DISTRIBUTION

Normal distribution plays a very important role in statistical theory because of the following reasons.

i. For large values of n, calculations of probability for discrete probability distribution (e.g. Binomial distribution, Poisson distribution, hyper geometric distribution etc,) become quite tedious and time consuming. In such cases, normal approximation can be used with great ease and convenience.

ii. For large value of n (i.e. $n \to \infty$), almost all of the exact sampling distribution e.g.- t-distribution , f-distribution and the chi-square distribution conform to normal distribution.

iii. The whole theory of exact sample (small sample) tests namely t, f, χ^2 tests etc is based on the assumption that the parent population from which the sample are drawn follows normal distribution.

iv. For large value of n (i.e. $n \to \infty$), the central limit theorem follows normal distribution.

v. It is extensively used in large sampling theory, to find the estimates of parameters from statistic and confidence limits etc.

vi Many of the distributions which are not normal can be made normal by simple transformation.

vii. It finds many applications in statistical quality control and industrial experiments for setting control limits.

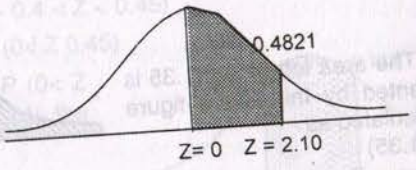
Example 1

Find the area when the value of Z is known.

a. Z = 0 and z = 2.10 (i.e. area of Z = 2.10)

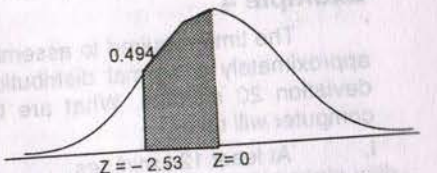
Solution

From the standard normal table, the area between Z = 0 and z = 2.10 is 0.4821. The shaded portion in the following figure denotes the area between. Z = 0 and z = 2.10



Thus, area bounded by the Z – curve, Z - axis and the ordinate Z = 0 and Z = 2.10 from the normal table is 0.4821. Symbolically it can be written as $P(0 < Z < 2.10) = 0.4821$

b. Z = -2.53 and Z = 0 (i.e. area of Z = -2.53)



The table gives at the end of the book doesn't contain entries corresponding to negative value of Z. But, since the curve is symmetrical we can find the area between Z = -2.53 and Z = 0 by looking the area corresponding Z = 0 and Z = 2.53. The shaded portion shown in the above figure represents the area between Z = -2.53 and Z = 0.

Thus, area between Z = -2.53 and Z = 0 from normal table is 0.4943. Symbolically, it can be written as $P(-2.53 < Z < 0) = P(0 < Z < 2.53) = 0.4943$ (by symmetry of normal Curve)

Example 2

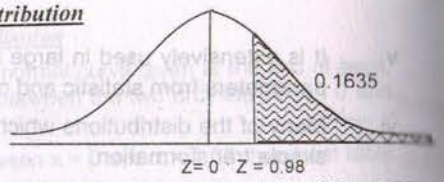
Find the area under the normal curve to the,

i) Right of Z = 0.98 and **(ii)** Left of Z = -1.05

Solution

In order to find the area under normal curve to the right of Z = 0.98 (i.e. right from the positive value of Z) , we proceed as,

$$P(Z > 0.98) = P(0 < Z < \infty) - P(0 < Z < 0.98) = 0.5 - 0.3365 = 0.1635$$



The Above shaded portion represents the area to the right of Z = 0.98.
Similarly, area left to Z = - 1.05 is

$P(Z < -1.05) = P(-\infty < Z < 0) - P(-1.05 < Z < 0)$

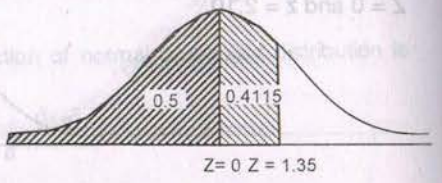
$= P(0 < Z < \infty) - P(0 < Z < 1.05)$

$= 0.5 - 0.3531 = 0.1469$. (By symmetry of normal curve)

Example 3

Find the area under the normal curve to the left of Z = 1.35

Solution



The area left of Z = 1.35 is represented by the above figure and calculated as

$P(Z < 1.35)$

$= P(-\infty < Z < 0) + P(0 < Z < 1.35)$

$= P(0 < Z < \infty) + P(0 < Z < 1.35)$

$= 0.5 + 0.4115 = 0.9115$.

Example 4

The time required to assemble a computer is a random variable having approximately a normal distribution with mean 129 minutes and standard deviation 20 minutes. What are the probabilities that the assembling of a computer will take

i. At least 124 minutes.

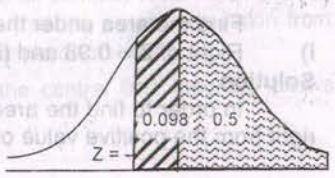
ii. Any where from 121 to 138 minutes.

Solution

The time required for assemble of a computer is a random variable (X)follows normal distribution with mean 129 minutes and standard deviation 20 minutes i.e. $X \approx N(\mu, \sigma^2)$, where $\mu = 129$ minutes and $\sigma = 20$ minutes.

i. The probability that time required to assemble a computer is at least 124 minutes i.e. $P(X \geq 124)$.

For $X = 124$, let $Z = \frac{X - \mu}{\sigma}$ is standard normal variate then



$Z = \frac{124 - 129}{20} = -0.25$, therefore the require probability is

$P(X \geq 124) = P(Z \geq -0.25) = P(-0.25 < Z < 0) + P(0 < Z < \infty)$

$= P(0 < Z < 0.25) + P(0 < Z < \infty)$ by symmetry of normal curve

$= 0.0987 + 0.5 = 0.5987$

ii. $P(121 < X < 138)$

For $X = 121$, let $Z = \frac{x - \mu}{\sigma}$ is standard normal variate then

$Z = \frac{121 - 129}{20} = -0.40$

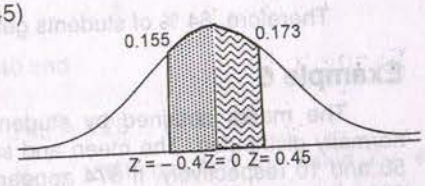
And $X = 138$, $Z = \frac{138 - 129}{20} = 0.45$, therefore the require probability is

$P(121 < X < 138) = P(-0.4 < Z < 0.45)$

$= P(-0.40 < z < 0) + P(0 < Z\ 0.45)$

$= P(0 < Z < 0.40) + P(0 < Z\ 0.45)$ by symmetry of the normal curve.

$= 0.1554 + 0.1736 = 0.329$.



Example 5

Students of a class are given a test. Their marks were normally distributed with mean 60 and standard deviation of 5. What percentage of students scored?

i. More than 60 marks.

ii. Less than 56 marks.

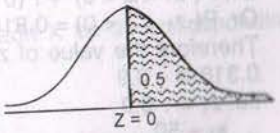
iii. Between 45 and 65 marks.

Solution

Let X denote the marks secure by student then X is a normal variate with mean $\mu = 60$ and standard deviation $(\sigma) = 5$. Let Z be the standard normal variate with mean 0 and variance 1 defined as

Then $Z = \frac{x - \mu}{\sigma} = \frac{x - 60}{5}$

i. For $P(X > 60) = P(\frac{x - \mu}{\sigma} > \frac{x - 60}{5}) = 0$

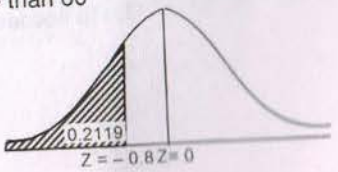


$\therefore P(X > 60) = P(Z > 0) = P(0 < Z < \infty) = 0.5$.

Therefore 50% students get marks more than 60

ii. $P(X < 56) = P(\frac{x - \mu}{\sigma} < \frac{56 - 60}{5}) = -0.8$

$\therefore P(X < 56) = P(Z < -0.8)$



$= P(-\infty < Z < 0) - P(-0.8 < Z < 0)$

$= P(0 < Z < \infty) - P(0 < Z < 0.8)$

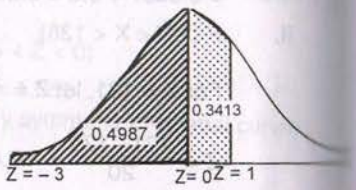
$= 0.5 - 0.2881 = 2119$

Therefore, 21.19 % of students gets marks less than 56.

iii. $P(45 < X < 65) = ?$

For $X = 45$, $Z = \frac{X - 60}{5} = \frac{45 - 60}{5} = -3$

For $X = 65$, $Z = \frac{X - 60}{5} = \frac{65 - 60}{5} = 1$,



$\therefore P(45 < X < 65) = P(-3 < Z < 1)$

$= P(-3 < z < 0) + P(0 < Z < 1)$

$= P(0 < Z < 3) + P(0 < Z < 1)$ by symmetry of normal curve

$= 0.4987 + 0.3413 = 0.84$

Therefore, 84 % of students get marks between 45 to 65.

Example 6

The marks obtained by students in B. E. entrance examination are normally distributed. The mean and standard deviation of the distribution are 50 and 10 respectively. If 674 appeared in the examination and out of these 550 are to be declared passed. What should be the minimum pass marks?

Solution

Let X denotes the normal variable marks. Let Z be standard normal variate of X. Then $Z = \frac{X - \mu}{\sigma}$, here $\mu = 50$ and $\sigma = 10$ $\therefore Z = \frac{X - 50}{10}$

Let minimum pass marks be x_1 .

Then $P(X > x_1) = \frac{550}{674} = 0.816$.

Let Z_1 be the corresponding SNV of x_1 then $Z_1 = \frac{x_1 - 50}{10}$

Now $P(X > x_1) = 0.816$.

Or $P(Z > -z_1) = 0.816$

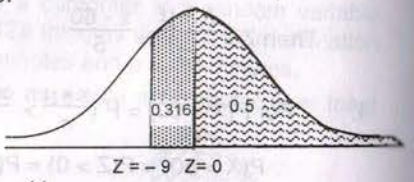
Or, $P(-z_1 < Z < 0) + P(0 < Z < \infty) = 0.816$.

Or, $P(-z_1 < Z < 0) = 0.816 - 0.5 = 0.316$

Therefore the value of z_1 at probability 0.316 is – 0.9

i.e. $z_1 = -0.9$

Or, $\frac{x_1 - 50}{10} = -0.9$



$X_1 = 41$. Hence the minimum pass marks = 41.

Example 7

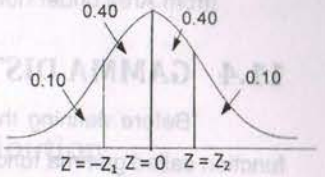
The mean and variance of height of 500 engineering students were found to be 165 cm and 25 cm^2 respectively. Find the range of height of middle 80 % of the students.

Solution

Let X denotes the normal variable of height of engineering students and Let Z be standard normal variate of X. Then

$Z = \frac{X - \mu}{\sigma}$, here $\mu = 165$ and $\sigma = 25$

$\therefore Z = \frac{X - 165}{5}$



Let x_1 and x_2 be the lower and upper limit of height of middle 80% of the students. Then $P(x_1 < X < x_2) = 0.80$ ... (i)

Let Z_1 and Z_2 be the SNV of x_1 and x_2 respectively

Then, $-Z_1 = \frac{x_1 - 165}{5}$ and $Z_2 = \frac{x_2 - 165}{5}$

$\therefore P(x_1 < X < x_2) = 0.80$

or, $P(-Z_1 < Z < Z_2) = 0.80$

or, $P(-Z_1 < Z < 0) = P(0 < Z < Z_1) = 0.40$ and

$P(0 < X < Z_2) = 0.40$

$\therefore$ The value of Z_1 at probability 0.40 is 1.28 $\therefore Z_1 = 1.28$.

Or, $\frac{x_1 - 165}{5} = -1.28$, Which implies $X_1 = 158.6$ cm And the value Z_2 at probability 0.40 is 1.28 $\therefore Z_2 = 1.28$.

Or, $\frac{x_2 - 165}{5} = 128$

Which implies $X_2 = 171.4$ cm, Therefore, required range of height of middle 80% students is equal to 171.4 – 158.6 = 12.8 cm.

Example 8

The probability that an electronic component will fail in less than 1,000 hours of continuous use is 0.25. Use the normal approximation to find the probability that among 200 such components fewer than 45 will fail in less than 1,00 hours of continuous use

Solution

let ,X be the continuous normal variable adn X be its corresponding Standard normal variate then $Z = \frac{X - \mu}{\sigma}$

Here Mean $\mu = np = 200 \times .25 = 50$

Standard Deviation $\sigma = npq = 200 \times 0.25 \times 0.75 = 37.5$

Here We have to calculate the P (X < 45), for Normal Approximation of Binomial distribution we use the continuity correction of 0.5

Here, $Z = \frac{44.5 - 50}{37.5} = 0.146$

so $P(X < 45) = P(X \leq 44.5) = P(Z \leq -0.146)$

$= 0.4404,$

[from Area under normal table]

11.4 GAMMA DISTRIBUTION

Before defining the gamma distribution it is essential to introduce a function called gamma function which plays an important role in many ranches of mathematics, which is as follows,

$$\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx \qquad \text{for all } n > 0$$

The most important properties of gamma function are

i. For any $n > 0$ we have $\Gamma n = n \Gamma n-1)$

ii. For any positive integer n we have $\Gamma n = (n-1)!$

iii. $\Gamma\frac{1}{2} = \sqrt{\pi}$

11.4.1 Definition of Gamma Distribution With Two Parameters

A continuous random variable X is said to have a gamma distribution with two parameter λ and α, if it assumes only non- negative values and its probability density function is defined as

$$F(x) = f(x, \alpha, \lambda) = \begin{cases} \frac{\alpha^\lambda e^{-\alpha x} x^{\lambda-1}}{\Gamma(\lambda)}, & 0 \leq x < \infty \\ 0, \text{ otherwise} \end{cases} \qquad \lambda > 0, \ \alpha \quad 0,$$

If we take $\alpha = 1$ then pdf is reduces as

$$f(x, \lambda) = \begin{cases} \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)}, & \lambda > 0, 0 \leq x < \infty \\ 0, \text{ otherwise} \end{cases}$$

11.4.2 Definition of Gamma Distribution with Single Parameter

A continuous random variable X is said to have a gamma distribution with single parameter λ, if it assumes only non- negative values and its probability density function is defined as

$$f(x) = f(x, \lambda) = \begin{cases} \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)}, \lambda > 0, 0 \leq x < \infty \\ 0, \text{ otherwise} \end{cases}$$

11.4.3 Properties of Gamma Distribution

Following are the important properties of gamma distribution.

i. The Game distribution depends upon the two constants α and λ are called the parameter of the Gamma distribution.

ii. Like normal distribution gamma distribution is also a continuous probability distribution.

iii. Any random variable which follows gamma distribution is known as gamma variate.

iv. We shall use the notation $X \approx G(\alpha, \lambda)$ to denote that random variable X follows Gamma distribution with parameters α and λ and also $X \approx G(\lambda)$ for single parameter λ.

iv. The $f(x, \lambda)$ defined in gamma distribution is pdf because

$$\int_0^\infty f(x, \lambda)\, dx = \frac{1}{\Gamma(\lambda)} \int_0^\infty e^{-x} x^{\lambda-1} dx = \frac{\Gamma(\lambda)}{\Gamma(\lambda)} = 1$$

vi. The cumulative distribution function, called incomplete gamma function is defined as

$$F(x) = P(X \leq x) = \int_0^x f(x, \lambda) dx = \frac{1}{\Gamma(\lambda)} \int_0^x e^{-x} x^{\lambda-1} dx, \ x>0$$

vii. **Mean of Gamma distribution with two parameters**

We have $f(x, \alpha, \lambda) = \begin{cases} \frac{\alpha^\lambda e^{-\alpha x} x^{\lambda-1}}{\Gamma(\lambda)}, \lambda > 0, \alpha \leq 0, 0 \leq x < \infty \\ 0, \text{ otherwise} \end{cases}$

$$\text{Mean} = E(X) = \int_a^b x f(x)\, dx = \int_0^\infty x \frac{\alpha^\lambda e^{-\alpha x} x^{\lambda-1}}{\Gamma(\lambda)} dx$$

$$= \frac{\alpha^{\lambda}}{\Gamma(\lambda)} \int_0^{\infty} e^{-\alpha x} x^{\lambda+1-1}\, dx$$

Let, $\alpha x = y$ Then, $dx = \frac{dy}{\alpha}$

$$\text{Mean} = E(X) = \frac{\alpha^{\lambda}}{\Gamma(\lambda)} \int_0^{\infty} e^{-y} \left(\frac{y}{\alpha}\right)^{\lambda+1-1} \frac{dy}{\alpha} = \frac{\lambda}{\alpha}$$

$$\text{Variance}(X) = E(X^2) - \{E(X)\}^2$$

$$\text{Now, } E(X^2) = \int_a^b x^2 f(x)\, dx = \int_0^{\infty} x^2 \frac{\alpha^{\lambda} e^{-\alpha x} x^{\lambda-1}}{\Gamma(\lambda)} dx$$

$$= \frac{\alpha^{\lambda}}{\Gamma(\lambda)} \int_0^{\infty} e^{-\alpha x} x^{\lambda+2-1}\, dx$$

Let, $\alpha x = y$ Then, $dx = \frac{dy}{\alpha}$

$$E(X^2) = = \frac{\alpha^{\lambda}}{\Gamma(\lambda)} \int_0^{\infty} e^{-y} \left(\frac{y}{\alpha}\right)^{\lambda+2-1} \frac{dy}{\alpha} = \frac{\lambda(\lambda+1)}{\alpha^2}$$

$$\therefore \text{variance}(X) = E(X^2) - \{E(X)\}^2 = \frac{\lambda(\lambda+1)}{\alpha^2} - \left(\frac{\lambda}{\alpha}\right)^2 = \frac{\lambda}{\alpha^2}$$

Mean of Gamma distribution with single parameter

$$\text{We have } f(x, \lambda) = \begin{cases} \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)}, \lambda > 0, 0 \le x \le \infty \\ 0, \text{ otherwise} \end{cases}$$

$$\text{Mean} = E(X) = \int_a^b x f(x)\, dx = \int_0^{\infty} x \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)} dx$$

$$= \int_0^{\infty} \frac{e^{-x} x^{\lambda+1-1}}{\Gamma(\lambda)} dx = \frac{\Gamma(\lambda+1)}{\Gamma(\lambda)} = \lambda$$

$$\text{Variance}(X) = E(X^2) - \{E(X)\}^2$$

$$\text{Now, } E(X^2) = \int_a^b x^2 f(x)\, dx = \int_0^{\infty} x^2 \frac{e^{-x} x^{\lambda-1}}{\Gamma(\lambda)} dx$$

$$= \int_0^{\infty} \frac{e^{-x} x^{\lambda+2-1}}{\Gamma(\lambda)} dx = \frac{\Gamma(\lambda+2)}{\Gamma(\lambda)} = \lambda(\lambda+1)$$

$$\therefore \text{Variance}(X) = E(X^2) - \{E(X)\}^2 = \lambda(\lambda+1) - \lambda^2 = \lambda$$

viii. As $\lambda \to \infty$ gamma distribution tends to normal distribution.

ix. If $x_1, x_2 \ldots\ldots\ldots x_n$ are identically independently distributed (iid) gamma variate with parameter α, λ_i i= 1, 2.....n then $\Sigma x_i \approx G(\alpha, \Sigma\lambda_i)$

Remarks:

Like Poisson distribution, the mean and variance of gamma distribution are also equal. However Poisson distribution is discrete while gamma distribution is continuous one.

11.4.4 Moment Generating Function Of Gamma Distribution

The moment generating function of gamma distribution is defined as follows

$$M_x(t) = E(e^{tx}) = \int_0^{\infty} e^{tx} f(x) dx = \frac{1}{\Gamma(\lambda)} \int_0^{\infty} e^{tx} e^{-x} x^{\lambda-1}\, dx$$

$$= \frac{1}{\Gamma(\lambda)} \int_0^{\infty} e^{-(1-t)x} x^{\lambda-1}\, dx$$

$$= \frac{\Gamma}{\Gamma(\lambda)} \frac{\Gamma(\lambda)}{(1-t)^{\lambda}} = (1-t)^{-\lambda}$$

Mean and variance of gamma distribution by using gamma distribution

$$\therefore \text{Mean} = E(X) = \left.\frac{\partial M_x(t)}{\partial t}\right|_{t=0} = -\lambda(1-t)^{-\lambda-1}(-1)\Big|_{t=0} = \lambda$$

$$\text{Again } E(X^2) = \left.\frac{\partial^2 M_x(t)}{\partial t^2}\right|_{t=0} = \lambda(1-\lambda)(-\lambda)^{-\lambda-2}(-1)\Big|_{t=0} = \lambda(1+\lambda)$$

$$\therefore \text{Var}(X) = E(X^2) - [E(X)]^2 = (\lambda + \lambda^2) - \lambda^2 = \lambda$$

$\therefore$ Mean and variance of gamma distribution is equal to λ

Example 8

If a random variable has the gamma distribution with $\alpha = 2$, $\lambda = 2$, find the mean and the standard deviation of the distribution. Also find the probability that the random variable will take on a value less than 4.

Solution

X follows gamma distribution with two parameter $\alpha = 2$, $\lambda = 2$ then its pdf is defined as

$$f(x) = \frac{a^{\lambda} e^{-ax} x^{\lambda-1}}{\Gamma(\lambda)} = \frac{2^2 e^{-2x} x^{2-1}}{\Gamma(2)} = \frac{2^2 e^{-2x} x^{2-1}}{\Gamma(2)} = 4 \times e^{-2x}$$

i. Mean = $E(X) = \int_0^\infty x\, 4e^{-2x}\, x\, dx = 1$

And $E(X^2) = \int_0^\infty x^2\, 4e^{-2x}\, x\, dx = \frac{6}{4}$

Then $var(X) = E(X^2) - [E(X)]^2 = 6/4 - (1)^2 = 0.5$

ii. The probability that the random variable will take on a value less than 4.

$$P(X < 4) = \int_0^4 4e^{-2x}\, x\, dx = 0.247$$

Example 9

The daily consumption of milk in a city in excess of 20,000 liters is approximately distributed as gamma variate with parameter 1/100,000 and 2. The city has a daily stock of 30,000 liters. What is the probability that the stock is insufficient as a particular day?

Solution

If the r.v. x denotes the daily consumption of milk (in liter) in a city then r.v y = x – 20,000 has a gamma distribution with p.d.f

$$g(y) = \frac{\alpha^\lambda}{\Gamma(\lambda)} e^{-\alpha y} y^{\lambda-1} = \frac{1}{(10,000)^2\, \Gamma 2} y2\text{-}1\, e^{-\frac{y}{10,000}}$$

$$= \frac{y\, e\text{-}y/10,000}{(10,000)2}, \quad 0 \le y < \infty$$

Since the daily stock of the city in 30,000 liters, the required probability 'p' that the stock is insufficient on a particular day is given by

$$P(x > 30,000) = P(y > 10,000)$$

$$= \int_{10,000}^{\infty} g(y)\, dy = \int_{10,000}^{\infty} \frac{y\, e^{-y/10,000}}{(10,000)^2}\, dy = \frac{2}{e}$$

11.5 CHI- SQUARE DISTRIBUTION

The chi-square distribution was first discovered by Helmert in 1876 and later independently by Karl person in 1900. It is denoted by χ^2.

The square of a standard normal variate is known as a chi-square variate (pronounced as ki - sky without s) with 1 degree of freedom (d.f.). Thus if Z is SNV with mean 0 and variance 1 than Z^2 is known as the chi-square variate.

Further, If $x \approx N(\mu, \sigma^2)$, then $Z = \frac{x-\mu}{\sigma} \approx N(0,1)$ and $Z^2 = \left(\frac{x-\mu}{\sigma}\right)^2$ is a chi- square variate with 1 degree of freedom

In general if $x_1, x_2 \ldots\ldots\ldots x_n$ are n independent normal variates with mean μ_i and variances. σ_i (i = 1, 2.....n), then $Z_1, Z_2 \ldots\ldots\ldots Z_n$ are independent SNV &

$$\chi^2 = Z_1^2 + Z_2^2 + \ldots + Z_n^2 = \sum Z_i^2 = \sum \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

is a chi-square variate with n. d.f. and it's p.d.f. is defined as,

$$f(\chi^2) = \frac{1}{2^{\frac{n}{2}}\, \Gamma_{n/2}}\, e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}; \quad o \le \chi^2 < \infty.$$

If we suppose the chi-square variate $\chi^2 = x$ then, the probability density function can be written as

$$f(x) = \frac{1}{2^{\frac{n}{2}}\, \Gamma n/2}\, e^{-\frac{x}{2}}\, x^{\frac{n}{2}-1} \quad o \le x < \infty$$

11.5.1 Derivation of the Probability Density of Chi-Square Distribution

If $x_1, x_2, \ldots\ldots x_n$ are independent $N(\mu i, \sigma i^2)$ we want the distribution of

$$\chi^2 = \sum \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 = \sum z_i^2 \text{ where } z_i = \frac{x_i - \mu_i}{\sigma_i} \approx N(0,1)$$

Since x_i are independent then z_i's are also independent. Therefore,

$M_{\chi}^2(t) = M_{\sum z_i}^2(t) = \prod M_{z_i}^2(t) = [M_{z_i}^2(t)]^n$

[Since z_i are identically independent distributed N(0,1)]

Also $M_Z^2(t) = E(e^{tz_i^2}) = \int_{-\infty}^{\infty} e^{tz_i^2} f(x_i)\, dx_i$

$$= \int_{-\infty}^{\infty} e^{tz_i^2} \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{(x_i - \mu_i)^2}{2\sigma_i^2}}\, dx_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tz_i^2}\, e^{-\frac{z_i^2}{2}}\, dz_i \qquad \text{where } z_i = \frac{x_i - \mu_i}{\sigma_i}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{1-2t}{2} z_i^2}\, dz_i$$

$$= \frac{1}{\sqrt{2\pi}} \frac{\sqrt{\pi}}{\left(\frac{1-2t}{2}\right)^{\frac{1}{2}}} = (1-2t)^{-\frac{1}{2}} \quad \text{Since } \int_{-\infty}^{\infty} e^{-a^2x^2} dx = \frac{\sqrt{\pi}}{a}$$

$$\therefore M_{\chi}^2(t) = (1-2t)^{-\frac{n}{2}}$$

Which is the moment generating function of a gamma variate with parameter $\frac{1}{2}$ and $\frac{n}{2}$.

Hence by uniqueness theorem of moment generating function of chi-square distribution as

$\chi^2 = \sum\left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$ is a gamma variate with parameter $\frac{1}{2}$ and $\frac{n}{2}$

$$\therefore f(\chi^2) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\frac{n}{2}} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1} \quad f(\chi^2) = \frac{1}{2^{\frac{n}{2}}\Gamma_{n/2}} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}; \; o \le \chi^2 < \infty.$$

$$f(\chi^2) = \frac{1}{2^{\frac{n}{2}}\Gamma_{n/2}} e^{-\frac{\chi^2}{2}} (\chi^2)^{\frac{n}{2}-1}; \; o \le \chi^2 < \infty.$$

11.5.2 Properties of Chi-Square Distribution

i. A random variable X which follows chi-square distribution is also known as chi-square variate.

ii. A chi-square variate X is continuous random variable which assumes non-negative variables only i.e. $o \le x < \infty$.

iii. We shall use the notation $X \approx \chi^2(n)$ to denote that random variable X follows chi- square distribution with n- degrees of freedom.

iv. Mean and variance of chi square distribution are **n** and **2n** respectively.

We have the probability density function of chi-square distribution with n-degrees of freedom as

$$f(x) = \frac{e^{-\frac{x}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} \quad o \le x < \infty$$

$$\text{Mean} = E(X) = \int_a^b x\, f(x)\, dx = \int_0^\infty x \frac{e^{-\frac{x}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} dx = \int_0^\infty \frac{e^{-\frac{x}{2}} x^{\frac{n}{2}}}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} dx.$$

Let $\frac{x}{2} = y$ which implies $dx = 2\, dy$

$$E(X) = \int_0^\infty \frac{e^{-y}(2y)^{\frac{n}{2}}}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} 2\, dy$$

$$= 2\int_0^\infty \frac{e^{-y}(y)^{\frac{n}{2}}}{\Gamma\frac{n}{2}} = 2\frac{\Gamma\left(\frac{n}{2}+1\right)}{\Gamma\frac{n}{2}} = n$$

$$\text{Variance}(X) = E(X^2) - \{E(X)\}^2$$

$$\text{Now, } E(X^2) = \int_a^b x^2 f(x)\, dx = \int_0^\infty x^2 \frac{e^{-\frac{x}{2}} x^{\frac{n}{2}-1}}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} dx = \int_0^\infty \frac{e^{-\frac{x}{2}} x^{\frac{n}{2}+1}}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} dx$$

Let $\frac{x}{2} = y$ which implies $dx = 2\, dy$

$$\therefore E(X^2) = \int_0^\infty \frac{e^{-y}(2y)^{\frac{n}{2}+1}}{2^{\frac{n}{2}}\Gamma\frac{n}{2}} 2\, dy = 4\int_0^\infty \frac{e^{-y}(y)^{\frac{n}{2}+1}}{\Gamma\frac{n}{2}}$$

$$= 4\frac{\Gamma\left(\frac{n}{2}+2\right)}{\Gamma\frac{n}{2}} = 4\, n\left(\frac{n+2}{2}\right)\frac{n}{2} = n(n+2)$$

$\therefore$ Variance(X) $= E(X^2) - \{E(X)\}^2 = n(n+2) - n^2 = 2n$

Thus, mean and variance of chi-square distributions with 'n' degrees of freedom is n and 2n respectively.

$\therefore$ Variance $= E(x^2) - E(x)^2 = n^2 + 2n - n^2 = 2n$

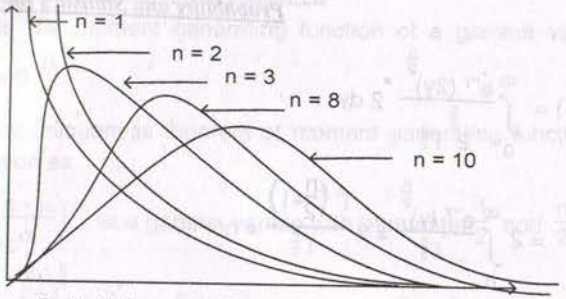
$\therefore$ Mean & variance of chi-square distribution is n & 2n respectively

ii. Mode of the chi-square distribution curve is lies at the point $\chi^2 = n - 2$ for $n > 2$

iii. Measure of Skewness is $\frac{8}{n}$ and Kurtosis is $\frac{12}{n}$

Therefore, Chi-square curve is positive Skewness & Leptokurtic

iv. The chi-square distribution curve with 1 and 2 degree of freedom is hyperbolic in shape where as for $n > 2$, the curve is un-symmetric bell shaped.



Probability curve of chi-square distribution

v. If $x_1, x_2, \ldots\ldots x_k$ are k iid chi square variate with n_i d.f then $\sum x_i$ in also a chi-square variate with $\sum n_i$ d.f. this is known as additive property of chi-square.

vi. It X & Y are two independent distribution chi square variate the with n_1 & n_2 d.f. then $(X \pm Y)$ is also a chi-square with $(n_1 \pm n_2)$ d.f.

vii. If n is large then $\lim_{n\to\infty} \chi_n^2 \to N(n, 2n)$. This is known as the limiting property of the chi-square.

viii. The chi-square distribution is used to test whether a hypothetical value of the population variance is true or not.

ix. It is used to test a test of goodness of fit.

x. It is used to test the independent of attributes.

xi. It is used to test the homogeneity of several population variances.

xii. It is used to test the equality of several population correlation coefficients.

11.5.3 Moment Generating Function of Chi-Square Distribution

We moment generating function of chi-square distribution as

$$M_X(t) = (1- 2t)^{-\frac{n}{2}}$$

Here, we alternatively find the value of mean and variance by using moment generating function.

$$\text{Mean} = E(x) = \frac{\partial M_x(t)}{\partial t}\Big|_{t=0} = -\frac{n}{2}(1 - 2t)^{-\frac{n}{2}\cdot 1}(-2)\Big|_{t=0} = n$$

$$\text{And} \quad E(x^2) = \frac{\partial^2 M_x(t)}{\partial t^2}\Big|_{t=0} = n\left(-\frac{n}{2} - 1\right)(1 - 2 + 1)^{-\frac{n}{2}\cdot 2}(-2)\Big|_{t=0}$$

$$= n\left(\frac{n+2}{2}\right)2 = n(n+2)$$

EXERCISE 11

1. Define the term Normal distribution, Standard normal distribution and state the importance properties of normal distribution.

2. Define normal distribution, standard normal distribution, Gamma distribution and Chi- square distribution.

Numerical of Normal Distribution

3. Suppose the temperature during may in Kathmandu is normally distributed with mean 35^{oc} and standard deviation 4^{oc}. Find the probability that the temperature during may in Kathmandu is (a) less than 32 (b) between 38 and 47

4. In a photo graphic process, the developing time of prints may be looked upon as a random variable having the normal distribution with a mean of 16.28 seconds and a standard deviation of 0.12 second. Find the probability that it will take
   i. Anywhere from 16.00 to 16.50 seconds to develop one of the prints.
   ii. At least 16.20 seconds to develop one of the prints.
   iii. At least 16.35 seconds to develop one of the prints.

5. Extruded plastic rods are automatically cut into normal lengths of 6 inches. Actual lengths are normally distributed about a mean of 6 inches and their standard deviation 0.06 inch. What percentages of the rods exceed the length of 6.05 inches?

6. The weekly income of a group of 10,000 persons was found to be normally distributed with mean Rs 750 and standard deviation of Rs 50
   i. If 5% are the poorest find the minimum income of the rich people.
   ii. If 5% are the richest find the minimum income of the richest people.
   iii. What was the lowest income among the richest 100 ?

7. Suppose the average length of stay in hospital for a certain chronic disease of a patient is 60 days with standard deviation of 15 days. If it is assumed that the length of stay is approximately normally distributed. Find the portability that the randomly selected patient from this group will have a length of stay
   i. More than 50 days. ii. Less than 30 days.
   iii. Between 30 and 60 days

xi. It is used to make the homogeneity of several population variances.

xii. It is used to test the equality of several population correlation coefficients.

Moment generating function of chi-square distribution

We moment generating function of chi-square distribution as

$$M_X(t) = (1- 2t)^{-\frac{n}{2}}$$

Here, we alternatively find the value of mean and variance by using moment generating function.

$$\text{Mean} = E(x) = \frac{\partial M_x(t)}{\partial t}\Big|_{t=0} = -\frac{n}{2}(1 - 2t)^{-\frac{n}{2}-1}(-2)\Big|_{t=0} = n$$

$$\text{And } E(x^2) = \frac{\partial^2 M_x(t)}{\partial t^2}\Big|_{t=0} = n\left(-\frac{n}{2} - 1\right)(1 - 2 + 1)^{-\frac{n}{2}-2}(-2)\Big|_{t=0}$$

$$= n\left(\frac{n+2}{2}\right)2 = n(n+2)$$

Exercise 12

1. Define the term Normal distribution, Standard normal distribution and state the importance properties of normal distribution.
2. Define normal distribution, standard normal distribution, Gamma distribution and Chi- square distribution.
3. Suppose the temperature during may in Kathmandu is normally distributed with mean 35^{oc} and standard deviation 4^{oc}. Find the probability that the temperature during may in Kathmandu is (a) less than 32 (b) between 38 and 47
4. In a photo graphic process, the developing time of prints may be looked upon as a random variable having the normal distribution with a mean of 16.28 seconds and a standard deviation of 0.12 second. Find the probability that it will take
   i. Anywhere from 16.00 to 16.50 seconds to develop one of the prints.
   ii. At least 16.20 seconds to develop one of the prints.
   iii. At least 16.35 seconds to develop one of the prints.

5. Extruded plastic rods are automatically cut into normal lengths of 6 inches. Actual lengths are normally distributed about a mean of 6 inches and their standard deviation 0.06 inch. What percentages of the rods exceed the length of 6.05 inches?
6. A fair die is tossed 180 times. Determine the probability that the face 6 will appear (a) between 29 and 32 times inclusive (b) between 31 and 35 times inclusive(c) less than 22 times.

   {Hints ; P(29 < X < 32) = P(28.5<X<31.5) convert X values (BD) to standard units}
7. The income of a group of 10,000 persons was found to be normally distributed with mean Rs 750 per month & standard deviation of Rs 50 of this group show that.
   i. About 95% had income exceeding Rs 668.
   ii. Only 5% had income exceeding Rs 832.
   iii. What was the lowest income among the richest 100 ?
8. Suppose the average length of stay in hospital for a certain chronic disease of a patient is 60 days with standard deviation of 15 days. If it is assumed that the length of stay is approximately normally distributed. Find the portability that the randomly selected patient from this group will have a length of stay
   i. More than 50 days. ii. Less than 30 days.
   iii. Between 30 and 60 days
9. In a normal distribution, 31% of the items are under 45 and 8% are over 64. Find the mean and the standard deviation of the distribution.
10. In a normal distribution 7% of the items are under 35 & 89 % are under 63. Find the mean and the standard deviation of the distribution.
11. A manufacturer known from experience that the resistance of resistors he produces is normal with mean $\mu = 100$ and standard deviation $\sigma = 2$ ohms. What percentages of resistors will have resistance between 98 ohms and 102 ohms ?
12. A sample of 100 mobile battery cells tested to find the length of life produced the following results as mean 13 months and standard deviation of 3 months. Assuming the data to be normally distributed what percentage of battery cells expected to have life?
    i. More than 15 months. ii. Less than 6 months.
13. The mean inside diameter of a sample of 200 washers produced by a machine is 0.502 cm and the standard deviation as 0.005 cm. The purpose for these washers are interned allows a maximum tolerance in

the diameter of 0.496 to 0.508 cm, otherwise the washers are considered defective. Determine the percentage of defective washers produced by the machine. Assuming the diameters are normally distributed.

14. The mean height of 500 male students in a college is 151 cm and the standard deviation is 15 cm. Assuming heights are normally distributed. Find how many students have heights between 120 cm to 155 cm.

15. Suppose the duration of trouble free operation of a new vacuum cleaner is normally distributed with mean 530 days and standard deviation of 100 days

    i. What is the probability that the vacuum cleaner will work for at least two years without trouble.

    ii. The company wishes to set the warranty period so that no more than 10% of the vacuum cleaners would need repair services while under warranty. How long a warranty period must be set ?

16. The number of calories in a salad on the lunch menu of a hostel of a college is normally distributed with mean 200 and standard deviation of 5. Find the probability that the salad you select will contain

    i. More than 208 calories ii. Between 190 and 200 calories

17. Let X is a normal variate with mean 25 and standard deviation 10. Find the value of constant 'b' such that the probability of the interval [25, b] is 0.4772.

18. The mean of normal distribution is 500 and 16 percentage of the values are greater than 600. What is the standard deviation of the distribution?

Answers:

3.	0.2266, 0.2253		
4.	**i.** 0.9568. **ii.** 0.7475, **iii.** 0.7201	**5.**	0.2033
6.	0.3094, 0.3245, 0.0446	**7.**	866.5
8.	**i.** 0.7487 **ii.** 0.0228, **iii.** 0.4772	**9.**	50; 10
10.	50.5 and 10.33	**11.**	68.26%
12.	**i.** 15.87 **ii.** 2.28%		
13.	23.02%	**14.**	300
15.	**i.** $P(X \geq 730) = 0.023$ **ii.** 658 days	**16.**	**i.** 0.0548 **ii.** 0.4772
17.	45	**18.**	100

Chapter 13

Theory of Estimation

Sampling

Sampling is a part of our day to day life which we use advertently or inadvertently. For example, a housewife takes one or two grains of rice from the cooking pan and decides whether the rice is cooked or not . A quality controller takes a few items and decides whether the lot is in accordance with the desired specifications or not. A pathologist takes a few drops of blood and tests for any change in blood of the whole body then normal. In all these situations sampling is inevitable and gives satisfactory results.

Even in those cases where complete enumeration is possible , it is not preferred due to the facts that it is much more time consuming and expensive, requires more skilled and technical personal, more errors are caused due to greater volume of work, measurement error etc. Complete enumeration is used only for various censuses or in case of small population

Sampling is a part or fraction of a population selected on some basis, Sample consists of a few items of a population, in principle a sample should be such that it is a true representative of the population. Usually a random sample is selected. If the population is reasonably homogenous a simple random sampling is most preferred one. But the moment one start identifying sampling using on the basis of their characteristic; it gives rise to different sampling methods.

The foremost purpose of sampling is to gather maximum information about the population under consideration at minimum cost, time and human power. This is best achieved when the sample contains all the properties of the population.

In many statistical studies, data are to be collected from the population under study. The population consists of a number of units, usually very large, and sometime infinitely many. In many cases it is practically not possible to include all unity of the populations to study. Therefore a few of the population units are selected as a representative of the whole population, the few selected unit is called sample and the method of selecting the sample is called the sampling technique.

Parameter

The statistical constants of whole population are called parameter such as population mean (μ), population variance (σ^2), population proportion (P). Population correlation coefficient (ρ), population regression coefficient (β_{YX}) etc. out of various parameters mean and variance are largely used. In a theoretical probability distribution by parameters one means those population constant which appear in probability density (mass) function.

12.3 STATISTICS

According to Ronald A. Fisher the statistical constants of the sample selected from the population are called statistics; in general statistics are the function of observable random variable and doesn't involve any unknown parameters. For example sample mean $\bar{x}$, sample variance (S^2 or s^2), sample correlation coefficient (r) etc.

Name	Parameter	Statistics
Size	N	n
Mean	μ	$\bar{x}$
Proportion	P	p
Variance	σ^2	S^2 or s^2
Correlation coefficient	ρ	r
Regression coefficient	β_{YX}	b_{yx}

In practice parameters values are not known and their estimate based on sample values are made, Thus statistics can be regarded as the estimate of population parameter.

12.4 ESTIMATOR

An estimator is a rule or a function of variates for estimating the population parameters. It is expressed as a function of sample variates. An estimator is itself a random variable and can have any value within its domain .For example the estimator for mean is $\Sigma x/n$, which depends on the sample value of x of a variable X, Similarly, the estimator of variance is $\frac{\Sigma(x - \bar{x})^2}{(n - 1)}$

A particular value of an estimator from a fixed set of values of a random sample is known as estimate , An estimate stands for the value of a parameter, For example sample mean $\bar{x}$ is an estimate of population mean μ and sample variance s^2 is an estimate of population variance σ^2 where as the functions for $\bar{x}$ & σ^2 are estimators.

12.5 SAMPLING DISTRIBUTION

If we draw a sample of size n from a given finite population of size N then the total number of possible sample is ${}^{N}C_n = \frac{N!}{(N - n)!n!} = k$ (say)

For each oF these k samples we can compute some statistics t_n a function of n observations, in particular the mean $\bar{x}$, the variance S^2etc are given below

Sample no	t_n	Statistics	
		$\bar{x}$	S^2
1	t_1	$\bar{x}_1$	S_1^2
2	t_2	$\bar{x}_2$	S_2^2
...	...	...	...
k	t_k	$\bar{x}_k$	S_k^2

The set of the values of the statistics so obtained, one for each sample, constitute what is called the sampling distribution of the statistic. For example the values $t_1, t_2 \ldots\ldots, t_k$ determine the sampling distribution of the statistics t_n. in other words , statistics t_n may be regarded as a random variable which can takes the various statistical constants like mean, variance , skewness, kurtosis, etc for its distribution. For example the mean and variance of the sampling distribution of the statistics t_n are given as

$$\bar{t}_n = \frac{t_1 + t_2 + \ldots. + t_k}{k} = \frac{\Sigma t_i}{k} \text{ and}$$

$$\text{Var}(t_n) = \frac{\Sigma(t_n - \bar{t}_n)^2}{k}$$

To bring the sample mean ($\bar{x}$) close to the population mean μ , a number of random samples are taken from the population and the mean of each sample is computed which give the number of several sample means. The probability distribution of all possible sample mean of a given size selected from a population is known as the sampling distribution of the mean. Similarly, the probability distribution of the sample variance and the probability distribution of the sample proportion of all possible random sample of a fixed size n is called the sampling distribution of the proportion.

12.5.1 Properties of the Sampling Distribution of the Mean

i. The mean of the sampling distribution and the mean of the population are equal, if all possible samples of a given size are taken from the population.

ii. The spread of the sample means in the sampling distribution is small than the spread in the population values.

iii. The shape of the sampling distribution of the sample means tends to be bell-shaped and approximates the normal distribution even when the population is not normally distributed; provide that the sample size is reasonably large.

12.6 STANDARD ERROR (SE)

The standard deviation of the sampling distribution of sample statistics is known as standard error (S.E.) of the statistics. The standard deviation of distribution of sample mean is called the standard error of the sample mean. It is denoted by $\sigma_{\bar{x}}$ or S.E.($\bar{x}$). Standard error of the mean is a measure of the distribution of sample mean, it measures the likely deviation of sample mean from the grand mean of the sampling distribution.

12.6.1 Standard Error of the Mean

If the random sample is drawn with replacement from large population, i.e. $x \sim N(\mu,\sigma^2)$ then we have $E(x) = \mu$ and variance$(X) = \sigma^2$. Now the mean of sample mean is defined as

$$E(\bar{x}) = E\left(\frac{\Sigma x_i}{n}\right) = \frac{1}{n} E(x_1 + x_2 + \ldots + x_n)$$

$$= \frac{1}{n} [E(x_1) + E(x_2) + \ldots + E(x_n)]$$

$$= \frac{1}{n}(\mu + \mu + \mu \ldots + \mu)\ \mu = \mu$$

Thus $E(\bar{x}) = \mu$

Now the variance of sample mean is

$$\text{Var}(\bar{x}) = \text{Var}\left(\frac{\Sigma x_i}{n}\right) = \frac{1}{n^2} \text{var}(x_1 + x_2 + \ldots + x_n)$$

$$= \frac{1}{n^2} [\text{var}(x_1) + \text{var}(x_2) + \ldots + \text{var}(x_n)] \quad [\text{Since } x_i\text{'s are iid}]$$

$$= \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Thus, The standard error of sample mean for large population is given by $SE(\bar{x}) = \frac{\sigma}{\sqrt{n}}$, where σ is population standard deviation.

If the random sample is drawn without replacement from a finite population size N, the standard error of sample mean for finite population is given by $S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$, where n is the sample size.

The factor $\sqrt{\frac{N-n}{N-1}}$ is also known as finite correction factor or finite population multiplier, the fraction $\frac{n}{N}$ is called the sampling fraction. When the sampling fraction is less than 0.05 the finite population multiplier need not be used.

12.6.2 Standard Error of the Proportion

The selection of the sampling distribution of proportion is known as the standard error of the proportion. It is denoted by σ_p or S.E.(p) if the random sample is drawn without replacement from a finite population size N then the standard error of the sample proportion for finite population is given by

$$S.E.(p) = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$$

Where P-is the population proportion of success and the Q = 1 – P and the standard error of the sample proportion for large population is given by

$$S.E.(p) = \sqrt{\frac{PQ}{n}}$$

Similarly the S.E. of the various statistics is given below.

i. For random sample drawn from a normal population with s.d. σ The standard error of sample S.D. (s) is given as $S.E.(s) = \sqrt{\frac{\sigma^2}{2n}}$.

ii. If $\bar{x}_1$ and $\bar{x}_2$ denote the means calculated from independent random sample of sizes n_1 and n_2 and drawn from two normal population with S.D. σ_1 and σ_2 respectively the S E of $(\bar{x}_1 - \bar{x}_2)$ the S.E. of difference between two mean is

$S\ E(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, in particular ,if the population have the same S.D = σ i.e. $\sigma_1 = \sigma_2 = \sigma$ then $S\ E(\bar{x}_1 - \bar{x}_2) = \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

iii. If p_1 and p_2 denote the sample proportion calculated from independent random samples of sizes n_1 and n_2 drawn from two independent population with proportion P_1 and P_2 respectively then the S.E. of difference between proportions is given by

$S.E.(p_1 - p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}}$, in particular if it is assumed that the two population proportion P_1 and P_2 are equal say $P_1 = P_2 = P$ then

$S.E.(p_1 - p_2) = \sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

iv. If s_1 and $s_2$denote the sample standard deviations calculated from independent random sample of sizes n_1 and n_2 and drawn from two population with S.D σ_1 and σ_2 respectively then

$$S.E(s_1 - s_2) = \sqrt{\frac{\sigma_1^2}{2n_1} + \frac{\sigma_2^2}{2n_2}}$$

The standard error indicated not only to observe the variability in the sample means but also the accuracy of the estimation population parameters. A distribution of sample means having larger S.E. is a better estimator of the population mean than a distribution of sample means having smaller standard error. Following table shows the different statistics with corresponding Standard errors

S.No.	Statistics	Standard error
1.	Sample mean $\bar{x}$	$\frac{\sigma}{\sqrt{n}}$
2.	Sample Proportion (p)	$\sqrt{\frac{PQ}{n}}$
3.	Sample s.d (s)	$\sqrt{\frac{\sigma^2}{2n}}$
4.	Sample Quartiles	$1.36263\frac{\sigma}{\sqrt{n}}$
5.	Sample median	$1.25331\frac{\sigma}{\sqrt{n}}$
6.	Sample correlation (r)	$1-\frac{\rho^2}{\sqrt{n}}$
7.	Sample Moments μ_3	$\sigma^3\sqrt{\frac{96}{n}}$
8.	Sample Moments μ_4	$\sigma^4\sqrt{\frac{96}{n}}$
9.	Sample coefficient of variation	$\frac{CV}{\sqrt{2n}}\sqrt{1+\frac{2(CV)^3}{104}}\approx\frac{CV}{\sqrt{2n}}$
10.	$\bar{x}_1-\bar{x}_2$	$\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$
11.	S_1-S_2	$\sqrt{\frac{\sigma_1^2}{2n_1}+\frac{\sigma_2^2}{2n_2}}$
12.	p_1-p_2	$\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}$

12.6.3 Utility of Standard Error

Standard Error plays vital role in the large sample theory and forms the basis of the estimation and the testing the hypothesis it is extensively used.

i. To determine the precision or reliability of the sample estimate of some population parameter. Thus if $\bar{x}$ is a sample mean and to estimate the population mean μ, the precision of $\bar{x}$ is $\frac{1}{S.E(\bar{x})}$

ii. To test if the sample statistic differs significantly from the corresponding hypothetical value in the population i.e. to test the significant of the difference $(\bar{x}-\mu)$. For large sample, the Z-statistic is as $Z=\frac{(\bar{x}-\mu)}{S.E.(\bar{x})}$

If calculated Z is less than or equal to tabulated Z , it is not significant at αlevel of significance and null hypothesis is accepted . This means that there is no significance difference between observed and expected value of statistics. If calculated Z is greater than tabulated Z, it is significant at α level of significant and null hypothesis is rejected . This means that there is significant difference between observed sample value and the hypothesized population value.

iii. To test the significance of the difference between two independent sample estimates of the same population parameter.

iv. To obtain the point estimates of the population parameters.

v. To obtained interval estimates of the population parameter i.e. to obtain probable limits between which the true value of the parameter may be expected to lie.

Example 1

There are 60 students in the B.E. class .Their average score in their semester end exam were 70 with a standard deviation of 10. A sample of 40 of these students is taken at random calculate the Standard error of the mean for this sample.

Solution

Here we are given, Population size (N) = 60, Sample size (n) = 40, Population standard deviation (σ) = 10 then the standard error of mean =? Here sampling fraction n/N = 0.6 is greater than 0.05 , the standard error of mean is given as

$$S.E(\bar{x})=\sigma_{\bar{x}}=\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}=\frac{10}{\sqrt{40}}\sqrt{\frac{60-40}{60-1}}=0.8512$$

Example 2

Simple random sample of size 20 is drawn without replacement from a finite population consisting of 60 units. If the number of defective units in the population be 8.Find the standard error of the sample proportion defectives.

Solution

Here we are given Population size (N) = 60, Sample size (n) = 20, Number of defective in the population(X) = 8. Then the Population proportion of defective unit (P) = X/N = 8/60 =0.134 and Q = 1- 0.134 = 0.866. The required Standard error of sample proportion of defective is

$$S.E\ (p) = \sigma_p = \sqrt{\frac{PQ}{n}}\sqrt{\frac{N-n}{N-1}}$$

$$= \sqrt{\frac{0.134 \times 0.866}{20}}\sqrt{\frac{60-20}{60-1}} = 0.063$$

Note: When the sampling fraction $\frac{n}{N}$ is less than 0.05, the finite population multiplier $\sqrt{\frac{N-n}{N-1}}$ need not be used.

12.7 SUM AND AVERAGE OF RANDOM VARIABLE

12.7.1 Random Variable

A rule that assigns a real number to each outcome of a random experiment is called a random variable (random variable.) . It can take any one of the various possible values each with definite probability. For example, in a throw of a dice if X denotes the number obtained then X is a random variable which can take any one of the values 1, 2, 3, 4, 5, 6, each with equal probability 1/6.

A random variable is usually denoted by any of the Capital Latin letters X, Y, Z, U, V... etc and particular values which the random variable takes are denoted by the corresponding small letters. Let us consider a random experiment of three tosses of a coin then the sample space S consist of $2^3 = 8$ sample points given as

S ={HHH, HHT,HTH,THH, HTT,THT, TTH, TTT}.

Let us consider the variable X, which is the number of heads obtained, then X is a random variable which can take any one of the values 0, 1, 2, 3.

Outcomes	HHH	HHT	HTH	THH	HTT	THT	TTH	TTT
Values of X	3	2	2	2	1	1	1	0

12.7.2 Types of Random Variables

There are two types of random variable namely. (a) Discrete random variable, (b) Continuous random variable,

12.7.3 Discrete random variable

A random variable, say X, which can take a finite or a countably infinite numbers of values in an interval of the real line is called discrete random

variable. For example, if we toss a coin, the variable can take only two values 0 and 1 assigned to tail and head respectively.

$$X = \begin{cases} 0, \text{ if X is Tail} \\ 1, \text{ If X is Head} \end{cases}$$

In a rolling of a die, only six values of the variable x, i.e. 1,2,3,4,5&6 are possible, hence the variable X is discrete , here the variable

$X = \{ x: x \text{ is } 1, 2, 3, 4,5 \text{ and } 6\}$

12.7.4 Continuous Random Variable

A random variable X, which can take infinite and uncountable set of values in an interval of real line is said to be a continuous random variable , the probability of a point x is zero i.e. P(X=x) = 0. But the probability is ascribable in an interval. for example the weight of middle aged people in Kathmandu lying between 40 kg and 150 kg is a continuous random variable . Symbolically, $X = \{ x: 40 \leq x \leq 150\}$

12.7.5 Properties of Random Variable

The important properties of random variable are as follows,

i. If X is a random variable and a and b are any two constant, then aX+b is also a random variable.

ii. If X is a random variable, then X^2 is also a random variable.

iii. If X is a random variable then 1/ X is also a random variable.

iv. If X and Y are random variables defined on the same sample space S then, X+Y, X-Y, aX, bY or aX+bY are also random variables with a and b are non negative constants.

v. If $X_1, X_2 \ldots X_n$ are n independent random variables then,
$U_n = \min(X_1, X_2 \ldots X_n)$ and
$V_n = \max(X_1, X_2 \ldots X_n)$ are also random variables.

12.7.6 Sum of Random Variables

Let $X_1, X_2 \ldots X_n$ are n random variables then their sum as

$X_1 + X_2 + \ldots + X_n = \sum_{i=1}^{n} x_i$ is also a random variable.

In normal distribution let, $X_1, X_2 \ldots X_n$ are n independently normally distributed random variables with mean μ_1 and variance σ_i^2 then their sum Σx_i is also a normally distributed random variable with mean $n\mu$ and variance $n\sigma^2$ for large value of sample size n.

12.7.7 Average of Random Variable

Let $X_1, X_2 \ldots\ldots X_n$ are n random variables then their simple average $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is also a random variable.

In normal distribution let $X_1, X_2 \ldots\ldots X_n$ are n independently normally distributed random variables with mean μ_i and variance σ_i^2 then their simple average $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ is also a normally distributed random variable with mean μ and variance $\frac{\sigma^2}{n}$ for large n. i.e. $E(X) = \mu$ and $Var(X) = \frac{\sigma^2}{n}$

12.8 CENTRAL LIMIT THEOREM

The central limit theorem in the mathematical theory of probability may be expressed as follows "If $X_1, X_2 \ldots\ldots X_n$ are n independent random variables following any distribution with mean $E(X_i) = \mu_i$ and variance $Var(X_i) = \sigma_i^2$, than under certain very general conditions, the random variable i.e. their sum $S_n = X_1 + X_2 + \ldots\ldots + X_n = \sum x_i$ is asymptotically normal distributed with mean μ and standard deviation σ Where $\mu = \Sigma\mu_i$ and $\sigma^2 = \Sigma\sigma_i^2$ i.e. $S_n \approx N(\mu, \sigma^2)$ as $n \to \infty$"

This theorem was first stated by Laplace in1812 and a rigorous proof under fairly general conditions was given by Liapounoff in 1901. .The general conditions or assumptions are as

i. The variables are independent and identically distributed.
ii. $E(X_i^2)$, exists for all i = 1, 2 .....

12.8.1 Application of Central Limit Theorem

If $X_1, X_2 \ldots\ldots X_n$ are n identically independently binomial distributed random variables with parameter r and p. then their sum

$S_n = X_1 + X_2 + \ldots\ldots + X_n = \sum_{i=1}^{n} x_i$ is normal variate with mean $E(S_n) = \Sigma E(Xi) = \Sigma rp = nrp$. And $Var(S_n) = V(\Sigma Xi) = (\Sigma V(X_i) = nrpq$

Example 10

Let $X_1, X_2 \ldots\ldots X_n$ be a identically independently Poisson variate with parameter λ Use the central limit theorem to estimate $P(12 < S_n < 160)$ where $S_n = X_1 + X_2 + \ldots\ldots + X_n$ and $\lambda = 2$ and $n = 75$.

Solution

Since X_i's are iid Poisson variates with parameter λ then $E(X) = \lambda$ and $V(X) = \lambda$ and hence $E(S_n) = \Sigma E(X_i) = n\lambda$ and $V(S_n) = var(X_1 + X_2 + \ldots\ldots + X_n) = \Sigma V(X_i) = n\lambda$ [Since X_i's are iid] Hence by Central limit theorem for large size of n

$S_n \approx N(n\lambda, n\lambda)$

$S_n \approx N(\mu = 150, \sigma^2 = 150)$

$[\because \lambda = 2 \ \& \ \nu = 50]$

$$\therefore P(120 < S_n < 160) = P\left(\frac{120 - 150}{\sqrt{150}} < Z < \frac{160 - 150}{\sqrt{150}}\right)$$

$= P(-2.45 < Z < 0.82)$

$Z \approx N(0,1)$

$= P(-2.45 < Z < 0) + P(0 < Z < 0.82)$

$= 0.4929 + 0.2939$

$= 0.7868$ from standard normal table.

Thus 78.68% values lies between 120 and 160 of sum of Poisson variate

12.9 ESTIMATION

The theory of estimation was founded by Prof. R.A. Fisher. Estimation is the process by which population characteristics are estimated from characteristics of the sample studies with desired degree of precision. The main objective of estimation is to obtain a guess or estimate of the unknown, true value from the sample data or past experience. There are many situations in our daily life where we make estimation. When crossing a road, we estimate the speed of taxi that is approaching, the distance between us and that taxi and our own speed. Having made these quick estimates, we decide whether to wait, walk or run. We must make quick estimates to save the accidents, some times, a decision maker make rational decisions without complete information and with a great deal of uncertainty about what the future will bring.

In estimation, the sample statistics are used to estimate the population parameters. For example a principal makes the estimates of enrollments next year in B.E. from current enrollments in the same college. The civil engineer estimates the strength of soil, Strength of rod, strength of brick from the past records. A businessman estimates his future sales of computer and electronic goods from the past records, in each case; some one is trying to draw conclusion about population from sample information.

The estimation of population parameter like mean, variance, proportion etc from the corresponding sample statistics with reasonable accuracy is very common problem of statistical inference. The sample distributions of a

statistics and its standard error have played a vital role in both the estimation of parameters and the testing of statistical hypothesis.

12.9.1 Types of Estimates

A random sample of a given size is selected from a given population and then computes a statistic which is a characteristic of the sample and this becomes an estimate of the similar characteristic of the population. There are two types of estimates about a population parameter namely a point estimate and an interval estimate.

12.9.2 Point Estimation

A point estimate of a parameter is a single numerical value compute from sample information. For example a sample mean ($\bar{x}$) and a sample variance s^2 is the point estimate of population mean μ and population variance σ^2.

We write as point estimator of μ is $\bar{x}$ as $\hat{\mu} = \bar{x}$ and that of σ^2 as $\hat{\sigma}^2 = s^2$

12.9.3 Interval Estimation

A range of values used to estimate a population parameter is called an interval estimate. There are two ways for indicating error by the extent of its range and by the probability of the true population parameter lying within that range. For example

$P(a < \mu < b) = 0.95$.

Example of interval estimation are given in the chapter of hypothesis testing and exact sampling test.

12.9.4 Properties of Good Estimator

A good estimator is one which is as close to the true value of the parameter as possible. The following are the some of the criteria that should be satisfied by a good estimator.

a. Unbiasedness
b. Consistency
c. Sufficiency
d. Efficiency

12.10 UNBIASEDNESS

An estimator or statistics t_n which is a function of the sample observations $x_1, x_2, \ldots x_n$ is said to be an unbiased estimator of the corresponding population parameter θ if $E(t_n) = \theta$ ... (i)

For example, sample mean ($\bar{x}$), sample median (M_d) and sample Mode (M_o) are unbiased estimator of population mean

i.e. $E(\bar{x}) = E(M_d) = E(M_o) = \mu$.

Similarly, sample proportion p is unbiased estimator of population proportion (P) i.e. $E(p) = P$

But $E(S^2) \neq \sigma^2$ where as $E(s^2) = \sigma^2$

Where $S^2 = \frac{\Sigma(x - \bar{x})^2}{n}$ and $s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$

If $E(t_n) \neq \theta$, then the statistics (t_n) is said to be biased estimator of θ.

Let $E(t_n) - \theta = B(t_n)$ then $B(t_n)$ is called the amount of bias. If $B(t_n) > 0$

i.e $E(t_n) > \theta$. Then t_n is said to be positive biased and

If $B(t_n) < 0$ i.e $E(t_n) < \theta$. it is said to be negative biased.

Therefore, Amount of Biased is $B(t_n) = E(t_n) - \theta$

Example 3

Let $x_1, x_2, \ldots\ldots\ldots x_n$ is a random sample from a normal population $N(\mu, 1)$ show that $t_n = \frac{1}{n}\Sigma x_i^2$ is an unbiased estimate of $(1+\mu^2)$

Solution

We are given $E(x_i) = \mu$ and $V(x_i) = 1$ for ,i= 1, 2,.....n

Now $E(x_i^2) = V(x_i) + [E(x_i)]^2 = (1+\mu^2)$

$\therefore E(t_n) = \frac{1}{n}\Sigma(x_i^2) = \frac{1}{n}\Sigma(1 + \mu^2) = 1 + \mu^2$,

Hence $t_n = \frac{1}{n}\Sigma x_i^2$ is an unbiased estimate of $(1+\mu^2)$

Example 4

If T is an unbiased estimator of θ, show that T^2 is a biased estimator of σ^2

Solution

Since T is an unbiased estimator for θ, we have $E(T) = \theta$

Also $Var(T) = E(T^2) - \{E(T)\}^2 = E(T^2) - \theta^2$

Which implies $E(T^2) = \theta^2 + Var(T)$ [since $Var(T) > 0$]

Since $E(T^2) \neq \theta^2$, T^2 is a biased estimator of θ^2

Example 5

Show that $\frac{\Sigma x_i (\Sigma x_i - 1)}{n(n-1)}$ is an unbiased estimate of θ^2, for the sample $x_1, x_2 \ldots\ldots\ldots\ldots x_n$ drawn on which X take the value 0 or 1 with respectively probability $(1-\theta)$ and θ.

Solution

Since $x_1, x_2 \ldots\ldots x_n$ is a random sample from Bernoulli population with parameter θ, Now $T = \Sigma x_i$ is a binomial distribution with parameter n and θ

$\Rightarrow E(T) = n\theta$ and $V(T) = n\theta(1-\theta)$

$$\text{Therefore } E\left(\frac{\Sigma x_i (\Sigma x_i - 1)}{n(n-1)}\right) = E\left(\frac{T(T-1)}{n(n-1)}\right) = \frac{[E(T^2) - E(T)]}{n(n-1)}$$

$$= \frac{[\{var(T) + (E(T))^2\} - E(T)]}{n(n-1)}$$

$$= \frac{[\{n\theta(1-\theta) + n^2\theta^2\} - n\theta]}{n(n-1)} = \frac{n\theta^2(n-1)}{n(n-1)} = \theta^2$$

Which shows that $\frac{\Sigma x_i (\Sigma x1 - 1)}{n(N-1)}$ is an unbiased estimator of θ^2.

12.11 CONSISTENCY

A statistics t_n based on sample of size n is said to be a consistence estimator of the parameter θ if it converges in probability to θ

i.e. if $t_n \rightarrow \theta$ as $n \rightarrow \infty$,

Symbolically, $\lim_{n\to\infty} P(t_n \rightarrow \theta) = 1$. In other word statistics t_n is said to be consistent estimator of the population parameter θ if as the sample size increase the value of the sample statistics become very close to the value of the population parameter.

For example a sample mean ($\bar{x}$), a sample proportion (p) and sample variance (s^2) comes closes to the respective consistence estimator of the population mean μ, population proportion (P) and population variance (σ^2) if

$\lim_{n\to\infty} E(\bar{x}) = \mu$, $\lim_{n\to\infty} E(p) = P$ and $\lim_{n\to\infty} E(s^2) = \sigma^2$

For calculation purpose of consistency of an estimator we use the following relation.

A statistics t_n is a consistence estimator of the parameter θ

If $E(t) = \theta$ and $V(t) = 0$ as $n \rightarrow \infty$

Example 6

Prove that in sampling from an $N(\mu, \sigma^2)$ population, the sample mean $\bar{x}$ is a consistent estimator of μ.

Solution

In a sampling form a $N(\mu, \sigma^2)$ population the sample mean $\bar{x}$ is also normally distributed as $N(\mu, \frac{\sigma^2}{n})$ i.e. $E(\bar{x}) = \mu$ and $Var(\bar{x}) = \frac{\sigma^2}{n}$.

Thus as $\lim_{n\to\infty} E(\bar{x}) = \mu$ and $Var(\bar{x}) = 0$. Hence $\bar{x}$ is a consistent estimator for population mena μ,

Also sample median is normally distributed with mean μ and variance $\frac{\pi\sigma^2}{2n}$ i.e. $E(M_d) = \mu$ and $Var(M_d) = \frac{\pi\sigma^2}{2n}$.

Thus as $\lim_{n\to\infty} E(M_d) = \mu$ and $Var(M_d) = 0$. Hence M_d is also an unbiased and consistent estimator for μ.

12.12 EFFICIENCY

Even if we confine ourselves to unbiased estimates there will in general exist more than one consistent estimator of a single parameter. Thus there is a necessity of some further criteria which will enables us to choose between the estimators with the common property of consistency such a criterion which is based on the variance of the sampling distribution of estimators usually known as efficiency.

An unbiased estimator t_n based on sample of n observation is said to be efficient estimator than any other unbiased estimator t_n^* of population parameter θ if and only if $Var(t_n) < Var(t_n^*)$.

Also the relative efficiency of t_n as compared to t_n^* is given as

$$R.E. = \frac{Var(t_n)}{Var(t_n^*)}$$

Example 7

Sample mean ($\bar{x}$) and sample median (M_d) are unbiased and consistent estimator of population mean μ i.e. $E(\bar{x}) = \mu$ and $E(M_d) = \mu$ but we also have $Var(\bar{x}) = \frac{\sigma^2}{n}$ and

Var (Md) = $\frac{\pi\sigma^2}{2n}$ Here Var ($\bar{x}$) < Var (M_d).

Therefore $\bar{x}$ is more efficient than M_d.

Also the relative efficiency of M_d is R.E (M_d.) = $\frac{\text{Var}(\bar{x})}{\text{Var}(M_d)} = \frac{\sigma^2/n}{\pi\sigma 2/2n}$ = 0.637

Therefore, the relative efficiency of M_d is 63.7% as compared to mean $\bar{x}$.

Example 8

A random sample (x_1, x_2, x_3, x_4, and x_5) of size 5 is drawn from a normal population with unknown mean μ. Consider the estimator to estimate μ.

i. $t_1 = \frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$, ii. $t_2 = \frac{x_1 + x_2}{2} + x_3$, iii. $t_3 = \frac{2x_1 + x_2 + \lambda x_3}{3}$

Where λ is such that t_3 is an unbiased estimator of μ. Find λ, Are t^1 and t_2 unbiased. State giving reasons, which estimator is best among t_1, t_2 and t_3

Solution

Since $X \approx N(\mu, \sigma^2)$ then $E(x_i) = \mu$ and $V(x_i) = \sigma^2$, i=1, 2,....5
and $Cov(x_i, x_j) = 0$ $i \neq j$

i. Now $E(t_1) = \frac{1}{5}\Sigma E(x_i) = \mu$, which implies t_1 is an unbiased estimator of μ

ii. $E(t_2) = \frac{1}{2}E(x_1 + x_2) + E(x_3) = 2\mu$, which implies t_2 is not an unbiased estimator of μ.

iii. Given that $E(t_3) = \mu, \Rightarrow \frac{E(2x_1 + x_2 + \lambda x_3)}{3} = \mu$ [$\because$ t_3 is unbiased estimator of μ]

or, $2\mu + \mu + \lambda\mu = 3\mu \Rightarrow \lambda = 0$

iv. Also, $var(t_1) = \frac{1}{25}(V(x_1) + V(x_2) + V(x_3) + V(x_4) + V(x_5)$

$= \frac{1}{5}\sigma^2$

$var(t_2) = \frac{(V(x_1) + V(x_2))}{4} + V(x_3) = \frac{1}{2}\sigma^2 + \sigma^2 = \frac{3}{2}\sigma^2$

v. And $var(t_3) = \frac{(4V(x_1) + V(x_2))}{9} = \frac{5}{9}\sigma^2$

Since var (t_1) is least therefore the estimator t_1 is the best estimator.

12. 13 SUFFICIENCY

A statistics t_n based on sample of n observation is said to be sufficient for a parameter θ if it contains all the information in the sample regarding the parameter. For example the mean use all the sample values in its computation while median uses only information of two class intervals and the mode uses only information of there class intervals. Thus the sample mean $\bar{x}$ is sufficient estimator of the population mean μ.

Also, A statistics t_n based on sample of n observation is sufficient for the parameter θ of a density function $f(x, \theta)$ if the conditional distribution of x_1, x_2..... x_n given value of t_n=k is independent of parameter. θ

i.e. $P(x_1 \cap x_2 \cap \ldots\ldots \cap x_n / t_n = k) = h(x)$, where h(x) is function of x, not contain θ

12. 15 METHOD OF ESTIMATION

So far we have been discussing the requisites of a good estimator. Now we shall briefly outline some of the important methods for obtaining such estimators. Some of the most commonly used method for obtaining the point estimates are listed below

i. Method of maximum likelihood
ii. Method of moments
iii. Method of least square
iv. Method of minimum variance
v. Method of minimum chi-square
vii. Method of inverse probability
viii. Bayes' estimation procedures

12.15 .1 Method of Maximum Likelihood

Method of maximum likelihood estimator (MLE) was initially formulated by C.F. Gauss but as a general method of estimation was first introduced by Prof. R. A. Fisher and later on developed by him in a series of papers.
Let x_1, x_2..... x_n be a random sample of size n drawn from a population with pdf $f(x, \theta)$ where θ is the unknown parameter. The principle of maximum likelihood consists in finding an estimator for the unknown parameter θ which maximizes the likelihood function 'L' of x_1, x_2..... x_n defined as

$$L = f(x_1, \theta)\, f(x_2, \theta) \ldots\ldots\ldots f(x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$$

Here we maximize L to find an estimator of θ. Thus for Maxima and minima,

$$\frac{\partial L}{\partial \theta} = 0 \text{ and } \frac{\partial^2 L}{\partial \theta^2} < 0$$

$$\text{Or } \frac{1}{L}\frac{\partial L}{\partial \theta} = 0 \text{ or } \frac{\partial \log L}{\partial \theta} = 0$$

Here Log L is a non-decreasing function of L and also L and Log L attains their maximum values at the same value therefore

$\frac{\partial \log L}{\partial \theta} = 0$ is known as the likelihood equations for estimation of parameter θ.

Example 9

If $x_1, x_2 \ldots\ldots x_n$ is a random sample of size n drawn from a Poisson population P (λ) where λ is unknown parameter. Find the MLE of λ.

Solution

Since each $x_i \approx P(\lambda)$ then $f(x_i, \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$, then likelihood function of $x_1, x_2 \ldots\ldots x_n$ is

$$L = f(x_1, \lambda)\, f(x_2, \lambda) \ldots f(x_n, \lambda) = \frac{e^{-\lambda}\lambda^{x_1}}{x_1!}\frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \ldots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$

$$L = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{\Pi x_i!}, \text{ taking log on both sides}$$

$$\text{Log } L = -n\lambda + \Sigma x_i \log \lambda\, x_1 + \Pi\,(x_1!)$$

And the likelihood equation to estimate the value of parameter λ is

$$\frac{\partial \log L}{\partial \lambda} = 0 \text{ or } -n + \frac{\Sigma x_i}{\lambda} = 0$$

$$\therefore \lambda = \bar{x}$$

Therefore, Maximum likelihood estimator of parameter λ is $\bar{x}$

Example 10

A random sample of size n is drawn from a normal population $N(\mu, \sigma^2)$. Estimate the parameter μ and σ^2 by the method of MLE.

Solution

Let $x_1, x_2 \ldots\ldots x_n$ is a random sample of size n drawn from a normal population $N(\mu, \sigma^2)$ where μ and σ^2 are unknown parameter then

$$f(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_i - \mu}{\sigma}\right)^2}, \quad x = o, 1 \ldots$$

Then the likelihood function of samples $x_1, x_2 \ldots\ldots x_n$ is

$$L = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_1 - \mu}{\sigma}\right)^2} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_2 - \mu}{\sigma}\right)^2} \ldots \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x_n - \mu}{\sigma}\right)^2}$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\Sigma\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

$$= \left(\frac{1}{\sigma^2(2\pi)}\right)^{n/2} e^{-\frac{1}{2}\Sigma\left(\frac{x_i - \mu}{\sigma}\right)^2}$$

Taking log on both sides, we get

$$\text{Log } L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{\sigma^2}\Sigma(x_i - \mu)^2$$

The likelihood equation to estimation of parameter μ is

$$\frac{\partial \text{ Log } L}{\partial \mu} = 0$$

$$\frac{1}{\sigma^2}\Sigma(x_i - \mu) = 0$$

$$\therefore \mu = \bar{x}$$

Therefore Maximum likelihood estimator of parameter μ is $\bar{x}$

Similarly, likelihood equation to estimate the value of parameter σ^2 is

$$\frac{\partial \text{ Log } L}{\partial \sigma^2} = 0$$

Or,
$$-\frac{n}{2\sigma^2}\left(\sigma^2 - \frac{\Sigma(x_i - \bar{x})^2}{n}\right) = 0$$

$$\therefore \sigma^2 = \frac{\Sigma(x_i - \bar{x})^2}{n} = s^2$$

Therefore Maximum likelihood estimator of σ^2 is s^2

Example 11

If n sample are drawn from Binomial population with parameter r and P. find the Maximum likelihood estimator of P

Solution

Let $x_1, x_2 \ldots\ldots x_n$ is a random sample of size n drawn from a B(r, P) then probability mass function of binomial distribution is

$$f(x, P) = {}^rC_x P^x Q^{r-x}$$

Then the likelihood function of samples $x_1, x_2 \ldots\ldots x_n$ is

$$L = f(x_1, P)\, f(x_2, P) \ldots f(x_n, P)$$

$$= {}^rC_{x_1} P^{x_1} Q^{r-x_1}\, {}^rC_{x_2} P^{x_2} Q^{r-x_2} \ldots {}^rC_{x_n} P^{x_n} Q^{r-x_n}$$

$$= \Pi({}^rC_{xi})\, P^{\Sigma xi}\, Q^{\Sigma(r-xi)}$$

Taking log on both sides,

$$\text{Log } L = \Sigma \log {}^{r}C_{x_i} + \Sigma x_i \log P + \Sigma (r - x_i) \log (1 - P)$$

The likelihood equation to estimate the parameter P is

$$\frac{\partial \text{Log } L}{\partial P} = 0$$

or, $$0 + \frac{\Sigma x_i}{P} - \frac{nr - \Sigma x_i}{1 - P} = 0$$

or, $$\frac{nr - \Sigma x_i}{1 - P} = \frac{\Sigma x_i}{P}$$

or, $$nrP - P \Sigma x_i = \Sigma x_i - P \Sigma x_i$$

$$\therefore P = \frac{\bar{x}}{r}$$

Therefore Maximum likelihood estimator of P is $\frac{\bar{x}}{r}$

Properties of MLE

Following are the important properties of MLE

i. A MLE is not necessarily unique.

ii. A MLE is not necessarily unbiased.

iii. A MLE may not be consistent in rare cases.

iv. A MLE may not be uniformly minimum variance unbiased estimator.

v. If a sufficient estimator exists, it is a function of the maximum likelihood estimator.

vi. If t_n is a sufficient statistics for parameter θ and a unique then MLE of θ exist and function of t_n and vice versa.

vii. If t_n is a MLE of θ and $\tau(\theta)$ is a one to one function of θ. $\tau(t_n)$ is a MLE $\tau(\theta)$ of .This property is known as invariance property of MLE.

viii. Under very general conditions, Maximum likelihood estimators are consistent. The general conditions are

a. The first order and second order derivative viz $\frac{\partial \text{Log} L}{\partial \theta}$ and $\frac{\partial^2 \text{Log} L}{\partial \theta^2}$ exists and are continuous function of θ in a range of R.

b. The third order derivative exists such that $\left|\frac{\partial^3 \text{Log} L}{\partial \theta^3}\right| < M(x)$, where $E(M(x)) < k$, a positive quantity.

ix. Under certain conditions, it has been observed that MLE is consistent but it is not generally unbiased.

x. When MLE exists, it is most efficient for large samples under the assumption that the distribution of the estimator tends to normal.

12.15.2 Method of Moments

This method was discovered and studied in detail by Karl Pearson in 1894. It is the oldest method of estimating the parameter of a distribution. The method of moments consists of equating the sample moments to the corresponding moments of the distribution, which are the function of unknown parameters .Here we equate as many sample moments as there are unknown parameters.

Solving the simultaneous equations, one obtains the estimates of the moments of the population in terms of sample variates.

Here population moments are $\mu_1' = E(X)$, $\mu_2' = E(X^2)$, $\mu_3' = E(X^3)$ etc and the sample moments are $m(1) = \frac{\Sigma x}{n} = \bar{x}$, $m(2) = \frac{\Sigma x^2}{n}$, $m(3) = \frac{\Sigma x^3}{n}$ etc.

12.15.3 Properties Of The Estimation Obtained By The Method Of Moments

i. Under the fairly general conditions, the estimates obtained by the method of moments will have asymptotically normal distribution for large n.

ii. The mean of the distribution of estimate will differ from the true value of the parameter by a quantity of order $\frac{1}{n}$.

iv. The variance of the distribution of estimate will be of the type $\frac{c^2}{n}$.

v. In general, the estimates obtained by the method of moments are less efficient than the maximum likelihood estimation. In particular cases, they are equivalent.

Example 12

Let $x_1, x_2 \ldots\ldots x_n$ are n samples drawn from a Binomial population with parameter r and P. Estimate the parameter r and P.

Solution

We have the population moments of Binomial distribution are

$\mu_1' = E(X) = r P$ and

$\mu_2' = E(X^2) = r^2P^2 - rP^2 + r P$

Now sample moments are $m(1)= \frac{\Sigma x}{n} = \bar{x}$ and $m(2) = \frac{\Sigma x^2}{n}$.

According to method of moments $\mu_1' = m(1)$

Or, $rP = \bar{x}$ which implies

$$P = \frac{\bar{x}}{r}$$

Also, $\mu_2' = m(2)$ or, $r^2P^2 - rP^2 + rP = \frac{\Sigma x^2}{n}$

Or, $\bar{x}^2 - \frac{\bar{x}^2}{r} + \bar{x} = \frac{\Sigma x^2}{n}$

$$\therefore \quad r = \frac{n\bar{x}^2}{n\bar{x}^2 - \Sigma x^2 + n\bar{x}}$$

Example 13

Find the estimate of the parameter λ of the Poisson distribution by method of moments.

Solution

let $x_1, x_2 \ldots\ldots x_n$ is a random sample of size n drawn from a Poisson distribution $P(x, \lambda)$ we know in case of Poisson distribution its mean and variance are equal. Then theoretical moments is

$$\mu_1' = E(X) = \Sigma x \frac{e^{-\lambda}\lambda^x}{x!} = \lambda \text{ and}$$

Now sample moments is $m(1) = \frac{\Sigma x^2}{n} = \bar{x}$

According to method of moments $\mu_1' = m(1)$

Or, $\lambda = \bar{x}$

Therefore estimate of λ is $\bar{x}$

Example 14

Estimate the parameter μ and σ^2 of the normal distribution by the method of moments

Solution

Let $x_1, x_2 \ldots\ldots x_n$ is a random sample of size n drawn from a normal population $N(\mu, \sigma^2)$

We have the population moments of normal distribution are

$\mu_1' = E(X) = \mu$ and

$\mu_2' = E(X^2) = Var(X) + [E(X)]^2 = \sigma^2 + \mu^2$

Now sample moments are $m(1)= \frac{\Sigma x}{n} = \bar{x}$ and $m(2) = \frac{\Sigma x^2}{n}$

According to method of moments $\mu_1' = m(1)$

Or, $\mu = \bar{x}$

Also, $\mu_2' = m(2)$

or, $\sigma^2 + \mu^2 = \Sigma x^2/n$

Or, $\sigma^2 = \Sigma x^2/n - \bar{x}^2$

12.15.4 Least Square Method

Least square method of estimation of the parameters of simple linear model

Consider the simple linear model $Y = \beta_0 + \beta_1 X + e$. Let there be n pair of observations $(x_1, y_1), (x_2, y_2) \ldots\ldots\ldots (x_n, y_n)$. In the above model Y is called the dependent variable and X, the independent variable and e is the random error. For the i^{th} pair of observation $Y_i = \beta_0 + \beta_i X_i + e_i$

Or $e_i = Y_i - \beta_0 + \beta_1 X_i$ may be positive or negative, to avoid the problem of sign of error, square both side and also take the sum over all pair then we have

$$Q = \Sigma e_i^2 = \Sigma(Y_i - \beta_0 - \beta_1 x_i)^2$$

To get the best estimate of β_0 and β_1 one has to minimize $Q = \Sigma e_i^2$, which amounts to minimize $\Sigma(Y_i - \beta_0 - \beta_1 X_i)^2$.

Thus value of β_0 and β_1 which minimize Q are known as least square estimator. To estimate parameters β_0 and β_1 differential Q partially with respect to β_0 and β_1 respectively and equate to zero

i.e. $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$ which gives

$$\Sigma Y_i = n\beta_0 - \beta_i \Sigma x_i \text{ and}$$

$$\Sigma x_i Y_i = \beta_0 \Sigma X_i - \beta_1 \Sigma X_i^2$$

These two equations are known as normal equation. On solving these two equations we obtain as

$$\beta_0 = \bar{Y} - \beta_1 \bar{X} \quad \text{and} \quad \beta_1 = \frac{n\Sigma X_i Y_i - \Sigma X_i \Sigma Y_i}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

{The best example of least square method is to estimate the parameters 'a' and 'b' of the regression line which is done in correlation and regression chapter}

Note: The Question for interval estimation are Dealt after the hypothesis testing at the end of each of the test.

EXERCISE 12

1 What are the properties of good estimator? Discusses in brief.

2 Define the Central limit theorem, sum of random variable and average of random variable.

3 What are the properties of good estimator? Write about method of maximum likelihood estimator to find the point estimator of a parameter.

4 What do you understand by the method of moments of estimation of the parameter?

5 Find the maximum likelihood estimator of mean (λ) of Poisson distribution.

6 x_1, x_2, x_3 is a random sample of size 3 from a population with mean μ and variance σ^2. Let T_1, T_2 and T_3 are the estimator used to estimate mean value μ where

$T_1 = x_1 + x_2 - x_3$, $T_2 = 2x_1 + 3x_2 - 4x_3$ and $T_3 = \frac{1}{3}(\lambda x_1 + x_2 + x_3)$

i) Are T_1 and T_2 unbiased estimator?

ii) Find the value of λ such that T_3 is unbiased estimator for μ.

iii) Which is the best estimator?

7 For Gamma distribution $f(x, \alpha, \lambda) = \frac{e^{-\alpha x} x^{\lambda - 1}}{\Gamma \lambda}$, find the MLE of α where λ is taken to be a unknown constant

8 Estimate the parameter of the exponential distribution

$f(x, \lambda) = \lambda e^{-\lambda x}$ for $0 \le x < \infty$ by the method of moments.

ANSWERS

6. i. yes, ii. $\lambda = 1$, iii. T_3 **7.** $\hat{\alpha} = \frac{\lambda}{\Gamma \lambda \bar{x}}$, **8.** $\hat{\lambda} = \frac{1}{\bar{x}}$

CHAPTER 13

13. Statistical Inference

13.1 INTRODUCTION

The theory of testing of parametric statistical hypothesis was originally set forth by J. Neyman in 1928 and Karl Pearson in 1933.

There is always some contention about the values of a parameter or the relationship between parameters. When parametric values are unknown, we estimate them through sample values. If the sample value is exactly the same as per our contention, there is no hitch in accepting it. And it is far from our contention; there is not reason to accept it. But the problem arises when the sample provides a value which is neither exactly equal to the parametric value, nor too far. In that situation one has to develop some procedures which enables one to decide whether to accept a contended (hypothetical) value or not on the basis of sample values. Such a procedure is known as testing of hypothesis

There can be more than one procedure to test a hypothesis, If so, which one to choose out of man tests at hand. For this purpose, many properties of tests have been put forward. A test with maximum virtues is most preferred one. so we have to deal with hypothesis test procedures and their properties.

A statistical test of a hypothesis is a rule or procedure which makes one to decide about the acceptance or rejection of the hypothesis.

13.2 STATISTICAL HYPOTHESIS

A quantitative statement about the population parameter is called a hypothesis. A statistical hypothesis is a statement about a population parameter or parameters of two or more populations. In other word, a statistical hypothesis is a statement about the distribution of one or more variables. For example, the life of the electric bulb in general is 2,500 hours. Two bulb manufacturing processes produce bulbs of the same average life etc. i.e.μ =2,500 hours, $\mu_1 = \mu_2$.

13.2.1 Types of Hypothesis

In general Statistical Hypotheses are of two types. Null hypothesis and alternative hypothesis

Null and alternative hypothesis

Such a hypothesis, which is usually a hypothesis of no difference is called null hypothesis and is usually denoted by H_0. According to Prof. R.A. Fisher, null hypothesis is the hypothesis which is tested for possible rejection under the assumption that it is true.

For example, in case of a single statistics, the null hypothesis H_0 will be that the sample statistics doesn't differ significantly from the hypothetical value of population parameter, and in the case of more than two statistics, the null hypothesis H_0 will be that the sample statistics don't differ significantly.

Having set up the null hypothesis we compute the probability 'p' that the deviation between the observed sample statistic and hypothetical parameter value might have occurred due to fluctuation of sampling. If the deviation comes out to be significant, null hypothesis is refuted or rejected at the particular level of significance adopted and if the deviation is not significant null hypothesis may be accepted at that level.

Any hypothesis which is complementary to the null hypothesis is called an alternative hypothesis, usually denoted by H_1, In other words, alternative hypothesis is a statement about the population parameter or parameters which provides an alternative to the null hypothesis within the range of pertinent value of the parameter. The idea of alternative hypothesis was propounded by J. Neyman. If H_0 is accepted, H_1 is rejected and vice versa.

For example, if we want to test the null hypothesis that the population has a specified mean μ_0 (say) i.e. $H_0 : \mu = \mu_0$ then the alternative hypothesis could be,

i. $H_0: \mu \neq \mu_0$ (i.e., $\mu > \mu_0$ or $\mu < \mu_0$)
ii. $H_0 : \mu < \mu_0$
iii. $H_0: \mu > \mu_0$

The alternative hypothesis (i) is known as a two tailed alternative hypothesis and the alternative hypotheses (ii) and (iii) are known as left tailed and right tailed alternatives hypothesis respectively. The setting of alternative hypothesis is very essential since it enables us to decide whether we have to use a single tailed (right or left) or two tailed test.

13.3 ERRORS IN SAMPLING

The main objective in sampling theory is to draw valid conclusions about the population parameters on the basis of the sample results. In practice we decide to accept or reject the lot after examining a sample from it.

There are the four possibilities in the testing of statistical hypothesis,

i. Accepting the null hypothesis when the null hypothesis is true.
ii. Rejecting the null hypothesis when the null hypothesis is true.
iii. Accepting the null hypothesis when the null hypothesis is false.
iv. Rejecting the null hypothesis when the null hypothesis is false.

The above decisions can be presented in the following table

	Decision from sample	
	Accept H_0	**Reject H_0**
When H_0 is True	Correct decision (No error)	Wrong decision (Type I error)
When H_0 is False	Wrong decision (Type II error)	Correct decision (No error)

In the testing of hypothesis, we may commit two types of errors. The error committed in rejecting null hypothesis H_0 when it is true is called type I error or the error of the first kind and its probability is denoted by α (alpha). The error committed in accepting null hypothesis H_0 when it is false is called type II error or the error of second kind and its probability is denoted by β (beta), thus

Type I error = Reject null hypothesis when it is true.

Type II error = Accept null hypothesis when it is wrong i.e. accept null hypothesis when alternative hypothesis is true.

α = Probability (Reject null hypothesis when it is true) = P (reject H_0 when H_0 is true)

β = Probability (Accept null hypothesis when it is wrong) =P (accept H_0 when H_1 is true)

Where α and β are called the size of type I error and type II error respectively.

In practice, type I error amounts to rejecting a lot when it is good and type II error may be regarded as accepting the lot when it is bad.

Thus, P (reject a lot when it is good) = α

P (accept a lot when it is bad) = β

Where, α and β are referred as producers risk and consumer's risk respectively

Let us consider a practical problem. If a medicine is administered to a few patients of a particular disease to cure them and the medicine is curing the disease, but it is claimed that it has no effect or has an adverse effect and hence it is discontinued. This is type I error on the contrary, the medicine has adverse effect and is claimed to have good effect and the treatment is continued, this is type II error. If we delve into the consequences of both types of errors, we find type II error is more serious than type I error, Hence α may be relatively large than β. In view of these facts effort is made to minimize β even for certain risk of type I error usually 0.01 or more.

13.4 POWER OF A TEST AND POWER FUNCTION

For testing a hypothesis H_0 against H_1, the test with probabilities α and β of type I and II errors respectively, the quantity $(1 - \beta)$ is called the power of the test.

The power of the test depends upon the difference between the parameter value specified by H_0 and the actual value of the parameter. $1 - \beta$ can be expressed as a function of the true parameter, say θ , If T is a test of H_0 , the power function of T is defined as the probability of refection of H_0 when the distribution, from which the sample is drawn, is parameterized by θ and denoted by $P_T(\theta)$. In terms of second kind of error, the function $1 - \beta(\theta)$ is known as the power function, whereas the function $\beta(\theta)$ is known as the operating characteristic function or OC function.

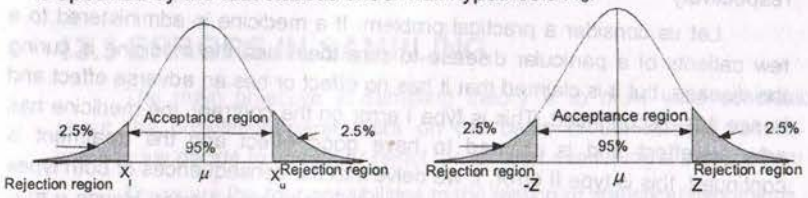
13.5 LEVEL OF SIGNIFICANCE

The level of significance may be defined as the probability of type of I error which we are ready to tolerate in making a decision about H_0.
I.e. P (reject H_0 when H_0 is true) = α.

It is our endeavour to carry out a test which minimizes both type of error; unfortunately for given set of observations, both the errors can't be controlled simultaneously. Hence it is a general practice to assign a bound to type I error and to minimize type II error, Thus one choose a value of α lying between 0 and 1 which is known as the level of significance . Hypotheses in general are tested at 1% or 5% level of significance. But the most commonly used level of significance in practice is 5%.

13.6 CRITICAL REGION OR REGION OF REJECTION

Let S denote the sample space of observations. Then sample space S is specified by the test statistic under null hypothesis H_0.

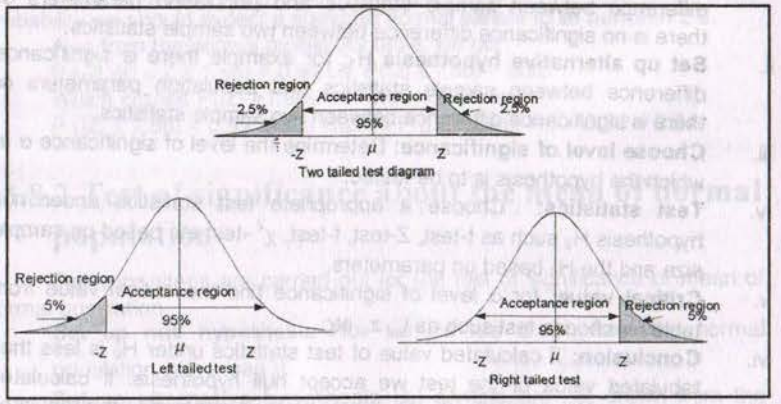


The space S is divided into two types of region ω and S-ω . The level of significance is the size of the critical regionω. The critical region is also known

as the region of rejection. The region of rejection may be situated on both the tails or on only one tail depending on the alternative hypothesis, also the region(s-ω) is called the acceptance region. If the value of the test statistics lies in the acceptance region, the null hypothesis H_0 is accepted and if it is lies in the critical region H_0 is rejected.

13.7 ONE TAILED AND TWO TAILED TEST

If an alternative hypothesis is such that it leads to two alternatives to the null hypothesis, it is said to be a two tailed test. For example, Test H_0: μ = 20 verses H_1: $\mu \neq 20$ leads to two sided test as μ can be greater than 20 or less than20. In this situation half of the area of critical region lies on the left tail and the half on the right. If α is the area of the critical region $\frac{\alpha}{2}$ is the area one both the tails.



Again, if the alternative hypothesis provides one sided alternative to H_0 e.g. H_0: μ =20 verses H_1 μ >20 or H_1: μ <20, the critical region or size α lies only on one tail. Specifically an area equal to α lies on the right tail given by ω:Z > Zα when H_1 is μ >20 and on the left tail when H_1 is μ < 20, an area equal to α given by ω : Z < Zα

The significance or critical value of Z for a single tailed test (left or right) at level of significance α is same as the critical value of Z for a two –tailed test at level of significance '2α'.

We give below, the critical value of Z at commonly used levels of significance of both two tailed and single tailed test, these values have been obtained from the normal probability table.

Critical value (Z_α)	level of significance (Z_α)				
	1%	2%	4%	5%	10%
Two tailed test	$\|Z_\alpha\| = 2.58$	$\|Z_\alpha\| = 2.33$	$\|Z_\alpha\| = 2.05$	$\|Z_\alpha\| = 1.96$	$\|Z_\alpha\| = 1.645$
Right tailed test	$Z_\alpha = 2.33$	$Z_\alpha = 2.05$	$Z_\alpha = 1.75$	$Z_\alpha = 1.645$	$Z_\alpha = 1.28$
Left tailed test	$Z_\alpha = -2.33$	$Z_\alpha = -2.05$	$Z_\alpha = -1.75$	$Z_\alpha = -1.645$	$Z_\alpha = -1.28$

Remarks: the above critical values are provide for large samples ($n > 30$). If n is small (i.e. $n \leq 30$) then the sample distribution of the test statistic Z will not be normal in that case we can't be use the above significant value. For small samples ($n \leq 30$) we use the significant values base on the exact sampling distribution of 't' statistics detail of this is given in next chapter of this book.

13.8 PROCEDURE FOR TESTING OF HYPOTHESIS OR TEST OF SIGNIFICANCE

We now summaries below the various steps on testing of a statistical hypothesis in a systematic manner,

i. **Set up null hypothesis H_o:** for example there is no significance difference between sample statistics and population parameters or there is no significance difference between two sample statistics.

ii. **Set up alternative hypothesis H_1:** for example there is significance difference between sample statistics and population parameters or there is significance difference between two sample statistics.

iii. **Choose level of significance**: Determine the level of significance α at which the hypothesis is to be tested.

iv. **Test statistics:** Choose a appropriate test statistics under null hypothesis H_0 such as t-test, Z-test, f-test, χ^2 -test etc based on sample size and the H_0 based on parameters.

v. **Critical value:** for α level of significance find the critical value from table for choose test such as t_α, z_α etc.

vi. **Conclusion:** If calculated value of test statistics under H_0 is less than tabulated value of the test we accept null hypothesis. If calculated value of test statistics under H_0 is greater than or equal to the tabulated value of the test we reject null hypothesis and accept our alternative hypothesis at level of significance α .

13.8.1 Test of significance for Large Sample

Large sample test is generally used when the sample size 'n' is greater than 30. The test statistic which is applied in case of large sample is called

Z-test or normal test and is defined as $Z = \frac{\text{difference}}{\text{S.E.}} = \frac{t - E(t)}{\text{S.E.}(t)}$

Where S.E (t) is the standard error of the sample statistics 't'.

The test of significance of parametric tests for sampling of variables and sampling of attributes of large sample are given by

a. Test of significance about the mean of a normal population.
b. Test of significance of a single mean.
c. Test of significance of a difference between two means.
d. Test of significance of a sample proportion.
e. Test of significance difference between two sample proportions.

In this chapter, we will discuss the test of significance when samples are large. We have seen that for large value of n the number of trials. Almost all the distribution e.g. binomial, Poisson, negative binomial etc are very closely approximated by normal distribution, Thus in this case we apply the normal test which is based up on the following fundamental property (area property) of the normal probability curve.

If $X \approx N(\mu, \sigma^2)$ then $Z = \frac{X - \mu}{\sigma} \approx N(0, 1)$

Thus from the normal probability table we have $P(-3 \leq Z \leq 3) = 0.9927$ i.e. $P(|Z|) \leq 3) = 0.9973$,

Which implies $P(|Z|) > 3) = 1 - P(|Z| \; P(|Z|) \leq 3) = 0.0027$ i.e. in all probability we should expect a standard normal variate to lie between ± 3.

Also from the normal probability table, we get

$P(-1.96 \leq Z \leq 1.96) = 0.95$ i.e. $P(|Z| \leq 1.96) = 0.95$

Which implies $P(|Z| > 1.96) = 1 - 0.95 = 0.05$ and

$P(|Z| \leq 2.58) = 0.99$ which implies $P(|Z| > 1.96) = 1 - 0.99 = 0.01$

13.8.2 Test of significance about the mean of normal population

Following steps are carried out for the test of significance of mean of normal population

i. **Set up null hypothesis Ho:** sample are drawn from the normal population with mean μ

ii. **Set up alternative hypothesis H_1:** sample are not drawn from the normal population with mean μ

iii. **Test statistics:** Compute the Z- statistics under H_0 as

$Z = \frac{X - u}{\sigma}$, follows normal distribution with mean 0 and variance 1 i.e. N(0, 1)

iv. **Conclusion:** If $|Z| > 3$, H_0 is always rejected and If $|Z| \leq 3$, we test its significance at certain level of significance, usually at 5% and some times at 1% level of significance. Thus for a two tailed if

$|Z| > 1.96$ H_0 is rejected at 5% level of significance.

Similarly if $|Z| > 2.58$ H0 is contradicted at 1% level of significance and if $|Z| \leq 2.58$ H_0 may be accepted at 1 % level of significance.

From the normal table we have
$P(Z > 1.645) = 0.5 - P(0 \le Z \le 1.645) = 0.5 - 0.45 = 0.05$ and
$P(Z > 2.33) = 0.5 - P(0 \le Z \le 2.33) = 0.5 - 0.0.49 = 0.01$
Hence, for a single tail test (Right or left tail test) we compare the computed value of |Z| with 1.645 at 5% level of significance and 2.33 at 1% level of significance and accept or reject H_0 accordingly.

13.8.3 Test of Significance for Single Mean for Large Sample

We have for random sample $x_1, x_2 \ldots x_n$ of size n from a normal population with mean μ and variance σ^2, then the sample mean $\bar{x}$ is distributed normally with mean μ and variance $\frac{\sigma^2}{n}$ i.e. $\bar{x} \approx N(\mu, \sigma^2)$. However, this result holds even in random sample from non-normal population provided the sample size 'n' is large (Central Limit Theorem). Then for large sample the SNV corresponding to $\bar{x}$ is

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

Following are the steps for test of significance of single mean in large sample

i. **Set up null hypothesis H_o**: $\mu = \mu_0$ (say) or $H_0 : \mu = \bar{x}$ i.e. population mean has specified value say μ_0 or there is no significance difference between population mean μ and sample mean $\bar{x}$

ii. **Set up alternative hypothesis H_1** : $\mu \neq \mu_0$ (say) or $H_1 : \mu \neq \bar{x}$ (two tailed test) there is no significance difference between population mean $\mu$and sample mean $\bar{x}$

Or, $H_1: \mu > \mu_0$ (say) or $H_1: \mu > \bar{x}$ (right tailed test) population mean μ is greater than sample mean $\bar{x}$

Or, $H_1: \mu < \mu_0$ (say) or $H_1: \mu < \bar{x}$ (left tailed test) population mean μ is less than sample mean $\bar{x}$.

iii. **Choose level of significance**: Determine the level of significance α at which the hypothesis is to be tested.

iv. **The test statistics:** under the null hypothesis $H_0 : \mu = \mu_0$ the test statistics is given by $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$

If the population standard deviation is unknown then we use its estimate provided by the sample variance given by $\hat{\sigma}^2 = s^2 \Rightarrow \hat{\sigma} = s$ (for large samples size) i.e. $Z = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \approx N(0, 1)$

v. **Critical value:-** for α level of significance, find the critical value of Z form standard normal table as Z_α

vi. **Conclusion:-**If calculated value of Z i.e. |Z| is less than Z_α we accept our null hypothesis H_0 at α level of significance and reject otherwise. And at last draw valid conclusion on the basis of hypothesis testing.

13.8.4 Confidence Limits or Fiducial Limits

The approximate $100(1 - \alpha)$ % confidence limits for estimation of population mean μ from large sample at α level of significance is given by the definition of critical region.

$$P(-Z_\alpha < Z < Z_\alpha) = 1 - \alpha$$

$$P\left(-Z_\alpha < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_\alpha\right) = 1 - \alpha$$

The next step is to rearrange the terms in brackets so that the mean remains alone in the center. The probability statement then becomes

$$P\left(\bar{X} - Z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

This probability statement yields the confidence interval. The probability is $1 - \alpha$ that the random interval from $\bar{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}$ to $\bar{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}$ will cover mean. The observed value of this interval is called a $100(1 - \alpha)$% confidence interval for mean μ

$\therefore$ Upper limit for population mean $(\mu) = \bar{X} + Z_\alpha \frac{\sigma}{\sqrt{n}}$ and

Lower limit for population mean $(\mu) = \bar{X} - Z_\alpha \frac{\sigma}{\sqrt{n}}$

Thus the 95% and the 99% confidence limits for population mean μ are

$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$ and $\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}}$

However, in sample from a finite population of size N, the corresponding 95% and 99% confidence limits for mean μ are respectively

$\bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$ and $\bar{X} \pm 2.58 \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$

Example 1

A sample of size 400 was drawn and the sample mean was found to be 99. Test whether this sample could have come from a normal population with mean 100 and variance 64 at 5% level of significance.

Solution

We are given population mean $\mu = 100$ and population variance $\sigma^2 = 64$, sample size $n = 400$ and sample mean $\bar{x} = 99$.

We use the following steps to test the significance of mean as

i. **Null hypothesis H_0:** $\mu = 100$ i.e. samples are drawn from a normal population with mean 100 and variance 64.

ii. **Alternative hypothesis H_1:** $\mu \neq 100$ i.e. samples are not drawn from a normal population with mean 100 and variance 64.

iii. Choose the level of significance α as 5%.= 0.05

iv. **Test statistics:** under The null hypothesis H_0: $\mu = 100$, the test statistics is $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$

$$Z = \frac{99 - 100}{8/\sqrt{400}} = 2.5$$

$|Z| = 2.5$

v. **Critical value:** For $\alpha = 5\%$ level of significance and two tailed test the critical value of Z ie $Z_\alpha = Z_{0.05} = 1.96$

vi. **Conclusion:** Here $|Z| = 2.5$ is greater than $Z\alpha = Z_{0.05} = 1.96$, we reject our null hypothesis H_0: $\mu = 100$ at 5% level of significance and accept our alternative hypothesis and conclude that sample are not drawn from the normal population with mean 100 and variance 64.

Example 2

Each days, Nepal customs service has historically intercept about Rs. 28 thousand in contraband goods being smuggled in to the country, with a standard deviation of Rs. 16 thousand per day. On 64 randomly chosen days in 2005, the Nepal's customs service intercepted an average of Rs. 30.3 thousand in contraband goods. Does this sample indicate that the customs commissioner should concerned that smuggling has increased above its historical level at 5% level of significant.

Solution

We are given sample size (n) = 64 and sample mean $\bar{x}$ = Rs. 30.3 thousand, population mean = Rs 28 thousand and population standard deviation σ = Rs. 16 thousand.

We proceeded to test of significance of mean as following stages

i. **The Null hypothesis H_0:** $\mu = 28$ i.e. average contraband goods being smuggled in the country per day is Rs.28 thousand with variance Rs.16. thousand.

ii. **Alternative hypothesis H_1:** $\mu > 28$ (right tailed) i.e. average contraband goods being smuggled in the country per day is increased.

iii. Choose the level of significance α = as 5%.=0.05

iv. **Test statistics**: under the null hypothesis H_0: $\mu = 28$ the test statistics is $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$

$$Z = \frac{30.3 - 28}{16/\sqrt{64}} = 1.15$$

$Z = 1.15$

v. **Critical value:** For $\alpha = 5\%$ level of significance and right tailed test the critical value of Z i.e. $Z_\alpha = Z_{0.05} = 1.645$

vi. **Conclusion:** Here $|Z| = 1.15$ is less than $Z_\alpha = Z_{0.05} = 1.164$, we accept our null hypothesis H_0: $\mu = 28$ at 5% level of significance and conclude that average contraband goods being smuggled in the country per day is not increased.

Example 3

A sample of 100 light bulbs from a manufacturing lot had average life of 1416 hours with standard deviation of 30 hours.

i. Calculate the Standard error of the mean.

Ii. Set up 95% and 99% confidence limit or fiducial limits for the true population mean.

Solution

We are given sample size (n) = 100 and sample mean $\bar{x}$ = 1416 hrs, sample standard deviation's' = 30hrs.

i. The standard error of mean is given by $S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}} = \frac{30}{\sqrt{100}} = 3$

Since for large sample $\hat{\sigma} = s$

ii. The $(1 - \alpha)\%$ confidence limits for estimation of population mean μ is given by $\bar{X} \pm Z\alpha\ S.E.(\bar{X})$

Thus for 95% confidence limits $(1 - \alpha)\% = 95\% \Rightarrow \alpha = 5\% = 0.05$

$\therefore Z_\alpha = 1.96$ from table, Hence 95% confidence limits for estimating population mean are given by $\bar{X} \pm 1.96.S.E(\bar{X}) = 1416 \pm 1.96 \times 3$

$\therefore$ Upper limit for population mean $\mu = 1416 - 1.96 \times 3 = 1421.83$ hrs.

Lower limit for population mean $\mu = 1416 - 1.96 \times 3 = 1410.12$ hrs.

Similarly 99% confidence limits for population mean μ are

$\bar{X} \pm 2.58 S.E.(\bar{X}) = 1416 \pm 2.58 \times 3$

$\therefore$ Upper limit for population mean $\mu = 1416 + 2.58 \times 3 = 1423.74$ hrs.

Lower limit for population mean $\mu = 1416 - 2.58 \times 3 = 1408.26$ hrs.

13.8.5 Hypothesis Testing Using Binomial Distribution

Here we shall consider sampling from a population which is divided into two mutually exclusive and collectively exhaustive classes one class possessing a particular attribute say A, and other class not possessing that attribute and then note down the number of persons in the sample of size n possessing that attribute. The presence of an attribute in a sampled unit may be termed as success and its absence as failure. In this case a sample of 'n' observations is identified with that of a series of n independent Bernoulli trials with constant probability 'P' of success for each trial. Then the probability of x-success in n trials as given by the binomial probability distribution as

$P(x) = {}^nC_x p^x Q^{n-x}$, x = 0, 1 .....n

If x is the number of success in n independent trials with constant probability p of success for each trial then E(X) = n P and V(X) = n PQ where Q = 1 - P is the probability of failure. It has been proved that for large n, the binomial distribution tends to normal distribution, hence for large n

$$X \approx N(nP, nPQ) \text{ i.e. } z = \frac{X - E(X)}{\sqrt{V(X)}} = \frac{X - nP}{\sqrt{nPQ}} \approx N(o,1)$$

And we can apply the normal test by using the following steps

i. **Set up null hypothesis H_0**: Samples are drawn from Binomial population.
ii. **Set up alternative hypothesis H_1**: Samples are not drawn from Binomial population.
iii. Fix the level of significance α.
iv. **Test statistics**: Under the null hypothesis H_0 , the test statistics for X success is $Z = \frac{X - nP}{\sqrt{nPQ}} \approx N(o,1)$
v. **Critical Value:** For level of significance α find the critical value of Z from table as Z_α.
vi. **Conclusion:** If calculated value of Z is less than Z_α we accept our null hypothesis, reject otherwise. And draw a valid conclusion on the basis of test.

Example 4

A die is thrown 9,000 times and a throw of 3 or 4 is observed 3,240 times. Show that the die can't be regarded as an unbiased one.

Solution

If the coming of 3 or 4 is called a success, then in usual notation we are given (n) = 9,000

X, number of success = 3240,

Probability of success = P (of getting 3 or 4) $= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

We proceeded in following stages to test the unbiasedness of the coin

i. **Null hypothesis H_0**: is that the coin is unbiased one i.e. Probability of success = P(of getting 3 or 4) $= \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$
ii. **Alternative hypothesis H_1:** the die is biased i.e. $P \neq 1/3$
iii. Choose that the level of significance α = 5% = 0.05.
iv. **Test statistics**; Under the null hypothesis H_0 , the test statistics for X success is $Z = \frac{X - nP}{\sqrt{nPQ}} \approx N(o,1)$

$$Z = \frac{3240 - 9000 \times 1/3}{\sqrt{9000 \times 1/3 \times 2/3}} = 5.36$$

v. **Conclusion:** Since |Z| > 3, our null hypothesis is rejected and we conclude that the die is almost certainly biased.

13.8.6 Test of Significance for Single Proportion

In a sample of size n, let X be the number of persons possessing the given attribute. Then observed proportion of success $= \frac{X}{n} = p$ (say)

$E(p) = E\left(\frac{X}{n}\right) = \frac{1}{n} E(X) = \frac{nP}{n} = P$, which showed that sample proportion 'p' is an unbiased estimate of the population proportion "P".

Also, $V(p) = V\left(\frac{X}{n}\right) = \frac{1}{n^2} V(X) = \frac{nPQ}{n^2} = \frac{PQ}{n}$

Now standard error of p = S.E.(p) $= \sqrt{V(p)} = \sqrt{\frac{PQ}{n}}$

Thus for large n, x and consequently $\frac{X}{n}$ is asymptotically normal, the normal test for proportion of success being

$$Z = \frac{p - E(p)}{\sqrt{V(p)}} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \approx N(0, 1)$$

In order to test whether there is significance difference between the sample proportion 'p' and the population proportion 'P' we proceeded as follows

i. **Set up null hypothesis H_0:** $P = P_0$ (say) i.e. population proportion has specified value say P_0.
ii. **Set up alternative hypothesis H_1**: $P \neq P_0$ (two tailed) i.e. population proportion is not equal to specified value say P_0.
Or H_1: $P > P_0$ (right tail tailed) i.e. population proportion is greater than specified value say P_0.
Or H_1: $P < P_0$ (left tailed) i.e. population proportion is less than specified value say P_0

iii. Fix the level of significance α.

iv. **Test statistics:** Under the null hypothesis $H_0 : P = P_0$, the test statistics for proportion of success(p) is

$$Z = \frac{p - E(p)}{\sqrt{V(p)}} = \frac{p - P}{\sqrt{\frac{PQ}{n}}} \approx N(0, 1)$$

If we have sample from a finite population of size N, then

$$S.E.(p) = \sqrt{V(p)} = \sqrt{\frac{(N - n)}{N - 1}\frac{PQ}{n}}$$

v. **Critical value:** For α level of significance find the critical value of Z as Z_α from normal table.

vi. **Conclusion:** If $|Z| < Z_\alpha$ we accept our null hypothesis at α level of significance and If $|Z| \geq Z_\alpha$ we reject our null hypothesis and accept alternative hypothesis.

13.8.7 Confidence Limits for Population Proportion

Since the probable limits for a normal variate X are $E(X) \pm 3\,S.E(X)$, the probable limits for the observed proportion of success are

$$E(p) \pm 3\,S.E(p) \text{ i.e. } P \pm 3\sqrt{\frac{PQ}{n}}$$

If population proportion P is not known taking sample proportion p as an estimate of population proportion P, the probable limits for the population proportion are $p \pm 3\sqrt{\frac{pq}{n}}$

And the approximate $100(1 - \alpha)$ % confidence limits for estimating the population proportion are $p \pm Z\alpha\sqrt{\frac{pq}{n}}$

In particular, the 95% confidence limits for 'P' are given by $p \pm 1.96\sqrt{\frac{pq}{n}}$

And 99% confidence limits for 'P' are given by $p \pm 2.58\sqrt{\frac{pq}{n}}$

However, in sample from a finite population of size N, the corresponding 95% and 99% confidence limits for population proportion P are respectively

$$p \pm 1.96\sqrt{\frac{pq}{n}}\sqrt{\frac{N - n}{N - 1}} \text{ and } p \pm 2.58\sqrt{\frac{pq}{n}}\sqrt{\frac{N - n}{N - 1}}$$

Example 5

In a sample of 1,000 people in Kathmandu valley 540 are MO:MO eaters and the rest are Piazza eaters. Can we assume that both MO: MO and Piazza are equally popular in Kathmandu valley at 1% level of significance.

Solution

In the usual notation, we are given sample size (n) = 1,000

Let MO:MO eaters are success X-The number of MO:MO eaters = 540,

p = sample proportion of MO:MO eaters = $\frac{X}{n} = 0.54$

To test the significance of single proportion we proceeded in the following stages,

i. **Null hypothesis H_o:** P= 0.5=Q i.e. Both MO: MO and Pizza are equally popular in Kathmandu valley. So that P = population proportion of MO: MO eaters in Kathmandu = $0.5 \Rightarrow Q = 1-P= 0.5$.

ii. **Alternative hypothesis H_1 :** $P \neq 0.5$ (two tailed test) i.e. MO: MO and Pizza are not equally popular in Kathmandu valley.

iii. Given the level of significance $\alpha = 1\% = 0.01$.

iv. **Test statistics:** Under the null hypothesis: P= 0.5=Q, the test statistics is

$$Z = \frac{p - P}{\sqrt{PQ/n}} \approx N(0, 1)$$

$$Z = \frac{0.54 - 0.5}{\sqrt{0.5 \times 0.5/1000}} = 2.529$$

v. **Critical value**: Here for $\alpha = 1\%$ level of significance , the critical value of Z for two tailed test is $Z_{0.01} = 2.58$

vi. **Conclusion**: Here $|Z| = 2.529$ is less than $Z_\alpha = 2.58$, we accept our null hypothesis H_o: P = 0.5 =Q at 1% level of significance and conclude that Both MO: MO and Pizza are equally popular in Kathmandu valley.

Example 6

Forty people were admitted in a B.E. Civil department and only 36 passed. Will you reject the hypothesis that the Pass rate is 85% in favour of the hypothesis that it is more, at 5% level of significance?

Solution

In the usual notation, we are given sample size (n) = 40

Let the students who passed the B.E. exam is success, X- number of students who passed the B.E. exam = 36, p = sample proportion of students passed the exam= X/n = 36/40 = 0.9

To test the significance of single proportion, we proceeded in the following stages,

i. **Null hypothesis H_o:** P = 0.85 i.e the proportion of students passed B.E. exam is 85% $\Rightarrow$ Q = 1– P= 0.15.

ii. **Alternative hypothesis H_1 : P >** 0.85 (right tailed test) i.e. passed proportion is increased

iii. Given the level of significance $\alpha = 5\% = 0.05$.

iv. **Test statistics:** Under the null hypothesis: P= 0.85, the test statistics is

$$Z = \frac{p - P}{\sqrt{PQ/n}} \approx N(0, 1)$$

$$Z = \frac{0.9 - 0.85}{\sqrt{0.85 \times 0.15/40}} = 0.885$$

v. **Critical value**: Here for α = 5% level of significance , the critical value of Z for right tailed test is $Z_{0.05} = 1.645$

vi. **Conclusion:** Here $|Z| = 0.885$ is less than $Z_\alpha = 1.645$, we accept our null hypothesis H_o: P= 0.85 at 5% level of significance and conclude that passed proportion of student in B.E. exam is 85%.

Example 7

A random sample of 500 compact discs (CDs) was taken from a large consignment and 60 were found to be defective. Obtain the 98% confidence limits for the percentage of defective CDs in the consignment.

Solution

In the usual notation, we are given sample size (n) = 500

Let, the defective CDs are success then, number of defective - CDs X = 60

$$p = \text{sample proportion of defective CDs} = \frac{X}{n} = \frac{60}{500} = 0.12$$

Here for 98% confidence limits the level of significance is 2% and from table the critical value of Z is $Z_\alpha = Z_{0.02} = 2.33$

Therefore 98% confidence limits for population proportion of defective CDs is $p \pm 2.33\sqrt{\frac{pq}{n}}$

$$0.12 \pm 2.33\sqrt{\frac{0.12 \times 0.88}{500}}$$

Here, P is unknown so P is estimated by sample proportion 'p'

$\therefore$ Upper limit = 0.12 + 2.33 x0.0145 = 0.154 and

Lower limit = 0.12 - 2.33 x0.0145 = 0.086

Hence 98% confidence limits for percentage of defective CDs in the consignments are (8.61, 15.4)%.

13.8.8 Test of Significance for Difference of Mean for Large Sample

Let two samples of sizes n_1 and n_2 are drawn from two normal population with mean μ_1 & μ_2 and variances σ_1^2 & σ_2^2 Let $\bar{x}_1$ & $\bar{x}_2$ be sample means of two samples , then for large sample $\bar{x}_1$ & $\bar{x}_2$ is normally distributed with means μ_1 & μ_2 and variances σ_1^2/n_1 & σ_2^2/n_2 Also($\bar{x}_1 - \bar{x}_2$) , being the difference of two independence normal variate also a normal variate the value of Z(SNV) corresponding to ($\bar{x}_1 - \bar{x}_2$) is given by

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.(\bar{x}_1 - \bar{x}_2)} \approx N(0, 1)$$

Then the stapes for test of significance of difference between two means are as follows.

i. **Null hypothesis H_o:** $\mu_1 = \mu_2$ i.e. two population means are equal or samples are drawn form two population having equal population mean.

ii. **Alternative hypothesis H_1:** $\mu_1 \neq \mu_2$ (two tailed test) i.e. two population means are not equal or samples are drawn form two population having different population mean.

Or H_1: $\mu_1 > \mu_2$ (right tailed test) i.e. population mean of first is greater than second

Or H_1: $\mu_1 < \mu_2$ (left tailed test) i.e. Population mean of first is smaller than second

iii. Fix the level of significance α.

iv. **Test statistics:-** Under the null hypothesis H_o: $\mu_1 = \mu_2$ the test statistics is $Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E.(\bar{x}_1 - \bar{x}_2)} \approx N(0, 1)$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1) - E(\bar{x}_2)}{S.E.(\bar{x}_1 - \bar{x}_2)} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \approx N(0, 1)$$

[$\because E(\bar{x}_1) = E(\bar{x}_2) = \mu$ and

$S.E.(\bar{x}_1 - \bar{x}_2) = Var(\bar{x}_1) + var(\bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$, the covariance term vanishes since the sample mean are independent

Remarks 1

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ i.e. if the samples have been drawn from the population with common variance then under null hypothesis

H_o: $\mu_1 = \mu_2$ the test statistics is $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

Remarks 2

If σ is not known, then its estimate based on the sample variance is used .if the sample size are not sufficient large then an unbiased estimate of σ^2 is given as $\hat{\sigma}^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$

But, since sample size are large $S_1^2 \approx s_1^2$, $S_2^2 \approx s_2^2$, $(n_2 - 1) \approx n_2$ and $(n_2 - 1) \approx n_2$, therefore in practice for large sample , the following estimate of σ^2 is used as $\hat{\sigma}^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2}$

Remarks 3

If $\sigma_1^2 \neq \sigma_r^2$ and are not known, then they are estimated from sample value. This result is in some error which is practically immaterial; if samples are large these estimates for large samples are given by

$\sigma_1^2 \approx S_1^2 \approx s_1^2$ and $\sigma_2^2 = S_2^2 \approx s_2^2$, since sample are large, thus under null hypothesis H_o: $\mu_1 = \mu_2$ the test statistics is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0, 1)$$

v. **Critical value:** For α level of significance find the critical value of Z as Z_α from normal table.

vi. **Conclusion:** if |Z| less than Z_α, we accept our null hypothesis H_o: $\mu_1 = \mu_2$ at α level of significance, reject otherwise.

13.8.9 Confidence Limits for Difference of Means

The approximate (1 - α)100% confidence limit for estimation of the difference of population mean is given by $(\bar{x}_1 - \bar{x}_2) \pm Z_\alpha$ S.E. $(\bar{x}_1 - \bar{x}_2)$ i.e.

$P\{(\bar{x}_1 - \bar{x}_2) - Z_\alpha SE.(\bar{x}_1 - \bar{x}_2) < (\mu_1 - \mu_2) < (\bar{x}_1 - \bar{x}_2) + Z_\alpha SE(\bar{x}_1 - \bar{x}_2)\} = (1 - \alpha)$

In particular, the approximate 95% confidence limits for difference of mean is $(\bar{x}_1 - \bar{x}_2) \pm 1.96$ S.E $(\bar{x}_1 - \bar{x}_2)$

The approximate 99% confidence limits for difference of mean is

$(\bar{x}_1 - \bar{x}_2) \pm 2.58$ S.E $(\bar{x}_1 - \bar{x}_2)$

Example 8

In a certain factory there are two independent processes manufacturing the same items. The average weight in a sample of 250 items produced from one process is found to be 120 grams with a s.d. of 12 grams, while the corresponding figures in a sample of 400 items from the other process are 124 and 14. Find the S.E. of the difference of means and also test whether two mean weights differ significantly or not at 10% level of significance.

Solution

With the usual notation we are given

Sample I: $n_1 = 250$ $\bar{X}_1 = 120$ grams $s_1 = 12$ grams

Sample II: $n_2 = 400$ $\bar{X}_2 = 124$ grams $s_2 = 14$ grams

Then we proceeded in the following stapes for test of significance of difference between two mean weights as follows.

i. **Null hypothesis H_o:** $\mu_1 = \mu_2$ i.e. there is no significance difference between two mean weights.

ii. **Alternative hypothesis H_1:** $\mu_1 \neq \mu_2$ (two tailed test) i.e. there is significance difference between two mean weights.

iii. Given the level of significance $\alpha = 10\% = 0.10$

iv. **Test statistics:** Under the null hypothesis H_o: $\mu_1 = \mu_2$ the test statistics is $Z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0, 1)$

Here for large sample size $\sigma_1^2 \approx S_1^2 \approx s_1^2$ and $\sigma_2^2 \approx S_2^2 \approx s_2^2$

$$Z = \frac{120 - 124}{\sqrt{\frac{144}{250} + \frac{196}{400}}} = -3.87$$

Where S.E. $(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{144}{250} + \frac{196}{400}} = 1.032$

v. **Critical value:** For $\alpha = 10\%$ level of significance the critical value of Z as $Z_{0.10} = 1.645$ from normal table.

vi. **Conclusion:** Here |Z| = 3.87 is greater than $Z_{0.10} = 1.645$, we reject our null hypothesis H_o: $\mu_1 = \mu_2$ at 10% level of significance and accept our alternative hypothesis H_1: $\mu_1 \neq \mu_2$ and concluded that there is significant difference between two mean weights.

Example 9

If 60 B.E. Civil students are found to have a mean height of 63.6 inches and 50 B.E. Computer students have a mean height of 69.51 inches. Would you conclude that the computer students are taller than civil students? Assume the standard deviation of height of B.E. students of the college is 2.48 inches.

Solution

We are given as, Mean height of B.E. Civil students $(\bar{X}_1) = 63.6$ inches. And number of B.E. Civil students $n_1 = 60$,

Mean heights of B.E. Computer students $(\bar{X}_2) = 69.51$ inches. And number of B.E. Computer students $n_2 = 50$

Population standard deviation height of students of B.E. students is $\sigma = 2.48$ inches

Then we proceeded in the following stapes for test of significance of difference between two mean heights as follows.

i. **Null hypothesis H_o:** $\mu_1 = \mu_2$ i.e. mean height of B.E. Civil students and B.E. Computer students are equal.

ii. **Alternative hypothesis H_1:** $\mu_1 < \mu_2$ (Left tailed test) i.e. mean height of B.E. Civil students is less than that of B.E. Computer students.

iii. Let the level of significance $\alpha = 5\% = 0.05$

iv. **Test statistics:** Under the null hypothesis H_o: $\mu_1 = \mu_2$ the test statistics is

$$Z = \frac{\bar{x}_1 - \bar{x}_2}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{63.6 - 69.51}{2.48\sqrt{\frac{1}{60}+\frac{1}{50}}} = -12.44$$

v. **Critical Value**: For α = 5% level of significance and for left tailed test the critical value of Z as $Z_{0.05}$ = 1.645 from normal table.

vi. **Conclusion:** Here |Z| = 12.44 is greater than $Z_{0.05}$ = 1.645, we reject our null hypothesis H_o: $\mu_1 = \mu_2$ at 5% level of significance and accept our alternative hypothesis H_1: $\mu_1 \neq \mu_2$ and concluded that B.E. Computer students are taller than B.E. Civil students.

Example 10

The means heights of two samples of 1000 and 2000 individual are 67.5 inches and 68 inches respectively. Find the 95% confidence limits for the difference of population mean if two samples are drawn from the population having s.d. of 2.5 and 2.7 respectively.

Solution

With the usual notation we are given

	Sample I	Sample II
Sample size	n_1 =1000	n_2 =2000
Averages	$\bar{X}_1$ = 67.5 inches	$\bar{X}_2$ = 68 inches
Population s. d.	σ_1 = 2.5	σ_2 = 2.7

We have the approximate 100(1 – α)% confidence limit for estimation the difference of population mean is given by $(\bar{x}_1 - \bar{x}_2) \pm Z_\alpha$ S.E $(\bar{x}_1 - \bar{x}_2)$, here critical value of Z for 5% level of significance is $Z_{0.05}$ = 1.96

Therefore 95% confidence limits for difference of mean is

$$(\bar{x}_1 - \bar{x}_2) \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$$

$$(67.5 - 68) \pm 1.96\sqrt{\frac{(2.5)^2}{1000}+\frac{(2.7)^2}{2000}}$$

$$(-0.5) \pm 1.96 \times 0.0995$$

Lower limit for difference of mean = – 0.305

And upper limit for difference of mean = – 0.695

13.8.10 Z-tests for Difference Between two Population Proportions

Suppose we want to compare two distinct populations with respect to the prevalence of a certain attribute, say (A) among their member. Let x_1 and x_2 be the number of persons. Possessing the given attribute A in random

sample of sizes n_1 and n_2 from the two population respectively. Thus sample proportion are give by $p_1 = \frac{x_1}{n_1}$ and $p_2 = \frac{x_2}{n_2}$

If P_1 and P_2 are population proportion then

$E(p_1) = P_1$ and $E(p_2) = P_2$

$$V(p_1) = \frac{P_1Q_1}{n_1} \text{ and } V(p_2) = \frac{P_2Q_2}{n_2}$$

Since for large sample, p_1 and p_2 are independently and asymptotically normally distributed then $(p_1 - p_2)$ is also normally distributed. Then the standard variable corresponding to the different $(p_1 - p_2)$ is given by

$$Z = \frac{(p_1-p_2) - E(p_1-p_2)}{\sqrt{\text{var}\,(p_1 - p_2)}} \approx N(0,1)$$

$$= \frac{(p_1 - p_2) - (P_1 - P_2)}{\sqrt{\frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}}}$$

Where var $(p_1 - p_2) = V(p_1) + V(p_2) - 2\,\text{Cov}\,(p_1.p_2)$

Since sample proportion are independent therefore Cov $(p_1,p_2) = 0$

$$\therefore \text{Var}(p_1 - p_2) = V(p_1) + V(p_2) = \frac{P_1Q_1}{n_1}+\frac{P_2Q_2}{n_2}$$

Following are the steps in test of significance for difference of two proportions for large sample size are as follow

i. **Null hypothesis $H_0 = P_1 = P_2$** = P(say) i.e. two population proportion are same in other words there is no significance difference between two sample proportion p_1 and p_2

ii. **Alternative hypothesis $H_1 = P_1 \neq P_2$** (Two tailed test i.e. Two population proportion are not same or two samples are drawn from difference population or

H_1: $P_1 > P_2$ (right tailed test)i.e. population proportion of 1st population is greater than that of second or

H_1: $P_1 < P_2$ (Left tailed test)

That is population proportion of first is less then that of second

iii. Fix level of significance α

iv. **Test Statistics:** Under H_0 : $P_1 = P_2$ the test statistics is

$$Z = \frac{(p_1-p_2) - E(p_1-p_2)}{\sqrt{\text{var}\,(p_1 - p_2)}} \approx N(0,1)$$

$$Z = \frac{(p_1 - p_2)}{\sqrt{PQ\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \quad [\because P_1 = P_2 = P]$$

In general if common population proportion of two populations is unknown, then an unbiased estimate of the population proportion P, based on both the sample is given by

$$\hat{P} = \frac{n_1p_1 + n_2p_2}{n_1 + n_2} = \frac{x_1 + x_2}{n_1 + n_2}$$

v. **Critical value:** For α level of significance we find the critical value of Z from Table as $Z\alpha$.

vi. **Conclusion:** If $|Z| < Z\alpha$ we accept the null hypothesis H_0 at α % level of significance and If $|Z| \geq Z\alpha$ we reject the null hypothesis H_0 and accept the alternative hypothesis H_1.

13.8.11 Confidence Interval for Difference of Population Proportion

The approximate (1 - α) 100% confidence limits to estimate the difference of population proportion (P_1 - P_2) is given.

$(p_1 - p_2) \pm Z\alpha$ S.E.$(p_1 - p_2)$

$P\{(p_1 - p_2) - Z\alpha$ S.E$(p_1 - p_2) < (P_1 - P_2) < (p_1 - p_2) + Z\alpha$ S.E $(p_1 - p_2)\} = 1 - \alpha$

The 99% and 95% confidence limits are calculate using the formula

$(p_1 - p_2) \pm 2.58$ S.E $(p_1 - p_2)$ and

$(p_1 - p_2) \pm 1.96$ S.E.$(p_1 - p_2)$

Example 11

Random sample of 400 men and 600 women were asked whether they world like certain T.V. program. 200 men and 325 women were in favour of the proposal. Test the hypothesis that proportion of men & women in favour of the proposal are same against that they are not, at 5% level.

Solution

We are given $n_1 = 400$, X_1- number of men favoring the proposal = 200

$n_2 = 600$ & X_2ber of women favoring the proposal = 325.

p_1 - proportion of men favoring the proposal in sample = $\frac{x_1}{n_1} = \frac{200}{400} = 0.5$

p_2 - proportion of women favoring the proposal = $\frac{x_2}{n_2} = \frac{325}{600} = 0.541$

i. **The Null hypothesis:** H_0: $P_1 = P_2$ (say) i.e there is no significant different between open in man and women as for as proposal of certain T.V. program is concerned

ii. **The Alternative hypotheses** H_1: $P_1 = P_2$ (Two-tailed test).

iii. The level of significant α = 5% = 0.05

iv. **The Test statistics:** Under H_0: $P_1 = P_2$ the test statistics is given by

$$Z = \frac{(p_1 - p_2)}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx N(0,1)$$ Here population proportion is not given

and it's value is estimates as $\hat{P} = \frac{x_1+x_2}{n_1+n_2} = 0.525$

$\therefore \hat{Q} = 1 - \hat{P} = 1 - 0.525 = 0.475$

$$\therefore Z = \frac{0.5 - 0.541}{\sqrt{0.525 \times 0.475 \times \left(\frac{1}{400} + \frac{1}{600}\right)}} = \frac{-0.041}{\sqrt{0.001039}} = -1.269$$

v. **Critical value**: Critical value of Z for α= 5% level of significant is 1.96.

vi. **Conclusion:** Here $|Z| = 1.269 < Z_{0.05} = 1.96$. We accept H_0 at 5% of level. Thus we conclude that men & women do not differ significantly as regards proposal of certain T.V. program concerned.

Example 12

A machine produced 20 defective articles in a batch of 400 items. After overhauling it produced 10 defective articles in a batch of 300. Was the machine improved ?

Solution

From the question we are given that

The sample size before over hauling $n_1 = 400$ item

Numbers of defective items before over-hauling $x_1 = 20$

p_1 proportion of defect before over hauling = $\frac{x_1}{n_1} = \frac{20}{400} = \frac{1}{20}$

Sample size after over hauling = $n_2 = 300$

Numbers of defective after over hauling $x_2 = 10$

p_2 proportion of defect after over hauling = $\frac{x_2}{n_2} = \frac{10}{300} = \frac{1}{30}$

And estimated value of population proporti0on of machine is

$$\hat{P} = \frac{x_1+x_2}{n_1+n_2} = \frac{20 + 10}{400 + 300} = \frac{30}{700} = \frac{3}{70}$$

$$\hat{Q} = 1 - \frac{3}{70} = \frac{67}{70}$$

i. **The Null hypothesis** H_0: $P_1 = P_2$ i.e. there is no significance difference between proportion of defective item before & after over hauling

ii. **The alternative hypothesis** H_1: $P_1 > P_2$ i.e. the quality of machine is improved.

iii. The level of significance α = 5% = 0.05

iv. **Test statistics**: under H_0: $P_1 = P_2$, the test statistics is

$$Z = \frac{p_1 - p_2}{\sqrt{PQ\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx N(0,1)$$

$$Z = \frac{0.05 - 0.033}{\sqrt{0.043 \times 0.957\left(\frac{1}{100} + \frac{1}{300}\right)}} = 1.103$$

v. **Critical value:** the critical value of Z at $\alpha = 5\%$ for one tailed (right tailed test)in $Z\alpha = Z_{0.05} = 1.645$

vi. **Conclusion:** Here $|Z| = 1.103$ is less than that of $Z_{0.05} = 1.645$
We accept null hypothesis $H_0 : P_1 = P_2$ & we conclude that the quality of machine is improved.

Example 13

A candidate for election made a speech in city A but not in B. A sample of 500 voters from city A showed that 59.6% of the voters were in favour of him. Whereas a sample of 300 voters per city B showed that 50% of the voter favoured him. Find the 95% confidence limit for difference of proportion in city A and B.

Solution

For city A $n_1 = 500$, $p_1 = 59.6\% = 0.596$.

For city B $n_2 = 300$, $p_2 = 50\% = 0.50$

$$S.E\,(p_1\text{-}p_2) = \sqrt{\frac{P_1Q_1}{n_1} + \frac{P_2Q_2}{n_2}} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$$

$$= \sqrt{\frac{0.596 \times 0.404}{500} + \frac{0.5 \times 0.5}{300}} = 0.0362$$

The 95% confidence interval for difference of population proportion is

$p_1 - p_2 \pm 1.96\ S.E.(p_1\text{-}p_2)$

$(0.596 - 0.5) \pm 1.96 \times 0.0362$

0.096 ± 0.0710509

$\therefore$ Lower limit of population proportion is 0.02495 = 2.49%

$\therefore$ Upper limit of population proportion is 0.16706 = 16.70%

Thus, there is 2.49% to 16.7% difference in the city A and B at confidence level 95%

EXERCISE – 13

1. Define standard error and what the utility of standard error are.
2. What do you understand by statistical hypothesis? What do you know about the testing of statistical hypothesis?
3. Define null hypothesis alternative hypothesis. Write about the error in sampling, what you know about type I and type II error.
4. What is the power of test and level of significance in testing of statistical hypothesis?
5. What is critical region or region of rejection? Differentiate between one tailed and two tailed test.
6. Write down the Procedures for testing of hypothesis.
7. What are the steps for test of significance of single mean for large sample?
8. What do you know the hypothesis testing using the binomial distribution and write down the steps for this test.
9. Write down the steps of test of significance of difference of mean for large sample.

Numerical for single mean

10. The breaking strength of cable produced by a manufactures have a mean of 1800 pounds and standard deviation of 100 lbs. by a new technique in the manufacturing process, it is claimed that the breaking strength can be increased. To test this claim, a sample of 50 cables is tested and it is found that the mean breaking strength is 1850 lbs. can we support the claim at the 0.01 level of significance.
11. Dairy Development Corporation claims that the mean volume of its milk packets is 500ml. A random sample of 50 milk packets gave the mean volume of 490 ml with a standard deviation of 10.89. Test the Corporation's claim at 5% level of significance.
12. An educator claims that the average IQ of city college students is at most 110, and that in a study made to test these claim 150 city college students, selected at random, had an average IQ of 111.2 with a standard deviation of 7.2. Use a level of significance of 0.01 to test the claim of the educator.
13. The average commission charged by full service brokerage firms on a sale of common stock is Rs. 144, with a standard deviation of Rs. 52. A random sample of 121 traders is taken and determined that they paid an average commission of Rs. 151. Can it be conclude that the client's commission is higher than the industry average?
14. The quality control engineer at a light bulb factory needs to estimate the average life of a large shipment of light bulbs. The process standard deviation is known to be 100 hours. A random sample of 64 light bulbs indicated a sample average life of 350 hours. Calculate the standard error

found that the mean breaking strength is 1850 lbs. can we support the claim at the 0.01 level of significance.

11. Dairy Development Corporation claims that the mean volume of its milk packets is 500ml. A random sample of 50 milk packets gave the mean volume of 490 ml with a standard deviation of 10.89. Test the Corporation's claim at 5% level of significance.

12. An educator claims that the average IQ of city college students is at most 110, and that in a study made to test these claim 150 city college students, selected at random, had an average IQ of 111.2 with a standard deviation of 7.2. Use a level of significance of 0.01 to test the claim of the educator.

13. A company claims that its light bulbs are superior to those of a competitor on the basis of a study which showed that a sample of 40 of its bulbs had an average life time of 628 hours of continuous use with a standard deviation of 27 hours. While a sample of 30 bulbs made by the competitor had an average life time of 619 hours of continuous use with a standard deviation of 25 hours. Test at 5% level of significance, whether this claim is justified.

14. A random sample of boots worn by 40 combat soldiers in a Himalayan region showed an average life of 1.08 years with a standard deviation of 0.05 years. Under the standard conditions the boots are known to have an average life of 1.28 years. Is there reason to assert at a level of significance of 0.05 that use in the Himalayan cause the mean life of such boots to decrease ?

15. If the standard deviation of pulse rate in adults of the age group 35–40 is 9.9 and the normal average pulse rate is 69, would you conclude a significant difference at 2% level of significance if a group of 81 people of the same age group suffering from a disease were found to have average pulse rate of 74 ?

16. The quality control engineer at a light bulb factory needs to estimate the average life of a large shipment of light bulbs. The process standard deviation is known to be 100 hours. A random sample of 64 light bulbs indicated a sample average life of 350 hours. Calculate the standard error of mean, set up 95 % and 99 % confidence interval estimate of the true average life of light bulbs in this shipment.

17. In a random sample of 64 of 600 road crossing in a town, the mean number of automobile accidents per year was found to be 4.2 and the sample standard deviation was 0.8. Construct a 95% confidence interval for the mean number of automobile accidents per crossing per year.

18. A ketchup manufactures is in the process of deciding whether to produce new extra spicy brand. The company's marketing research department uses a national telephone survey of 6,000 households and found that the extra-spicy ketchup would be purchased by 335 of them. A more extensive study made two years ago showed that 5% of he households would purchase the brand then. At a 2% significance level. Should the company conclude that there is an increase interest in the extra-spicy flavor?

19. A manager of a certain bank claims that at least 90% of the customers are satisfied with our new credit policy. A sample of 200 customers was taken and out of these 160 showed their satisfaction to the new credit policy. Determine if the manager's claim is legitimate at 1% level of significance.

20. A random sample of 500 apples is taken from a large consignment and 65 were found to be bad. Show that the S.E. of proportion of bad ones in a sample of this size is 0.015 and deduce that the percentage of bad apples in the consignment almost certainly lies between 8.5 and 17.5.

21. In a random sample of 300 households in a city 223 have computer. Set approximate 95% and 99% limits to the true value of proportion of households with computer in the whole city.

22. In a random sample of 400 items from a large consignment, 20 items were found to be defective. Estimate the standard error of the proportion of defective items and find 95% and 99% confidence interval for the percentage of defectives in the consignment.

23. In a survey of buying habits, 400 women shoppers are chose at random in super market 'A' located in Kathmandu, their average weekly food expenditure is Rs. 250 with a standard deviation of Rs. 40. For 400 women shoppers chosen at random in supermarket 'B' in Lalitpur, the average weekly food expenditure is Rs. 220 with a standard deviation of Rs. 55. Test at 1% level of significance whether the average weekly expenditure of the two populations of shoppers are equal.

24. India's Theaters knows that a certain hit movie ran an average of 84 days in each city and the corresponding standard deviation was 10 days. The manager of certain company in Bombay was interested in comparing the movie's popularity in his region with that in all of India's other theaters. He randomly chooses 75 theaters in his region and found that they ran the movie an average of 81.5 days. Test whether there was a significant difference in the length of Picture's run between theaters in the Bombay and all of India's others theaters at 1% level of significance.

25. The average hourly wage of a sample of 150 workers in Pokhara was Rs. 25.6 with a standard deviation of 10.8. The average hourly wage of a sample of 200 workers in Biratnagar was Rs. 28.7 with a standard deviation of Rs 12.8. Can an applicant safely assume that the hourly wage paid in Biratnagar is higher than those paid in Pokhara.

26. The mean height of 50 male students who showed above average participation in college athletics was 68.2 inches with a standard deviation. of 2.5 inches. While 50 male students who showed no interest in such participation had a mean height of 67.5 inches with a standard deviation of 2.8 inches. Test the hypothesis that male students who participate in college athletics are taller than other male students.

27. A sample of 100 electric light bulbs produced by manufacturer A showod a mean lifetime of 1190 hrs and a standard deviation of 90 hrs. A sample

of 75 bulbs produced by manufacturer B showed a mean of 1220 and a standard deviation of 80 hrs. Is there a significant difference between the mean lifetimes of the two brands of bulbs at 5% level of significance?

28. On an elementary school examination in spelling, the mean grade of 32 boys was 72 with a standard deviation of 8, while the mean grade of 30 girls was 75 with a standard deviation of 6. Test the hypothesis that the girls are better in spelling than the boys.

29. The average commission charged by full service brokerage firms on a sale of common stock is Rs. 144, with a standard deviation of Rs. 52. A random sample of 121 traders is taken and determined that they paid an average commission of Rs. 151. Can it be conclude that the client's commission is higher than the industry average?

30. The mean yield of wheat from Lalitpur district was 210 lbs with a standard deviation of 10 lbs. per acre from a sample of 100 plots. In Kathmandu district, the mean yield was 200 lbs with S.D. 12 lbs from a sample of 150 plots. Assume that the standard deviation of yield in the entire Bagmati zone was 11 lbs. Test whether there is any significantly difference between the mean yields of crops in the two districts.

31. In a certain factory, there are two independent processes manufacturing the same item. The average weight in a sample of 250 items produced from one process is found to be 120 gram with a standard deviation of 12 gram, while the corresponding figures in a sample of 400 items from the other process are 124 and 14. Find the standard error of the difference of means and also test whether the two mean weights differ significantly or not at 10 percent level of significance.

32. A whole sale dealer wanted to buy a large quantity of light bulbs from two brands label A and B. He bought 100 bulbs from each bulbs brand and found by testing that brand A had mean life time 1120 hours and standard deviation 75 hours and brand B had mean life time 1062 hours and standard deviation 82 hours. Find the 99% confidence limits for the difference in the average life of bulbs from the two brands.

33. Two groups, A and B, consist of 100 people each who have a disease. A serum is given to group A but not to group B (which is called the control), otherwise, the two groups are treated identically. It is found that in groups A and B 75 and 65 people respectively recover from the disease. At 1% level of significance, test the hypothesis that the serum helps cure the disease.

34. A sample pole of 300 voters from district A and 200 voters from district B showed that 56% and 48% respectively were in favour of a given

candidate. At 5% level of significance, test the hypothesis that there is a difference between the districts.

35. A random sample of 200 bolts manufactured by machine A and of 100 bots manufactured by machine B showed 19 and 5 defective bolts respectively. Test the hypothesis that machine B is performing better than A at 5% level of significance.

36. Two urns A and B, contain equal number of marbles, but the proportions of red and white marble in each urns in unknown. A sample of 50 marbles selected with replacement from each of the urns revealed 32 red marbles from A and 23 red marbles from B. using a 5% level of significance test the hypothesis that A has a greater proportion of red marbles than does B.

37. In Kathmandu city out of a random sample of 700 peoples 420 were found below poverty level while in Biratnagar out of 500 peoples 325 were found below poverty level. Is there any significant difference in the poverty level of the people in the two cities?

38. It is claimed that both tea and coffee are equally popular in Kathmandu district. If in a random sample of 600 persons, 310 were regular consumers of tea, is the claim justified at 1% level of significance?

39. Consider the case of Pharmaceutical Manufacturing Company testing two new compounds intended to reduce blood-pressure level. The compounds are administered two different sets of laboratory animals. In groups A, 71 of 100 animal tested response to drug first with lower blood-pressure levels. In group B, 58 of 90 animals tested respond to drug second with lower blood pressure level. The company wants to test at 0.05 levels whether there is a difference between the efficiencies of these two drugs.

40. The following results were obtained when two sets of items were subjected to two different treatments X and Y to enhance their tensile strength. Treatment X was applied on 400 items and 80 were found to have gained in strength. Treatment Y was applied on 400 items and 20 were found to have gained in strength. Is treatment Y superior to treatment X.

41. Random sample of 400 villagers and 900 municipal people were asked whether they would like certain TV program, 250 villagers and 540 municipal were in favour of the proposal. Set up the 95% and 99% confidence limits for the difference of proportion in whole country.

42. An student of a college select a random sample of 90 male and 80 female from a certain village and asked whether they would smoke or not, 63 male and 28 female said they smoked. Set up the 95% and 90% confidence interval for difference of Population proportion of smoker in that particular village.

43. A sample of 100 engineers from small construction companies and a sample of 100 engineers from large construction companies were interviewed about having expense controls in place for use of cellular phones. Of the small companies, 47 had controls in place whereas 79 of he large companies had controls. Determine the 95% confidence in population proportion in expense controls in place for large and small companies' engineers.

Answers:

10. Z = 3.55, Significant
11. Z = - 6.49, Significant
12. Z = 2.04, Not Significant
13. Z = 1.44, Not Significant
14. Z = – 25.298, Significant
15. Z = 4.55, Significant
16. 12.5, (325.5, 374.5), (317.75, 382.25)
17 4.01, 4.39
18. Z = 2.07, Significant
19. Z = 1.54, not Significant
21. (0.694, 0.792), (0.6785, 0.80745)
22. 0.011, (2.86%, 7.14%), (2.19%, 7.81%)
23. Z = 8.82, Significant
24. |Z| = 2.161 Not Significant
25. Z = -2.46, Significant
26. Z = + 1.32, Not Significant
27. Z = 2.326, Yes
28. Z = -1.677, Significant
29. Z = 1.48, Not Significant
30. Z = -7.05, Significant
31. S.E. $(X_1 - X_2) = 1.032$; Z = – 3.87, Significant
32. 29.33, 86.67
33. Z = – 4.72, Significant
34. Z = 1.75, Not Significant
35. Z = 1.35, Not Significant
36. Z = 1.84, Significant
37. Z = -1.763 Not Significant
38. Z = 0.83, Not Significant
39. Z = 0.973 Not Significant
40. Z = 41.14, Significant
41. (0.0233, 0.0266) and (0.0227, 0.0272)
42. (0.20, 0.50) and (0.223, 0.476)
43. (0.186, 0.506}

Chapter 15

Exact Sampling Distribution

STUDENT 't' DISTRIBUTION

Student's t-distribution was given and used first by W.S. Gosset in 1908. The statistician W.S. Gosset is better known by his pseudonym student (Pen name) and later on developed and extended by Prof. R.A. Fisher (1926)

In the developments of large sample test for mean we discussed that the statistics $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx N(0,1)$

If the population variance is unknown then for large samples, its estimate provided by sample variance s^2 is used and normal test is applied. For small sample an unbiased estimate of population variance σ^2 is given by

$$S^2 = \frac{\Sigma(X - \bar{x})^2}{n-1} \text{ and } s^2 = \frac{\Sigma(X - \bar{x})^2}{n} \text{ which implies, } ns^2 = (n-1)s^2$$

It was customary to replace σ^2 by S^2 for small samples and then applies the normal test even for small sample. W.S. Gosset, who wrote under the pen name of student, obtained the sampling distribution of the statistics $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ for small sample and showed that it is for from normality. This discovers started a new field viz. "exact sampling tests' in the history of statistical inference.

When sample size n is less than 30, we consider sample is small and use t-test

Definition: Let $x_1, x_2 \ldots\ldots\ldots x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2. Then student 't' statistics is defined as.

$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ Or, $t = \frac{\bar{x} - \mu}{s/\sqrt{n-1}}$, Follows student t-distribution with (n-1) degrees of freedom and where $\bar{x} = \frac{\Sigma X}{n}$ and $S^2 = \frac{\Sigma(X - \bar{x})^2}{n-1}$ and $s^2 = \frac{\Sigma(X - \bar{x})^2}{n}$

And the probability density function of Statistics't' is defined as

$$f(t) = \frac{1}{\sqrt{\nu}} \frac{1}{\beta\left(\frac{1}{\nu}, \frac{\nu}{2}\right)(1 + t^2/\nu)^{\frac{\nu+1}{2}}} \; ; \; -\infty < t < \infty$$

And $\nu = (n - 1)$ degrees of freedom.

14.1.1 Properties of t-Distribution

The important properties of t-distribution are given as follows

i. Like Z-distribution, t-distribution is a continuous distribution having symmetrical and bell shaped curve. The value of 't' ranges from. $-\infty$ to ∞ i.e. $-\infty < t < \infty$

ii. We use the notation $t \approx t_n$ to show that the statistics't' follows student's t-distribution with n degrees of freedom.

iii. If $\nu = 1$ i.e. degrees of freedom =1we get

$$f(t) = \frac{1}{\beta\left(\frac{1}{2}, \frac{1}{2}\right)(1+t^2)} = \frac{1}{\pi}\frac{1}{(1+t^2)} \quad -\infty < t < \infty,$$

This is the pdf of standard Cauchy distribution. Hence when degrees of freedom is one, Student's t-distribution reduces to Cauchy distribution.

iv. As $n \to \infty$ the t-distribution tends to standard normal distribution.

v. Mean of the t-distribution is zero and variance is equal to $\frac{\nu}{\nu - 2}$; $\nu > 2$

vi. The t-distribution is symmetrical i.e. $\beta_1 = 0 \Rightarrow \lambda_1 = 0$.

vii. For t-distribution $\beta_2 = 3\left(\frac{\nu - 2}{\nu - 4}\right)$, for large degrees of freedom $\beta_2 = 3$, hence for large degrees of freedom t-distribution tends to normal distribution.

viii. The t-distribution curve is unimodal with mean = mode = median = 0.

ix. The t-axis is the asymptote of the t-distribution.

x. The t-distribution is flatter than the normal distribution and there is a different t-distribution for every possible sample size. For example a t-distribution for a sample size of 15 is different than the t-distribution for a sample size of 2 and so on.

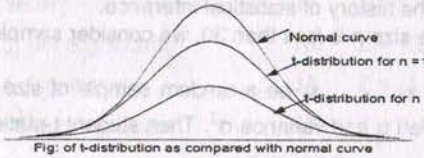


Fig: of t-distribution as compared with normal curve

xi. The t-distribution can be used even in case of large sample but the large sample theory can't be used for small sample.

xii. The very interesting property of the sampling distribution of t is that it does not depend on the population parameter and depends only on degrees of freedom $(\nu) = (n - 1)$ i.e. on the sample size n.

14.1.2 Degrees of Freedom

The number of independent observations in a set is called degrees of freedom (d.f.)

In other words, degrees of freedom may be defined as the number of observations in a set minus the number of restrictions imposed on it. It is denoted by

ν(nu.) or d.f with the help of degree of freedom, we know something about samples. For example, we are selected two samples values x_1 and x_2 such that their mean is 30 then $\frac{x_1 + x_2}{2} = 30$. How can we find what value of x_1 and x_2 can take on in this situation?

The answer is that x_1 and x_2 can be any two values whose sum is 60 because $60 \div 2 = 30$.

Suppose x_1 has value of 20 then x_2 is no longer free to take on any value but must have value of 40 i.e. if $x_1 = 20$ then $\frac{x_1 + x_2}{2} = 30$ or $20 + x_2 = 60$ then $x_1 = 40$.

This example shows that when there are two elements in a simple and we know the sample mean of these two elements; we are free to specify only one of the elements to find the other elements. So, we can say that there is one degree of freedom.

So with two sample values, number of degree of freedom = 2 -1 = 1.

If we have three sample values, number of degree of freedom = 3 -1 = 2.

If we have four sample values, number of degree of freedom = 4 -1 = 3.

In general for n sample values, number of degree of freedom = n -1 .

For t-distribution in estimating population mean, the number of degree of freedom = n-1 where n is the sample size and for estimating correlation coefficient and regression coefficient the degrees of freedom is (n – 2) here we are given sum of variable x and y respectively.

14.1.3 Assumption about t-Distribution

For derivation of student's t-distribution following basic assumption are made

i. The parent population from which the samples are drawn is normal with mean μ and variance σ^2.

ii. All observations in the sample are independent i.e. one item selected in the sample doesn't effect the other items included in the sample.

iii. The sample size is small i.e. less than 30 as a usual practice. Also the sample should not certain less tan 5 observations.

iv. The hypothetical value μ_0 of μ is a correct value of population mean of parent population form which the sample are drawn.

v. The sample values are correctly measured and recorded.

vi. The value of population standard deviation σ is unknown.

14.1.4 Critical Value of t-Test

The critical (or significance) value of t at α level of significance and υ degree of freedom for two tailed test are given by the equation $P(|t| > t_\nu(\alpha)) = \alpha$ and $P\{|t| > t_\nu(\alpha)\} = 1 - \alpha$

The value $t_\nu(\alpha)$ have been tabulated in t-table for different values of α and ν and are given at the end of this book

Since t-distribution is symmetric about t =0 we get

$P(t > t_\nu(\alpha)) + P[t_\nu - t_\nu(\infty)] = \alpha$

$P(t > t_\nu(\alpha)) = \frac{\alpha}{2} \therefore P(t > t_\nu(2\alpha)) = \alpha$

Then $t_\nu(2\alpha)$) from the table at the end of this book gives the significance value of 't' for a single tail test (right tail or left tail, since the distribution is symmetrical curve) at level of significance α and ν degree of freedom.

14.1.5 Application of t-Distribution

The student's t-distribution has wide number of application in statistics, same of which are enumerated below

i. To test it the sample mean ($\bar{x}$) differ significantly from the hypothetical value μ of the population mean, Population variance being unknown.
ii. To test the significance of the difference between two independent mean, The population variance being equal but unknown.
iii. Paired t-test for difference of two means.
iv. Test of significance of an observed sample correlation coefficient
v. Test of significance about the slope parameter β_0 of regression equation
vi. Test of significance about the regression coefficient β_{yx}.

14.2 SMALL SAMPLE TEST FOR SINGLE MEAN

We can use the student t-test under the following two cases

a. If random samples $x_1, x_2, \ldots, x_n$ of size n has been drawn from a normal population with a specified mean say μ_0
b. if the sample mean differ significantly from the hypothetical value μ_0 of the population mean.

Following are the steps for test of significance of a single mean in small sample (i.e. for t-test for small single mean)

i. **Set up the null hypothesis H_0:** $\mu = \mu_0$ i.e. the sample has been drawn from the population with mean μ_0 or there is no significant difference between the sample mean $\bar{x}$ and the populations mean μ_0.
ii. **Set up Alternative hypothesis** $H_1 : \mu \neq \mu_0$ (Two tailed)
or, $H_1 : \mu > \mu_0$ (right tailed)
or, $H_1 : \mu < \mu_0$ (left tailed)
i.e. there is significance difference between the sample mean and population mean
iii. Fix the level of significance α
iv. **Test statistics**: Under the null hypothesis $H_0 : \mu = \mu_0$ the test statistics

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$, follows student's t-distribution with (n-1) d.f.

Where $\bar{x} = \frac{\Sigma x}{n}$ in sample mean and

$$S^2 = \frac{\Sigma(X - \bar{x})^2}{n-1} = \left[\frac{\Sigma(x^2)}{n-1} - \frac{(\Sigma x)^2}{n(n-1)}\right]$$ and

$$s^2 = \frac{\Sigma(X - \bar{x})^2}{n} = \frac{\Sigma x^2}{n} - \left(\frac{\Sigma x}{n}\right)^2$$

v. **Critical value**: For α level of significance and (n-1) d.f. find the critical value of 't' from student t-table as $t(\alpha)$.
vi. **Conclusion:** If calculated value of $|t| <$ tabulated value t_α we accept our null hypothesis H_0 at α level and If $|t| \geq t_\alpha$ we reject null hypothesis H_0 and accept alternative hypothesis H_1

Example 1

A machinist is making engine parts with axle diameter of 0.7 inch. A random sample of 10 parts shows a mean diameter of 0.742 inch with a standard deviation of 0.04 inches. Tests whether sample are dram from the population or note

Solution

Here we are given

Population mean = 0.700 inches.

Sample mean $\bar{x}$ = 0.742 inch.

Sample s.d. (s) = 0.040 inch & sample size n = 10

Here sample size n < 30 & population s.d. unknown. Therefore we use student t-test and follow the following steps.

i. **The Null hypothesis H_0:** μ = 0.700 inches i.e. the product is confirming to specification.
ii. **The Alternative hypothesis H_1:** $\mu \neq$ 0.700 i.e. the product is doesnot confirming to specification.
iii. Let level of significance α = 5%
iv. **Test Statistics:** Under, the null hypothesis H_0: μ = 0.700, the test statistics is

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n-1}}}$$, follows student's t- distribution with n-1degrees of freedom

$$= \frac{0.742 - 0.700}{\frac{0.046}{\sqrt{10-1}}} = 3.15$$

$\therefore |t| = 3.15$

v. **Critical value**: Here, for α = 5% level of significance and (n - 1) = 10 - 1 = 9 d.f. the critical value of t is $t_{0.05}(9) = 1.833$

vi. **Conclusion**: Here the calculated value of $|t| = 3.15$ is greater than $t_{0.05}(9) = 1.833$, we reject our null hypothesis H_0 & accept the alternative hypothesis H_1 and thus we conclude that the product is not meet the specification.

14.3 Confidence Limit for Estimating Population Means

The approximate $(1 - \alpha)100\%$ confidence limit to estimate the population mean μ for small sample also known as fiducial limits is given by the following relation.

$$\bar{x} \pm t_v(\alpha)\frac{S}{\sqrt{n}} \text{ or, } \bar{x} \pm t_v(\alpha)\frac{s}{\sqrt{n-1}}$$

i.e. $P(\bar{x} - t_v(\alpha)\frac{S}{\sqrt{n}} < \mu < \bar{x} + t_v(\alpha)\frac{S}{\sqrt{n}}) = 1 - \alpha$

Thus, the approximate 95% & 99% confidence or fiducial limits

$\bar{x} \pm t_v(0.05)\frac{S}{\sqrt{n}}$ and $\bar{x} \pm t_v(0.01)\frac{S}{\sqrt{n}}$ respectively.

Example 2

The heights of 10 males of a given locality are found to be 70, 67, 62, 68, 61, 68, 70, 64, 64, 66 inch is it reasonable to believe that the average height is greater than 64 inch ? Test at 5% level significance.

Solution

To test the believe that the average height is greater than 64 inches from small sample we use student t-test as follows

i. **The Null hypothesis H_0 : $\mu = 64$ inches**
ii. **The Alternative hypothesis H_1: $\mu > 64$ inch** (right tailed test)
iii. Level of significance $\alpha = 5\%$
iv. **Test Statistics**: Under the null hypothesis H_0: $\mu = 64$ the test statistics is

$$t = \frac{\bar{x} - \mu}{\frac{S}{\sqrt{n}}} \approx t(n-1)$$

Calculation of sample mean $\bar{x}$ and sample s.d.

x	70	67	62	68	61	68	70	64	64	66	**660**
$x-\bar{x}$	4	1	-4	2	-5	2	4	-2	-2	0	**0**
$(x-\bar{x})^2$	16	1	16	4	25	4	16	4	4	0	**90**

Sample mean $\bar{x} = \frac{\Sigma x}{n} = \frac{660}{10} = 66$

$S^2 = \frac{1}{n-1}\Sigma(x - \bar{x})^2 = \frac{90}{9} = 10$

$\therefore t = \frac{\bar{x} - \mu}{\sqrt{\frac{S^2}{n}}} = \frac{66 - 64}{\sqrt{\frac{10}{10}}} = 2$

v. **Critical value**: Here for $\alpha = 5\%$ and $(n-1) = 9$ d.f the critical value of t for single tailed test is $t_9(0.05) = 1.833$.
vi. **Conclusion**: Here $|t| = 2 > t_9(0.05) = 1.833$, we reject the null hypothesis H_0 and accept the alternative hypothesis H_1
Thus we conclude that the average height in greater than 64 inches.

Example 3

The average breaking strength of steel rods in specified to be 18.5 thousand pound. To test this sample of 14 rods was tested. The mean and standard deviation obtained were 17.85 & 1.955. Obtained the 95% confidence interval for true mean diameter of a rod

Solution

We are given population mean $\mu = 18.5$

$\bar{x} = 17.85$, sample standard deviation (s) = 1.955 and sample size n =14

$1 - \alpha = 95\%$ which implies $\alpha = 5$

$\therefore$ The approximate 95% confidence limits are given by

$\bar{x} \pm t_v(\alpha)\frac{s}{\sqrt{n-1}}$ where $t_v(\alpha) = t_{13}(0.05) = 1.771$,(from table)

$17.85 \pm 1.771 \times \frac{1.955}{\sqrt{13}}$

$\therefore$ Lower limit 17.85 - 0.96 = 18.81
Upper limit 17.85 + 0.96 = 16.89

14.43.1 Assumptions for t- Test for Difference of Means

Following basic assumptions are considered to test of significance for difference of means for small samples.

i. Parent population from which samples have been drawn are normally distributed
ii. The two sample are random and independent of each other.
iii. Population variances are equal and unknown i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$ but unknown
iv. The sample values are correctly measured and recorded.
v. The hypothetical values μ_1 & μ_2 are correct values of population means of two populations.

14.4 TEST OF SIGNIFICANCE OF DIFFERENCE OF MEANS

Suppose we want to test if two independent small samples have been drawn from two normal populations having the same means and the population variances being equal but unknown in such a case we use t-test for difference of means.

Let $x_1, x_2 \ldots x_{n_1}$ & $y_1, y_2 \ldots y_{n_2}$ be two independent random samples from the given normal population when sample size of each sample in less than 30 then the test statistics for test of significance of difference mean is

$t = \frac{(\bar{x} - \bar{y}) - E(\bar{x} - \bar{y})}{S.E.(\bar{x} - \bar{y})}$, follow student's t-distribution with $(n_1 + n_2 - 2)$ d.f.

Where $\bar{x} = \frac{\sum x}{n_1}$ $\bar{y} = \frac{\sum y}{n_2}$

$S.E.(\bar{x} - \bar{y}) = S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ And $S^2 = \frac{\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2}{n_1 + n_2 - 2}$

We can use the following steps for test of significance for different of mean in small sample.

i. **The Null hypothesis H_0 :** $\mu_1 = \mu_2$ i.e. sample has been drawn from the normal population with same mean.

ii. **The Alternative hypothesis** H_1: $\mu_1 \neq \mu_2$ (Two tailed)
or, H_1: $\mu_1 > \mu_2$ (right tailed test)
or, H_1: $\mu_1 < \mu_2$ (left tailed test) i.e. sample are drawn from different normal population.

iii. Fix the level of significance α.

iv. **Test Statistics:** Under the null hypothesis H_0: $\mu_1 = \mu_2$ & the assumption that $\sigma_1^2 = \sigma^2_2 = \sigma^2$ two population variance are equal but unknown the test statistics is

$$t = \frac{\bar{x} - \bar{y}}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{n_1 + n_2 - 2}$$

Also when sample variances s_1^2 and s_2^2 are known than value of S^2 is calculate by using the formula $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}$

v. **Critical value:** For α level of significance and $(n_1 + n_2 - 2)$ d.f. find the critical value of 't' from table.

vi. **Conclusion:** If $t_{cal} < t_{tab}$ we accepts the null hypothesis H_0 at α level of significant and If $t_{cal} \geq t_{tab}$ we reject the null hypothesis H_0 and accept the alternative hypothesis H_1.

14.4.1 Confidence Limit for Difference of Population Mean

To calculate the approximate 100(1-α)% confidence limit for difference of population mean from small sample we use the following relation.

$(\bar{x} - \bar{y}) \pm t_{tab}\, S.E.(\bar{x} - \bar{y})$

$(\bar{x} - \bar{y}) \pm t_{tab}\, S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

Thus, 95% and 99% confidence limit for different of population mean in small sample is given as

$(\bar{x} - \bar{y}) \pm t_{\alpha}(0.05)\, S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, for 95 % confidence limit

$(\bar{x} - \bar{y}) \pm t_{\alpha}(0.01)\, S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, for 99 % confidence limit

Example 4

We have the following results from the two samples drawn from normal populations with same variance but unknown as

$n_1 = 6$ $\bar{x}_1 = 25$ $s_1^2 = 36$

$n_2 = 8$ $\bar{x}_2 = 20$ $s_2^2 = 25$

Whether the two samples are drawn from the normal populations with equal means?

Solution

To test whether the samples are drawn from the normal populations with equal means we use student t-test for difference of mean as follows

i. **The Null Hypothesis** is $H_0 : \mu_1 = \mu_2$ i.e. two samples are drawn from the normal populations with equal means.

ii. **The Alternative Hypothesis is** H_1: $\mu_1 \neq \mu_2$ (two tailed test) i.e. two samples are not drawn from the normal populations with equal means.

iii. Level of significance $\alpha = 5\%$

iv. **Test Statistics:** Under the null Hypothesis H_0 : $\mu_1 = \mu_2$ the test statistics is

$$t = \frac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \approx t_{n_1 + n_2 - 2}$$

Here, $S^2 = \frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \frac{6 \times 36 + 8 \times 25}{6 + 8 - 2} = 34.67$

$$\therefore t = \frac{25 - 20}{\sqrt{34.67\left(\frac{1}{6} + \frac{1}{8}\right)}} = 1.57$$

v. **Critical value:** Here critical value of 't' for $\alpha = 5\%$ and $(n_1 + n_2 - 2) = 12$ d.f. is $t_{Tab} = 2.18$

vi. **Conclusion**: Here $t_{cal} = 1.57$ is less than $t_{Tab} = 2.18$, we Accept the null hypothesis H0 : $\mu_1 \neq \mu_2$ and concluded that two samples are drawn from the normal populations with equal means.

Example 5

The nicotine content in milligrams of two samples of tobacco were found to be as follows

Sample A:	24	27	26	21	25	
Sample B:	27	30	28	31	22	36

Can it be said that two samples come from normal population having the same mean?

Solution

To test whether the samples are drawn from the normal populations with equal means we use student t-test for difference of mean as follows

i. **The null hypothesis** H_0: $\mu_1 = \mu_2$ i.e. the two samples have been drawn from the normal population with the same mean.

ii. **The Alternative hypothesis** H_1: $\mu_1 \neq \mu_2$ i.e. the two samples have been drawn from the normal population with different mean.

iii. Let the level of significance is $\alpha = 5\%$

iv. Test Statistics: Under the null hypothesis H_0: $\mu_1 = \mu_2$, the test statistics is $t = \dfrac{\bar{x}_1 - \bar{x}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ follows student t-distribution with $(n_1 + n_2 - 2)$ degrees of freedom

where $S^2 = \dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2} = \dfrac{\sum(x - \bar{x})^2 + \sum(y - \bar{y})^2}{(n_1 + n_2 - 2)}$

Calculation table

X	$x - \bar{x}$	$(x-\bar{x})^2$	Y	$(y-\bar{y})$	$(y-\bar{y})^2$
24	-0.6	0.36	27	-2	4
27	2.4	5.76	30	1	1
26	1.4	1.96	28	-1	1
21	-3.6	12.96	31	2	4
25	0.4	0.16	22	-7	49
			36	7	49
Total		21.2			108

Here $s^2 = \dfrac{21.2 + 108}{9} = 14.35$

$\therefore s = 3.788$

$\therefore t = \dfrac{24.6 - 29}{3.788\sqrt{\frac{1}{5} + \frac{1}{6}}} = 1.92,$

v. **Critical value:** Here critical value of 't' for $\alpha = 5\%$ and $(n_1 + n_2 - 2) = (5 + 6 - 2) = 9$ d.f. is $t_{Tab} = 2.262$

vi. **Conclusion**: Here $t_{cal} = 1.92$ is less than $t_{Tab} = 2.262$, we Accept the null hypothesis H0 : $\mu_1 \neq \mu_2$ and concluded that two samples are drawn from the normal populations with equal means.

14.5 PAIRED t-TEST FOR DIFFERENCE OF MEANS

In the t-test for difference of means, the two samples were independent of each other. However, there are many situation where the samples are *pair wise dependent* to each other. for example, if we are testing sales of goods after & before advertisement; a testing the productivity level of workers before & after a training program or checking memory capacity of person before & after training and so on. Then these values are related to each other. In such situation, we are concerned with

the difference between the pair of related observations instead of the value of the individual observation. Paired t-test for difference of means can be applied under the following situations.

i. The sample sizes are equal i.e. $n_1 = n_2 = n$ (say).

ii. The two samples are not independent but the sample observations are paired together i.e. the pair observations (x_i, y_i) i = 1, 2 ... n) corresponds to the same sample unit.

iii. The same set of sample is treating twice on the same subject matter.

The following are the steps in testing paired t-test for difference of mean or **t-test for paired wise dependent samples** is as follows

i. **Set up the null hypothesis** H_0: $\mu_1 = \mu_2$ or i.e. there is no significance difference in the observation before and after treatment.

ii. **Set up the Alternative Hypothesis** H_1: $\mu_1 \neq \mu_2$ (Two tailed test) i.e. there is significance difference in the observation before and after treatment

or H_1: $\mu_1 > \mu_2$ (Right tailed test) The treatment is not effective

or, H_1: $\mu_1 < \mu_2$ (Left tailed test) The treatment is effective

iii. Fixed the level of significant α

iv. **Test statistics:** under H_0: $\mu_1 = \mu_2$ the test statistics is

$t = \dfrac{\bar{d}}{\frac{S}{\sqrt{n}}}$, follows student t=distribution with (n-1) degrees of freedom

Where d = x - y difference between two set of observations.

$\bar{d} = \dfrac{\sum d}{n}$ and $S^2 = \dfrac{\sum(d - \bar{d})^2}{n-1} = \dfrac{\sum d^2}{n-1} - \dfrac{(\sum d)^2}{n(n-1)}$

v. **Critical value:** For α level of significant and (n -1) degrees of freedom find the critical value of 't' as $t_v(\alpha)$

vi. **Conclusion:** if calculated value of 't' as t_{cal} is less than tabulated value of 't' as $t_v(\alpha)$ We accept the null hypothesis, otherwise We reject null hypothesis at α level of significance and accept alternative hypothesis.

Example 6

Memory capacity of 10 students was tested before and after training state, whether the training was effective or not from the following score.

Before	12	14	11	8	7	10	3	0	5	6
After	15	16	10	7	5	12	10	2	3	10

Solution

To test the significance difference of mean i.e. whether the training was effective or not we proceed as follows

i. **The null hypothesis** H_0: $\mu_1 = \mu_2$ i.e the average memory capacity of students before and after training is same in other word training was not effective.

ii. The Alternative hypothesis H_1: $\mu_1 < \mu_2$ i.e. the average memory capacity of students after training is increased; in other word training was effective.

iii. Let the level of significance is $\alpha = 5\%$

iv. **Test Statistics:** Under null hypothesis H_0: $\mu_1 \neq \mu_2$. The test statistics is

$$t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}}$$, follows student t-distribution with (n-1) degrees of freedom

Where d = x - y difference between two set of observations.

$$\bar{d} = \frac{\sum d}{n} \text{ and } S^2 = \frac{\sum (d - \bar{d})^2}{n-1} = \frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}$$

Calculation table

X	Y	d=X-Y	d^2
12	15	- 3	9
14	16	- 2	4
11	10	1	1
8	7	1	1
7	5	2	4
10	12	- 2	4
3	10	- 7	49
0	2	- 2	4
5	3	2	4
6	10	- 4	16
Total		**- 14**	**96**

Thus $\bar{d} = \frac{\sum d}{n} = \frac{-14}{10} = -1.4$ and

$$S^2 = \frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)} = \frac{96}{9} - \frac{(-14)^2}{10 \times 9} = 8.49$$

$\therefore S = 2.913$

Now the test statistics is $t = \frac{\bar{d}}{S} \times \sqrt{n} = \frac{-1.4}{2.913} \times \sqrt{10} = -1.51$

v. **The critical value:** The critical value of 't' for 5% level of significance and (10-1) = 9 d.f. and for one tailed test is $t_{tab} = 1.833$

vi. **Conclusion:** Here tabulated value of $|t| = 1.51$ is less than tabulated value of $t_{tab} = 1.833$, we accept our null hypothesis
Thus concluded that the average memory capacity of students before and after training is same

Example 7

The measurement of diameter of 8 cylindrical rods by Vernier caliper (X) and Micrometer (Y) was as follows.

X	2.265	2.267	2.264	2.267	2.268	2.263	2.263	2.258
Y	2.270	2.268	2.269	2.273	2.270	2.270	2.268	2.268

Test whether the difference between measurements of diameter by Vernier and Micrometer is significant or not.

Solution

To test the significance difference of two paired mean i.e. the difference between measurements of diameter by Vernier and Micrometer is significant or not we proceed as follows

i. **The null hypothesis H_0:** $\mu_1 = \mu_2$ i.e. measurement of diameter by Vernier and Micrometer is not significant.

ii. **The Alternative hypothesis H_1:** $\mu_1 \neq \mu_2$ i.e. measurement of diameter by Vernier and Micrometer is significant.

iii. Let, the level of significance is $\alpha = 5\%$

iv. **Test Statistics:** Under the null hypothesis H_0: $\mu_1 = \mu_2$,

The test statistics is $t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}}$ where $\bar{d} = \frac{\sum d}{n}$ and

$$S^2 = \frac{\sum (d - \bar{d})^2}{n-1} = \frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)}$$

Calculation table

X	Y	d=(X-Y)	d^2
2.265	2.27	-0.005	0.000025
2.267	2.268	-0.001	0.000001
2.264	2.269	-0.005	0.000025
2.267	2.273	-0.006	0.000036
2.268	2.27	-0.002	0.000004
2.263	2.27	-0.007	0.000049
2.264	2.268	-0.004	0.000016
2.258	2.268	-0.010	0.000100
Total		0.04	0.000256

Thus $\bar{d} = \frac{\sum d}{n} = \frac{0.04}{8} = 0.005$ and

$$S^2 = \frac{\sum d^2}{n-1} - \frac{(\sum d)^2}{n(n-1)} = \frac{0.000256}{7} - \frac{(0.04)^2}{8 \times 7} = 0.000008$$

$\therefore$ The value of S = 0.0028

Now $t = \frac{\bar{d}}{\frac{S}{\sqrt{n}}} = \frac{0.005}{0.0028} \times \sqrt{8} = 5.05$

vi. **Critical value**: The critical value of 't' for $\alpha = 5\%$ level of significance and (8-1) = 7 degrees of freedom is $t_{tab} = 2.365$

vi. **Conclusion:** Here calculated value of $|t| = 5.05$ is much greater than tabulated value of $t_{tab} = 2.365$ we reject our null hypothesis and accept alternative hypothesis. Thus, we concluded that there is a significance difference between the reading made by Vernier and Micrometer at 5% level of significance.

14.6 TEST OF SIGNIFICANT OF AN OBSERVED SAMPLE CORRELATION COEFFICIENT

In order to test whether the sample correlation coefficient (r) is significant of any correlation between the variable in the population t-test for significant of an observed sample correlation coefficient is applied. The steps for testing of significance of an observed sample correlation coefficient are as follows.

i. **Set up null hypothesis H_0:** $\rho = 0$ i.e. the variable are uncorrelated in the population or population correlation coefficient is zero

ii. **Set up Alternative hypothesis H_1:** $\rho \neq 0$ i.e. the variable are correlated in the population or population correlation coefficient is not zero
   or, H_1: $\rho > 0$, (Right tailed test) i.e. there is the positive correlation in the population
   or, H_1: $\rho < 0$, (Left tailed test) i.e. there is the negative correlation in the population.

iii. Fix the level of significance α

iv. **Test statistics**: Under the null hypothesis $H_0 : \rho = 0$, the test statistics is
   $t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$, follows students t-distribution with (n -2) d.f.

v. **Critical value:** For α level of significance and (n-2) degrees of freedom find the critical value of t from table as t_{tab}.

vi. **Conclusion:** If calculated value of t as t_{cal} is less than tabulated value of t as t_{tab} we accept the null hypothesis H_0 at α level of significance and If $t_{cal} \geq t_{tab}$ we reject H_0 and accept alternative hypothesis H_1.

14.6.1 Confidence Limits for Population Correlation Coefficient (ρ)

The approximate $(1 - \alpha)$ 100% confidence limit for estimating population correlations coefficient (ρ) are given by the following relations

$r \pm t_\alpha(n-2)\ S.E.(r)$

$r \pm t_\alpha(n-2)\frac{1-r^2}{\sqrt{n}}$, Where, $S.E(r) = \frac{1-r^2}{\sqrt{n}}$

Example 8

A random sample of 27 pairs of observations from a normal population gives a correlation coefficient of 0.6. Is it likely that the variable in the population are

uncorrelated? Also compute 95% confidence limits for population correlation coefficient.

Solution

We are given sample size n = 27

And sample correlation coefficient (r) = 0.6

To test the significance of correlation coefficient we proceed in the following steps.

i. **Null hypothesis H_0:** $\rho = 0$ i.e. the variable in the population are uncorrelated.

ii. **Alternative hypothesis H_1:** $\rho \neq 0$, (two tailed) i.e. test variable in the population are correlated.

iii. Level of significance $\alpha = 5\ \%$

iv. **Test statistics**: under the null hypothesis H_0 the test statistics is

$$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} = \frac{0.6}{\sqrt{1-(0.6)^2}}\sqrt{27-2} = \frac{15}{4} = 3.75$$

v. **Critical value:** Degree of freedom = n - 2 = 27- 2 = 25

For $\alpha = 5\%$ and 25 d.f. for two tailed test critical value of as $t_{tab} = 2.060$

vi. **Conclusion:** Here $t_{cal} = 3.75$ is greater than the $t_{tab} = 2.060$ we reject H_0 & accept H_1 at 5% level. Then we conclude that the variable are correlated in the population.

Now, 95% confidence limits

$(1 - \alpha) = 95$ which implies $\alpha = 5\% = 0.05$

$r = 0.6$, d.f. = 25 $t_{0.05}(25) = 2.060$

$\therefore$ The approximate 95% confidence limits for estimating the population correlation coefficient is

$$r \pm t_{0.05}(25)\frac{1-r^2}{\sqrt{n}}$$

$$0.6 \pm 2.060 \times \frac{1 - 0.36}{\sqrt{27}}$$

0.6 ± 0.254

Upper limit = 0.6 + 0.254 = 0.854

Lower limit = 0.6 - 0.254 = 0.346

Thus the probability that the value of population regression coefficient lies between 0.346 to the 0.854 is 95%.

Symbolically $P(0.346 < \rho < 0.854) = 95\%$

14.7 TEST OF SIGNIFICANCE OF REGRESSION COEFFICIENT

We have the regression equation of Y on X is $Y = a + b_{YX}X$ and the population regression coefficient of Y on X as $Y = \beta_0 + \beta_{YX}\ X$, since there are two coefficient of regression we have two test of significance of regression coefficients as follows

i. Test of significance of Slope parameter β_0

ii. Test of significance of regression coefficient β_{YX}

14.7.1 Test of significance of Slope parameter β_0

To test the significance of slope parameter β_0 of regression line, we proceeded in the following stapes

i. **Set up null hypothesis H_0:** $\beta_0 = 0$ i.e. Slope parameter of regression Line in population is zero.

ii. **Set up alternative hypothesis H_1:** $\beta_0 \neq 0$ (two tailed) i.e. Slope parameter of regression Line in population is not equal to zero.

Or H_1: : $\beta_0 > 0$ (right tailed) i.e. Slope parameter of regression Line in population is greater than zero.

Or H_1: $\beta_0 > 0$ (left tailed) i.e. Slope parameter of regression Line in population is less than zero.

iii. Fix the level of significance α.

iv. **Test statistics;** Under the null hypothesis **H_0:** $\beta_0 = 0$, the test statistics v. for slope parameter β_0 is $t = \frac{a - \beta_0}{S_a}$,

follows student t-distribution with (n-2) degrees of freedom , where S_a is the standard error of an estimator a, and given as $S_a^2 = S_e^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum u^2}\right]$,

$S_e^2 = \frac{1}{(n-2)}[\sum v^2 - b_{YX}\sum uv]$, $u = x - \bar{x}$ and $v = y - \bar{y}$

v. **Critical value**: - For α level of significance and the (n-2) d.f. find the critical value of t as t_α from table.

vi. **Conclusion:** - If $|t| < t_\alpha$ we accept our null hypothesis at α level of significance and If $|t| \geq t_\alpha$ we reject our null hypothesis and accept alternative hypothesis.

Confidence limit for Slope parameter of regression line β_0

The $(1 - \alpha)$ 100% confidence limit for slope parameter of regression line β_0 is given as : $a \pm t_\alpha(n-2)\, S_a$

14.7.2 TEST OF SIGNIFICANCE OF REGRESSION COEFFICIENT β_{YX}

To test the significance of regression coefficient β_{YX}, we proceeded in the following stapes.

i. **Set up null hypothesis H_0:** $\beta_{YX} = 0$ i.e. regression coefficient in population is zero.

ii. **Set up alternative hypothesis H_1:** $\beta_{YX} \neq 0$ (two tailed) i.e. regression coefficient in population is not equal to zero.

Or H_1: $\beta_{YX} > 0$ (right tailed) i.e. regression coefficient in population is greater than zero.

Or H_1: $\beta_{YX} > 0$ (left tailed) i.e. regression coefficient in population is less than zero.

iii. Fix the level of significance α.

iv. **Test statistics;** Under the null hypothesis **H_0:** $\beta_{YX} = 0$, the test statistics for regression coefficient β_{YX} is $t = \frac{b - \beta_{YX}}{S_b}$, follows student t-distribution with (n-2) degrees of freedom , where S_b is the standard error of an estimator b_{YX}, and given as $S_b^2 = \frac{S_e^2}{\sum u^2}$, $S_e^2 = \frac{1}{(n-2)}[\sum v^2 - b_{YX}\sum uv]$

v. **Critical value:** For α level of significance and the (n-2) d.f. find the critical value of t as t_α from table.

vi. **Conclusion:** - If $|t| < t\alpha$ we accept our null hypothesis at α level of significance and If $|t| \geq t_\alpha$ we reject our null hypothesis and accept alternative hypothesis.

Confidence limit for regression coefficient β_{YX}

The approximate $(1 - \alpha)$ 100% confidence limit for regression coefficient β_{YX} is given as $b_{YX} \pm t_\alpha(n-2)\, S_b$

Example 9

Following table shows X is the tensile force applied to a steel and specimen in thousands of pounds, and Y is the resulting elongation in thousands of an inch

X	1	2	3	4	5	6
Y	14	33	40	63	76	85

i. Find the equation of regression line of Y on X

ii. Estimate the elongation when the tensile force is 3.5 thousand pounds

iii. Test the significance that the value of constant of regression in population is greater than 1.2 at 5% level of significance

iv. Test the significance that the value of regression of Y on X on whole population is greater than 14 at 5% level of significance

Solution

Calculation table for regression line of Y on X

X	Y	X^2	Y^2	XY
1	14	1	196	14
2	33	4	1089	66
3	40	9	1600	120
4	63	16	3969	252
5	75	25	5776	380
6	85	36	7225	510
21	**311**	**91**	**19855**	**1342**

Here $n = 6$, $\sum X = 21$, $\sum Y = 311$, $\sum X^2 = 91$, $\sum Y^2 = 19855$ and $\sum XY = 1342$

i. The regression equation of Y on X i.e. $Y = a + b\,X$ ... (i), where Y is elongation in thousand inch, X is tensile force applied to a steel and 'a' and 'b' are constant obtained by solving following two normal equations.

$\Sigma Y = na + b\Sigma X$

$\Sigma XY = a\Sigma X + b\Sigma X^2$

$\therefore$ $311 = 6a + 21b$ ... (ii)

$1342 = 21a + 19855$ ... (iii)

On solving equation (ii) and (iii) we get the value of a and b as

$a = 1.133$ and $b = 14.4857$

$\therefore$ The required line of regression is $\hat{Y} = 1.133 + 14.4857\ X$.

ii. The estimated elongation in thousand inches for the given tensile force applied to steel as 3.5 thousand pound is

$\hat{Y} = 1.133 + 14.4857 \times 3.5 = 51.8332$ thousand inches.

iii. For the test the significance that the value of constant of regression in population is greater than 1.2 at 5% level of significance, we proceeded in the flowing stapes.

a. **Set up null hypothesis H_0:** $\beta_0 = 1.2$ i.e. Slope parameter of regression line in population is 1.2

b. **Set up alternative hypothesis H_1:** $\beta_0 > 1.2$ (right tailed) i.e. Slope parameter of regression line in population is greater than 1.2

c. The level of significance $\alpha = 5\% = 0.05$

d. **Test statistics:** Under the null hypothesis **H_0:** $\beta_0 = 1.2$, the test statistics for slope parameter β_0 is

$t = \dfrac{a - \beta_0}{S_a}$, follows student t-distribution with (n-2) degrees of freedom, where S_a is the standard error of an estimator 'a', and given as

$$S_a^2 = S_e^2\left(\frac{1}{n} + \frac{\bar{x}}{\Sigma u^2}\right)$$

Calculation of deviation values

X	Y	u = X – 3	v = Y – 40	u^2	v^2	uv
1	14	-2	-26	4	676	52
2	33	-1	-7	1	49	7
3	40	0	0	0	0	0
4	63	1	23	1	529	23
5	75	2	35	4	1225	70
6	85	3	45	9	2025	135
21	**310**	**3**	**70**	**19**	**4504**	**287**

Now, $S_e^2 = \dfrac{1}{(n-2)}[\Sigma v^2 - b_{YX}\Sigma uv]$

$= \dfrac{1}{(6-2)}[4504 - 14.4857 \times 287] = 86.65$

Then $S_a^2 = S_e^2\left(\dfrac{1}{n} + \dfrac{\bar{x}}{\Sigma u^2}\right) = 86.65\left(\dfrac{1}{6} + \dfrac{(3.5)^2}{19}\right) = 70.308$

$\therefore$ $S_a = 8.385$.

Then, $t = \dfrac{a - \beta_0}{S_a} = \dfrac{1.133 - 1.2}{8.385} = -0.00799$

e. **Critical value**: For $\alpha = 5\%$ level of significance and the (6 - 2) = 4 degree of freedom and one tailed test the critical value of t as $t_{0.05}(4) = 2.132$

f. **Conclusion:** Here $|t| = 0.00799$ is less than $t_{0.05}(4) = 2.132$ we accept our null hypothesis at $\alpha = 5\%$ level of significance and Thus, concluded that the intercept of regression line in population is 1.2 it means that value of β_0 may not greater than 1.2

iv. To test the significance that the value of regression of Y on X on whole population is greater than 14 at 5% level of significance, we proceed in the following steps

a. **Set up null hypothesis H_0:** $\beta_{YX} = 14$ i.e. the value of regression coefficient in population is 14.

b. **Set up alternative hypothesis H_1:** $\beta_{YX} > 14$ (right tailed test i.e. regression coefficient in population greater than 14.

c. The level of significance $\alpha = 5\% = 0.05$.

d. **Test statistics:** Under the null hypothesis H_0: $\beta_{YX} = 14$, the test statistics for regression coefficient β_{YX} is

$t = \dfrac{b - \beta_{YX}}{S_b}$ follows student t-distribution with (n-2) degrees of freedom, where S_b is the standard error of an estimator b_{YX}, and given as $S_b^2 = \dfrac{S_e^2}{\Sigma u^2} = \dfrac{86.65}{19} = 4.5605$

$\therefore$ $S_b = 2.135$

Therefore, $t = \dfrac{b - \beta_{YX}}{S_b} = \dfrac{14.4857 - 14}{2.135} = 0.2275$

e. **Critical value**: For $\alpha = 5\%$ level of significance and the (6 - 2) = 4 degree of freedom and one tailed test the critical value of t as $t_{0.05}(4) = 2.132$

f. **Conclusion:** Here tabulated value of $t = 0.2275$ is less than $t_{0.05}(4) = 2.132$ we accept our null hypothesis at $\alpha = 5\%$ level of significance and Thus, concluded that the regression coefficient in population is not greater than 14.

Example 10

The following measurements show the respective height in inches of 10 fathers and their eldest sons

Father (X)	67	63	66	71	69	65	62	70	61	72
Son (Y)	68	66	65	70	67	67	64	71	62	63

a. Find the regression line of son's height on father's heights

b. Estimate the height of son for the given height of father as the 70 inches.
c. Test the significance of the regression coefficient.
d. Test the significance of the intercept of line of regression.

Solution

Calculation table

X	Y	X^2	Y^2	XY
67	68	4489	4624	4556
63	66	3969	4356	4158
66	65	4356	4225	4290
71	70	5041	4900	4970
69	69	4761	4761	4761
65	67	4225	4489	4355
62	64	3844	4096	3968
70	71	4900	5041	4970
61	60	3721	3600	3660
72	63	5184	3969	4536
666	663	44490	44061	44224

Here n = 10, $\Sigma X = 666$, $\Sigma Y = 663$, $\Sigma X^2 = 44490$, $\Sigma Y^2 = 44061$ and $\Sigma XY = 44224$

a. The regression equation of Y on X i.e. $Y = a + b X$ ... (i), where Y is height of son , X is height of father and 'a' and 'b' are constant obtained by solving following two normal equations.

$\Sigma Y = na + b \Sigma X$

$\Sigma XY = a \Sigma X + b \Sigma X^2$

$\therefore$ $663 = 10a + 666b$ ... (ii)

$44224 = 666a + 44490$ ... (iii)

On solving equation (ii) and (iii) we get the value of a and b as $a = 32.53$ and $b = 0.507$

$\therefore$ The required line of regression is $\hat{Y} = 32.53 + 0.507 X$.

b. The estimated height of son for the given height of father as 70 inches is

$\hat{Y} = 32.53 + 0.507 \times 70 = 68.02$ inches.

c. For the test of significance of observed regression coefficient. We proceeded in the flowing stapes.

i. **Set up null hypothesis H_0:** $\beta_{YX} = 0$ i.e. regression coefficient in population is zero.

ii. **Set up alternative hypothesis H_1:** $\beta_{YX} \neq 0$ (two tailed) i.e. regression coefficient in population is not equal to zero.

iii. The level of significance $\alpha = 5\% = 0.05$.

iv. **Test statistics:** Under the null hypothesis H_0: $\beta_{YX} = 0$, the test statistics for regression coefficient β_{YX} is $t = \frac{b - \beta_{YX}}{S_b}$ follows student t-distribution with

(n-2) degrees of freedom , where S_b is the standard error of an estimator b_{YX}, and given as $S_b^2 = \frac{S_e^2}{\Sigma u^2}$

Calculation table

X	Y	u=x-66.6	v=Y-66.3	u^2	v^2	uv
67	68	0.4	1.7	0.16	2.89	0.68
63	66	- 3.6	- 0.3	12.96	0.09	1.08
66	65	- 0.6	- 1.3	0.36	1.69	0.78
71	70	4.4	3.7	19.36	13.69	16.28
69	69	2.4	2.7	5.76	7.29	6.48
65	67	- 1.6	0.7	2.56	0.49	- 1.12
62	64	- 4.6	- 2.3	21.16	5.29	10.58
70	71	3.4	4.7	11.56	22.09	15.98
61	60	- 5.6	- 6.3	31.36	39.69	35.28
72	63	5.4	-3.3	29.16	10.89	- 17.82
666	**663**	**0**	**0**	**134.4**	**104.1**	**68.2**

Here n = 10, $\Sigma u = 0$, $\Sigma v = 0$, $\Sigma u^2 = 134.4$, $\Sigma v^2 = 104.1$and $\Sigma uv = 68.2$,

Now we have $b = 0.507$

$$S_e^2 = \frac{\Sigma v^2 - b_{YX} \Sigma uv}{(n-2)} = \frac{104.1 - 0.507 \times 68.2}{(10-2)} = 8.69$$

$$S_b^2 = \sqrt{\frac{S_e^2}{\Sigma u^2}} = \frac{8.69}{134.4} = 0.0646$$

$\therefore S_b = 0.2543$

Therefore $t = \frac{b - \beta_{YX}}{S_b} = \frac{0.507 - 0}{0.2543} = 1.99$

v. **Critical value:** For $\alpha = 5\%$ level of significance and the (10 - 2) = 8 d.f. the critical value of t as $t_{0.05}(8) = 2.306$

vi. **Conclusion:** - Here $|t| = 1.99$ is less than $t_{0.05}(8) = 2.306$ we accept our null hypothesis at $\alpha = 5\%$ level of significance and Thus concluded that the regression coefficient in population is not significance.

d. To Test of significance of Slope parameter β_0, We proceeded in the following stapes

i. **Set up null hypothesis H_0:** $\beta_0 = 0$ i.e. Slope parameter of regression Line in population is zero.

ii. **Set up alternative hypothesis H_1:** $\beta_0 \neq 0$ (two tailed) i.e. Slope parameter of regression Line in population is not equal to zero.

iii. The level of significance $\alpha = 5\% = 0.05$

iv. **Test statistics:** Under the null hypothesis H_0: $\beta_0 = 0$, the test statistics for slope parameter β_0 is $t = \frac{a - \beta_0}{S_a}$, follows student t-distribution with (n-2) degrees of freedom , where S_a is the standard error of an estimator a, and given as

$$S_a^2 = S_e^2\left(\frac{1}{n} + \frac{\bar{x}}{\Sigma U^2}\right) = 8.69\left(\frac{1}{10} + \frac{(66.6)^2}{134.4}\right) = 284.23$$

$\therefore S_a = 16.86.$

Then, $t = \frac{a - \beta_0}{S_a} = \frac{32.53 - 0}{16.86} = 1.929$

v. **Critical value**: For $\alpha = 5\%$ level of significance and the (10 - 2) = 8 d.f. the critical value of t as $t_{0.05}(8) = 2.306$

vi. **Conclusion:** Here $|t| = 1.929$ is less than $t_{0.05}(8) = 2.306$ we accept our null hypothesis at $\alpha = 5\%$ level of significance and Thus concluded that the intercept of regression line in population is not significance.

EXERCISE - 14

1. Define t-test. What are the assumptions made about t-distribution?
2. What are the properties of t-distribution? Write about the application of t-distribution.
3. What are the steps about the test of significance of single mean for small sample?
4. What are the assumption about the test of difference of mean in small sample and write down the steps of test of significance difference between two mean in small sample.
5. Write down the steps of test of significance of difference between two mean when sample are dependently paired.
6. What are the steps of test of significance of regression coefficient?
7. Define f-distribution and what are the assumptions made for f-distribution.
8. What are the applications of f-distribution?
9. Write down the steps for test of significance of equality of two population variance.

Numerical For Single Mean

10. Prices of shares of a company on the different days in a month were found to be 66, 65, 69, 70, 69, 71, 70, 63, 64 and 68. Discuss whether the mean price of the shares in the month is 65.

11. Ten students are chosen at random from a normal population and their heights are found to be: 63, 63, 66, 67, 68, 69, 70, 70, 71, 71, inches. Test if the sample belongs to the population whose mean height is 66 inches.
12. A salesman is expected to affect an average sale of Rs. 3,500. A sample test revealed that a particular salesman had made the following sales: Rs. 3700, 2500, 3400, 5200, 3000 and 2000

    Using 0.05 level of significance concludes whether his work is below standard or not.
13. A random sample of eight cigarettes of a certain brand has an average nicotine content of 18.6 milligrams and a standard deviation of 2.4 milligrams. Is this in line with the manufacturer's claim that average nicotine content doesn't exceed 17.5 milligrams at 1% level of significance and assume the distribution of nicotine contents to be normal? Also find 99% confidence limit for population mean.
14. Ten specimens of copper wires drawn from a large lot have the following breaking strength in kg weight: 578,572, 570, 568, 572, 571, 570, 572, 596, and 548. Test whether the mean breaking strength of the lot may be taken be 578 kg weight.
15. A time study engineer developed a new sequence of operation elements that he hopes will reduce the mean cycle time of a certain production process. The results of a time stud y of 20 cycles are given below :

    Cycle time in minutes:

12.25	11.97	12.15	12.08	12.31	12.28	11.94	11.89	12.16	12.04
12.09	12.15	12.14	12.47	11.98	12.04	12.11	12.25	12.15	12.34

    If the present mean cycle time is 12.5 minutes, Should he adopt the new sequence?
16. A random sample of size 25 showed a mean of 65 inches with a standard deviation of 25 inches. Determine 98% confidence intervals for the mean of the population.
17. A random sample of 12 female students in a certain dormitory showed an average weekly expenditure of Rs. 16 for snack foods with a standard deviation of Rs. 2. Construct 90% confidence limits for the average amount spent each week on snack foods by female students living on the dormitory.
18. The five randomly selected samples of the rubber calls have a mean breaking strength of 169.5 pounds with a standard deviation of 5.7 pounds. Test the assumption that mean breaking strength in a population is 180 pounds at 1% level of significance.

Numerical For Difference of Mean

19. The incomes of a random sample of six officers in bank First are Rs. 530, 650, 680, 690, 710 and 720 per month. The incomes of a similar sample from nine

Exercise - 15

1. Define t-test. What are the assumptions made about t-distribution?
2. What are the properties of t-distribution? Write about the application of t-distribution.
3. What are the steps about the test of significance of single mean for small sample?
4. What are the assumption about the test of difference of mean in small sample and write down the steps of test of significance difference between two mean in small sample.
5. Write down the steps of test of significance of difference between two mean when sample are dependently paired.
6. What are the steps of test of significance of regression coefficient?
7. Define f-distribution and what are the assumptions made for f-distribution.
8. What are the applications of f-distribution?
9. Write down the steps for test of significance of equality of two population variance.
10. Prices of shares of a company on the different days in a month were found to be 66, 65, 69, 70, 69, 71, 70, 63, 64 and 68. Discuss whether the mean price of the shares in the month is 65.
11. Ten students are chosen at random from a normal population and their heights are found to be: 63, 63, 66, 67, 68, 69, 70, 70, 71, 71, inches. Test if the sample belongs to the population whose mean height is 66 inches.
12. A salesman is expected to affect an average sale of Rs. 3,500. A sample test revealed that a particular salesman had made the following sales: Rs. 3700, 2500, 3400, 5200, 3000 and 2000

    Using 0.05 level of significance concludes whether his work is below standard or not.
13. A random sample of eight cigarettes of a certain brand has an average nicotine content of 18.6 milligrams and a standard deviation of 2.4 milligrams. Is this in line with the manufacturer's claim that average nicotine content doesn't exceed 17.5 milligrams at 1% level of significance and assume the distribution of nicotine contents to be normal? Also find 99% confidence limit for population mean.
14. Ten specimens of copper wires drawn from a large lot have the following breaking strength in kg weight: 578,572, 570, 568, 572, 571, 570, 572, 596, and 548. Test whether the mean breaking strength of the lot may be taken be 578 kg weight.

15. A time study engineer developed a new sequence of operation elements that he hopes will reduce the mean cycle time of a certain production process. The results of a time stud y of 20 cycles are given below :

    Cycle time in minutes:

    12.25, 11.97, 12.15, 12.08, 12.31, 12.28, 11.94, 11.89, 12.16, 12.04, 12.09, 12.15, 12.14, 12.47, 11.98, 12.04, 12.11, 12.25, 12.15, 12.34.

    If the present mean cycle time is 12.5 minutes, Should he adopt the new sequence?
16. A random sample of size 25 showed a mean of 65 inches with a standard deviation of 25 inches. Determine 98% confidence intervals for the mean of the population.
17. A random sample of 12 female students in a certain dormitory showed an average weekly expenditure of Rs. 16 for snack foods with a standard deviation of Rs. 2. Construct 90% confidence limits for the average amount spent each week on snack foods by female students living on the dormitory.
18. The five randomly selected samples of the rubber calls have a mean breaking strength of 169.5 pounds with a standard deviation of 5.7 pounds. Test the assumption that mean breaking strength in a population is 180 pounds at 1% level of significance.
19. The sales figure of an item in eight shops before and after advertisement is given as :

    Before: 70, 65, 48, 72, 80, 92, 98, 100

    After : 72, 70, 53, 75, 84, 95, 105,104

    Test whether advertisement was effective or not.
20. The incomes of a random sample of six officers in bank I are Rs. 530, 650, 680, 690, 710 and 720 per month. The incomes of a similar sample from nine officers in bank II are Rs. 610, 620, 650, 660, 690, 700, 710, 720 and 730 per month. Discuss the validity of the suggestion that bank I pays its officers much better than bank II.
21. The average breaking strength of steel rods is specified to be 18.5 thousand pounds. To test this, a sample of 14 rods was tested. The mean and standard deviation obtained were 17.85 and 1.955 thousand pounds respectively. Is the result of the experiment significant? Also obtained the 95% confident limits from the sample of the average breaking strength of steel rods.
22. seven housewives were randomly sampled and it was determined that they walked an average of 39.2 miles per week in their housework, with a standard deviation of 3.2 miles per week, Construct 95% confidence interval to the population mean.
23. Hotel's manager in Kathmandu wants to know the hotels average daily registration. The following table presents the numbers of guest registered

each of 27 randomly selected days. Calculate the sample mean, standard errors of mean and 95% confidence limits of population mean:

61	57	53	60	64	57	54	58	63
61	50	59	50	60	57	58	62	63
60	54	54	61	51	53	62	57	60

24. Two independent samples of 8 and 7 items respectively had the following values, whether these two samples are drawn from a same population or not.

Sample-I	9	11	13	11	15	9	12	14
Sample-II	10	12	10	14	9	8	10	

25. Two horses A and B were tested according to the time in seconds to run a particular track with the following results:

Horse A	28	30	32	33	33	29	34
Horse B	29	30	30	24	27	29	

Test whether the two horses have the same running capacity at 5% level of significance

26. The gain in weight of two random samples of rats fed on two different diets A and B are given below. Examine whether the difference in mean increase in weight is significant.

Diet A	13	14	10	11	12	16	10	8	
Diet B	7	10	12	8	10	11	9	10	11

27. A reading test is given to an elementary school class that consists of 12 boys and 10 girls. The result of the test are as follows:

	Boys	Girls
Mean	74	70
S. D.	8	10

Is the difference between the means of two groups significant at 5% level of significance?

28. A random sample of 16 values from a normal population has a mean of 41.5 inches and sum of squares of deviations from the mean is equal to 135 inches. Another sample of 20 values from an unknown population has a mean of 43 inches and sum of squares of deviations from the mean is equal to 171 inches. Show that the two samples may be regarded as coming from the same normal population.

	I Factory	II Factory
Sample size	12	10
Mean salaries(Rs.)	1,050	980
Sample standard deviation	68	74

29. A manufacturing company is interested in knowing if there is a difference in the average salary received by foremen in two factories accordingly sample of 12 foremen in the first factory and 10 foremen in the second factory are selected at random. Based upon experience foremen's salaries are known to be approximately normally distributed and the standard deviations are about the same.

30. A group of 5 Patients treated with medicine 'A' weight 42, 39, 48, 60, and 41 kg. Second group of 7 patients from the same hospital treated with medicine 'B' weight 38, 42, 56, 64, 68, 69 and 62 kg. Do you agree with the claim that medicine 'B' increases the weight significantly?

31. Two types of batteries are tested for their length of life and the following data are obtained.

	Sample size	Mean life time	Variance
Type A	9	600	121
Type B	8	640	144

Is there a significant difference in the two means?

32. The means of two random samples of size 9 and 7 are 196.42 and 198.82 respectively. The sums of the squares of the deviations from the mean are 26.94 and 18.73 respectively. Can the samples be considered to have been drawn from the same normal population?

33. The average numbers of articles produced by two machines per day are 200 and 250 with s.d 20 and 25 respectively. On the basis of records of 25 days production can you regard both the machines are equally efficient at 1% level of significance?

34. An IQ test was administered to 5 persons before and after they were trained. The results are given below:

Roll. No.	I	II	III	IV	V
IQ before training	110	120	123	132	125
IQ after training	120	118	125	130	121

Test whether there is any change in IQ after the training program

35. Ten soldiers visit a rifle range for two consecutive weeks. for the first and second week their scores are

First week	67	24	57	55	63	54	56	68	33	43
Second week	70	38	58	58	56	67	68	72	42	38

Examine if there is any significant difference in the performance.

36. Eleven students of B.E. were given a test in statistics. They were given 15 days special coaching and thereafter were given a second test. Marks obtained in the two tests are given below:

Students	I	II	III	IV	V	VI	VII	VII	IX	X	XI
I-test	23	20	19	21	18	20	18	17	23	16	19
II-test	24	19	22	18	20	22	20	20	23	20	18

Do the marks indicat4e the special coaching has benefited the students at 5% level of significance.

37. A certain diet newly introduced to each of the 12 pigs resulted in the following increase in body weight: 6, 3, 8, -2, 3, 0, -1, 1, 6, 0, 5, and 4. Can you conclude that the diet is effective in increasing the weight of pigs?

38. The following are the Brinell hardness values obtained for samples of two magnesium alloys.

Alloy I	66.3	63.5	64.9	61.8	64.3	64.7	65.1	64.5	68.4	63.2
Alloy II	71.3	60.4	62.6	63.9	68.8	70.1	64.8	68.9	65.8	66.2

Use the 5% level of significance to test the hypothesis that mean hardness in the population are same.

39. The following were obtained in experiment designed to check whether there is a systematic difference in the weights obtained with two different scales

Rock specimen		I	II	III	IV	V	VI	VII	VII	IX	X
wt. (gram)	Scale-I	11.23	14.36	8.33	10.50	23.42	9.15	13.47	6.47	12.40	19.38
	Scale-II	11.27	14.41	8.35	10.52	23.41	9.17	13.52	6.46	12.45	19.35

Test at the 5% level of significance whether the difference of the means of the weights obtained with the two scales is significant.

40. In a study of the effectiveness of physical exercise in weight reduction a group of 16 persons engaged in a prescribed program of physical exercise for one month showed the following results:

wt. in Pound																
	Before	209	178	169	212	180	192	158	180	170	153	183	165	201	179	243	144
	After	196	171	170	207	177	190	159	180	164	152	179	162	199	173	231	140

Use the 1% level of significance to test whether the prescribed program of exercise is effective

41. A random sample of pairs of observations from a normal population gives a correlation coefficient of 0.42 .Is it likely that the variables in the population are uncorrelated

42. A coefficient of correlation of 0.2 is derived from a random sample of 625 pairs of observations. Is this value of r significant? Also find the 95% and 99% confidence limits to the correlation coefficient in the population.

43. The following table gives the age and blood pressure of to women.

Age (y)	56	42	36	47	49	42	60	72	63	55
Pressure	147	125	118	128	145	140	155	160	149	150

i. Determine the least square regression equation of Y on X.
ii. Estimate the blood pressure of a women whose age is 45 years.
iii. Test the significance of slop parameter β_0.
iv. Test the significance of regression coefficient b_{yx}

44. Assuming that we conduct an experiment with eight fields planted with corn, four fields having no nitrogen fertilizer and four field having 80 kegs of nitrogen fertilizer, the resulting corn yields are shown in the table in bushels per acre.

Field	1	2	3	4	5	6	7	8
Nitrogen(kg)	0	0	0	0	80	80	80	80
Corn yield/hector	1	360	60	180	128	112	112	760

i. Determine the least square regression equation of Y on X.
ii. Estimate the corn yield when we use 60 kg of nitrogen fertilizer.
iii. Test the significance of slop parameter β_0 .
iv. Test the significance of regression coefficient b_{YX}

45 Ministry of Finance wants to compare the consistency level of price of share of two banks using one week's sample record. The following table shows the price per share of the two banks.

Price of Bank A	11	10	13	9	8	7	7
Price of bank B	15	14	9	7	11	6	10

Is there any significant difference in the consistency level of two banks?

46. It is desired to determine whether there is less variability in the silver planting done by company A than in that done by company B. If independent random samples of size 12 of the two companies were yield

0.035 mils and 0.062 mils standard deviation respectively, test the null hypothesis that sample are drawn from the normal population with equal variance.

47. The variances of the heat producing capacity of coal from two mines are 15,750 and 10,920 respectively with independent random samples of sizes 5 and 6. Test the hypothesis that sample are drawn from the normal population with equal variance at 1% level of significance

48. The time taken by the two secretaries of HMG to do decisions on five and seven sampled files are :
Time taken by first secretary in number of days: 7, 8,14,16,15,10.
Time taken by second secretary in number of days : 6,10,12,15,5,20,11.
Test whether the two secretaries have the same consistency level in taking decisions.

49. Pull-strength tests on 10 soldered leads for a semiconductor device yield the following results in pounds force required to rupture the bond: 15.8, 12.7, 13.2, 16.9, 10.6, 18.8, 11.1, 14.3, 17.0, and 12.5. Another set of eight leads was tested after encapsulation to determine whether the pull-strength has been increased by encapsulation of the device, with the following results: 24.9, 23.6, 19.8, 22.1, 20.4, 21.6, 21.8, and 22.5. Use 1% level of significance to test whether it is reasonable to assume that the two samples come from populations with equal variances

50. The following are the values in thousands of an inch obtained by two engineers in 10 successive measurements with the same micrometer. Is both engineer significantly equal consistent [Hint consistency are measured by variance]

Engineer A	503	505	497	505	505	495	502	499	493	510	501
Engineer B	502	497	492	492	498	499	495	497	496	498	

51. In one sample of 10 observations from a normal population, the sum of the square of the deviations of the sample values from the sample mean is 102.4 and in another sample of 12 observations from another normal population, the sum of the square of the deviations of the sample values from the sample mean is 120.5. Examine whether the two normal population have the same population.

52. The sales data of an item in six shops before & after a special promotional campaign are.

Shopes	A	B	C	D	E	F
Before	53	28	31	48	50	42
After	58	29	30	55	56	45

Can the campaign be judged to be a success ?

53. A certain stimulus administrated to each of the 12 patents resulted in the following change of blood pressure.
5, 2, 8, -1, 3, 0, -2, 1, 5, 0, 4, 6
Can it be conclude that the stimulus will be in general be accompanied by an increase in blood pressure.

54. A company has implemented a special trade promotion for its propane store and feels that the promotion should result in a price change for the consumer. Company knows that before the promotion began, the

average retail price of the store was Rs. 449.5 and the s.d. was 57.5. Company samples 25 of its retailers after the promotion begins and finds the mean price for the stores is now Rs. 429.5 at a 1% level of significance, does company have reason to believe that the average retail price to the consumer has decreased?

55. Generally electric has developed a new bulb whose design specifications call for a light output of 960 lumens compared to an earlier model that produced only 750 lumens. The company's data indicate that the s.d. of light output for this type of bulb is 18.4 lumens. From a sample of 20 new bulbs, the testing committee found an average light output of 954 lumens per bulb. At a 5% significance level, can general electric conclude that its new bulb is producing the specified 960 lumen output?

Answers:

10. t = 2.82, not significant
11. t =1.89, not significant
12. t = -0.44; not significant,
13. t =1.30, not significant [14.85, 20.15]
14. t =-1.72, not significant
15. t =-11.42 significant
16. 52.29, 77.71,
17. t =14.92, 17.08,
18. 4.12, not significance
19. t = -4.84; significant,
20. t= 0.099; not significant,
21. t =1.244, [16.92, 18.77]
22. 36.85, 41.55
23. 57.74, 0.79,[56.4,59.1]
24. t = 0.555, not significant
25. t =2.5, significance
26. t = 1.95, not significant
27. t = 0.994, not significant
28. t = -1.49, not significant
29. t = 2.202, not significant
30. t = -1.7 not significant
31. t =-6.735, significant
32. t = -2.64 significant
33. t =-7.65 significant
34. t = 0.816. Not significant
35. t =-2.039, not significance,
36. t =1.483, not significant
37. Not significant,
38. Significant
39. 2.205, not significant,
40. t =4.06, significant
41. 2.31, significant
42. t = 5.09, significant, {0.125, 0.275}, {0.101,0.299}
43. Y = 83.758 + 1.11 X, Y = 135,
44. Y = 180 + 11.12 X, Y = 847.5
45. f = 1.514, not significant
46. 3.14, significance
47. 1.44, not significance,
48. f = 1.346, not significant
49. 2.797, not significance
50. t=. 2.4,
51. 1.038
52. | t | = 2.78 significant
53. t_{cal} = 2.894, significant at 5% level.
54. t = 1.74, not significant.
55. t = – 1.458, not significant.

Chapter 16

Analysis of Categorical Data

Chi-square test

We have studied that the various test of significance like Z-test, t-test and f-test which are based on the assumption that the sample are drawn from the normally distributed population. These tests are made assumption about population parameter which are called parametric tests

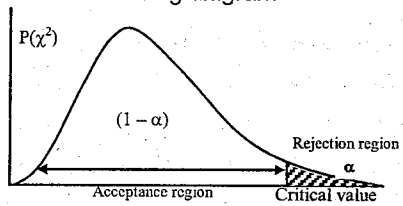
Chi-square (χ^2) is a non- parametric test because it depends only on the set of observed and expected frequencies and degree of freedom. Since χ^2 - test doesn't make any assumption about population parameters. It is also called distribution free test. χ^2 -test is a test which describes the magnitude of difference between observed frequency and theoretical or expected frequencies under certain assumptions. It is defined as

$$\chi^2 = \Sigma \frac{(O - e)^2}{e}$$, follows chi-square distribution with certain degree of freedom

Critical value of - χ^2 distribution

The critical values of χ^2 -test for various degrees of freedom and level of significance is given at the end of this book. The critical value of χ^2- test for α level of significance and (n-1) = ν (nu) degree of freedom $\chi^2_{\alpha}(\nu)$ is given by following probability

$P(\chi^2 > \chi^2_{\alpha}(\nu)) = \alpha$ as shown in the following diagram



Application of Chi-square distribution

Chi- square distribution has a large number of applications in statistics some of the most important applications are as follows.

i. To test if the hypothetical value of the population variance
ii. To test the 'goodness of fit'
iii. To test the independence of attributes.
iv. To test the homogeneity of independent estimate of the population variance.
v. To combine various probabilities obtained from independent experiments to given a single test of significance.

iv. To test the homogeneity of independent estimate of the population variance.

v. To combine various probabilities obtained from independent experiments to given a single test of significance.

15.2 CHI- SQUARE TEST FOR THE SINGLE POPULATION VARIANCE

Suppose we want to test if a random sample $X_1, X_2 \ldots\ldots X_n$ has been drawn from a normal population with a specified variance $\sigma^2 = \sigma_0^2$ (say) we can use the following steps.

i. **Set up the null hypothesis H_0:** $\sigma^2 = \sigma_0^2$ (say) i.e. sample are drawn from the normal population with specified variance σ_0^2

ii. **Set up the alternative hypothesis H_1:** $\sigma^2 \neq \sigma_0^2$ i.e. sample are not drawn from the normal population with specified variance σ_0^2

iii. Fix the Level of significance α.

iv. **Test statistics:** under null hypothesis Ho, the test statistics is

$$\chi^2 = \frac{\Sigma(x - \bar{x})2}{\sigma_0^2} = \frac{ns^2}{\sigma_0^2}$$, follows chi-square distribution with (n-1) d.f, where

$$s^2 = \frac{1}{n}\Sigma(x - \bar{x})$$

v. **Critical value:** for α% level of significance and (n-1) d.f find the critical value of chi-square from the chi-square table as $\chi^2\,\alpha(n-1)$

vi. **Conclusion:** If calculated value of χ^2 is greater that tabulated value of χ^2 as $\chi^2{}_\alpha(n-1)$ we reject null hypothesis otherwise accept it.

Example 1

A manufacturer claims that any of his lot of items can't have a variance more than 1 cm^2. A sample of 25 items has a variance of 1.2 cm^2. Test whether the claim of the manufacture is correct.

Solution

We are given population variance $\sigma^2 = 1\ cm^2$

Sample size n = 25

Sample variance $s^2 = 1.2\ cm^2$

We proceeds in the following steps to test the claim as

i. **Null hypothesis H_0 :** $\sigma^2 = 1\ cm^2$ i.e. the claim is correct

ii. **Alternative hypothesis H_1:** $\sigma^2 \neq 1\ cm^2$ i.e. the claim is not correct.

iii. The level of significance = 5%

iv. **Test statistics**: under null hypothesis H_0:$\sigma^2 = 1\ cm^2$ the test statistics is

$$\chi^2 = \frac{ns^2}{\sigma_0^2}$$,follows chi-square distribution with n-1 d.f

$$= \frac{25 \times 1.2}{1} = 30$$

v. **Critical value:** For 5% and (n-1) = 24 degrees of freedom the critical value of chi-square is $\chi^2{}_{0.05}(24) = 37.652$

vi. **Conclusion:** Here $\chi^2{}_{cal} = 30$ is less than $\chi^2{}_{tab} = 37.652$ therefore we accept null hypothesis at 5% level of significance and conclude that the sample have been drawn from the normal population with variance = 1 cm^2

Example 2

It is believed that precision (as measured by the variance) of an instrument is no more than 0.16. Write down the null hypothesis and alternative hypothesis for testing this belief. Carry out the test at 1% level given 11 measurements of the same subject on the instrument as 2.5, 2.3, 2.4, 2.3, 2.5, 2.7, 2.5, 2.6, 2.6, 2.7and 2.5

Solution

We are given population variance $\sigma^2 = 0.16$ and Sample size n = 11

We test the precision by using the following steps as

i. **Null hypothesis H_0 :** $\sigma^2 = 0.16$ i.e. the precision of the instrument is 0.16

ii. **Alternative hypothesis H_1:** $\sigma^2 \neq 0.16$ i.e. the precision of the instrument is not 0.16

iii. The level of significance = 1%

iv. **Test statistics**: under null hypothesis H_0 the test statistics is

$\chi^2 = \frac{ns^2}{\sigma_0^2}$ follows chi-square distribution with (n–1) d.f

X	$(X-\bar{X})$	$(X-\bar{X})^2$
2.5	– 0.009	0.000
2.3	– 0.209	0.044
2.4	– 0.109	0.012
2.3	– 0.209	0.044
2.5	– 0.009	0.000
2.7	0.191	0.036
2.5	– 0.009	0.000
2.6	0.091	0.008
2.6	0.091	0.008
2.7	0.191	0.036
2.5	– 0.009	0.000
27.6	**Total**	0.189

Here n= 11 mean $\bar{X} = \frac{27.6}{11} = 2.509$

$$\chi^2_{cal} \frac{\Sigma (X-\bar{X})^2}{\sigma^2} = \frac{0.189}{0.16} = 1.182$$

v. **Critical value:** For 1% and (n-1) = 10 degrees of freedom the critical value of chi-square is $\chi^2_{0.01}(10) = 23.2$

vi. **Conclusion:** Here $\chi^2_{cal} = 1.182$ is less than $\chi^2_{tab} = 23.2$ we accept null hypothesis at 1% level of significance and conclude that the sample have been drawn from the normal population with variance 0.16 or the precision of the machine is 0.16

15.3 CHI-SQUARE TEST OF GOODNESS OF FIT

In test of hypothesis or estimation of parameters it is usually assumed that the random variable follows particular distribution like Binomial, Poisson, normal distribution etc. But often the need is felt to conform whether our assumption is true or not.

So, on the basis of outcomes of a trial or observational data, a very powerful test for testing the significance of the discrepancy between theory and experiment was given by Prof. Karl Pearson in 1900 and is known as 'chi-square test of goodness of fit' it enables us to find if the deviation of the experiment from theory is just by chance or is it really due to the inadequacy of the theory to fit the observed data.

If O_i (i =1, 2 .....n) is a set of observed or experimental frequencies and e_i (i =1, 2 .....n) is the corresponding set of expected (theoretical or hypothetical) frequencies then Karl Pearson's chi-square test is given as

$\chi^2 = \Sigma \frac{(O - e)^2}{e}$, Follows chi-square distribution with (n-1) degrees of freedom. Where, $\Sigma O = \Sigma e = N$ total observations.

15.3.1 Steps for Test of Goodness of Fit

Following are the steps for test of goodness of fit

i. **Set up null hypothesis H_0:** the data have come from the assumed distribution or from the composite hypothesis or fitting of the certain distribution is good.

ii. **Set up alternative hypothesis H_1:** the fitting of the certain distribution in not good.

iii. Fix the level of significance α.

iv. **Test statistics:** under null hypothesis H_0 the test statistics is,

$\chi^2 = \frac{\Sigma(O - e)^2}{e}$, follows chi-square distribution with (n-1) degrees of freedom.

v. **Critical value**: For α level of significance and (n-1) degrees of freedom find the critical value of chi-square as χ^2 cal

vi. **Conclusion:** If calculated value of chi-square $\chi 2$ cal is greater than tabulated value of chi-square χ^2_{tab} we reject our null hypothesis otherwise accept it.

Remarks: the expected frequencies for composite hypothesis is calculated by using the composite hypothesis,

i. For ordinary case of equal number of observations expected frequencies are calculated as $e = \frac{\Sigma O}{n}$, where n is number of observations.

ii. If composite hypothesis is in the proportion or ratio of observed frequency. In that case the expected frequencies are calculated on the basis of given proportion.

iii. If there is a frequency distribution to which a theoretical frequency distribution (like Binomial, Poisson, normal) is applicable the expected frequencies would be derived by applying the appropriate theoretical distribution (i.e. the case of goodness of fit when probability is completely specified)

15.3.2 Condition for the Validity of Chi-Square Test

Chi-Square test is an appropriate test for large values of n. For the validity of chi-square test of 'goodness of fit' between theory and experiment, the following conditions must be satisfied. Which are also the assumptions of chi square distribution of goodness of fit.

i. The total frequency (N) should be reasonably large, say greater than 50.

ii. The sample observations should be independent.

iii. The total observed frequencies and expected frequencies should be equal i.e. $\Sigma O = \Sigma e = N$, where 'O' is observed frequency, 'e' is expected frequency and N- total frequency.

iv. No theoretical or expected frequency is less than 5. If any expected frequency is less than 5 we should be pooled the expected frequencies which are less than 5 with the succeeding or preceding frequencies so that the resulting sum is 5 or more and adjust the degrees of freedom accordingly.

15.3.3 Degrees of Freedom for Chi-Square Test of Goodness of Fit

The number of values that can choose freely is called degrees of freedom; it is denoted by ν (nu) or d.f. Thus d.f = (n-1) in general.

But, if we are needed to calculate the parameters of a theoretical distribution on the basis of given frequency distribution for obtaining the expected frequencies of the distribution, then we subtract 1 degrees of freedom for each parameter estimated. If we are required to fit Binomial and Poisson distribution to the given set of date, we lose 1 degrees of freedom for

applying χ^2 - test. Since for binomial distribution we have estimated only one parameter P, n being the given in the date and for the poison distribution we have to estimated one parameter λ.

In Binomial distribution if the hypothetical value of P is given then we don't lose any degrees of freedom for applying χ^2-test.

However, if we required fitting the normal distribution we lose 2 degrees of freedom for applying χ^2- test because of calculating two parameters μ and σ^2 from the given data.

Thus the degrees of freedom for the χ^2 - test of goodness of fit are given by **degrees of freedom = $n - 1 - k_1 - k_2$.**

i. 1 degrees of freedom is lost due to linear constraint $\Sigma O = \Sigma e = N$.

ii. K_1 is the number of parameters computed (if any) from the given data for estimation the expected frequency of the distribution.

iii. K_2 is the number of degrees of freedom lost due to pooling of expected frequencies which are less than 5.

Example 3

The following table give the number of failed student occurred during the past 7-semester. Find whether the number of failed student are uniformly distributed over the past 7-semester.

Semester	I	II	III	IV	V	VI	VII	Total
No. of failed students	14	17	12	13	15	14	13	98

Solution

i. **Null hypothesis H_o:** the given frequencies are consistent with the belief that number of failed students were the same during the past 7-semester.

ii. **Alternastive hypothesis H_1:** the given frequencies are not-consistent with the belief that number of failed students were the same during the past 7-semester.

iii. Level of significance = 5% = 0.05.

iv. **Test statistics:** under null hypothesis Ho, the test statistics is $\chi^2 = \Sigma\frac{(O - e)^2}{e}$, follows distribution with (n-1) d.f, where o- observed frequency and e-expected frequency. The expected frequency are calculate by using the formula $e = \frac{\Sigma o}{n}$, since under null hypothesis the number of failed students would be same

Calculation of chi-square

observed frequency (O)	expected frequency (e)	(O-e)	$(o-e)^2$	$\frac{(o-e)^2}{e}$
14	14	0	0	0
17	14	3	9	0.643
12	14	-2	4	0.286
13	14	-1	1	0.071
15	14	1	1	0.071
14	14	0	0	0
13	14	-1	1	0.071
98	**98**			**1.143**

Therefore $\chi^2 = \Sigma\frac{(O - e)^2}{e} = 1.143$, Degree of freedom = (n-1) = 7-1 = 6

v. **Critical value:** for 5% level of significance and 6 d.f the tabulated value of chi-square is = 12.592.

vi. **Conclusion:** here calculated value of chi-square = 1.143 is less than tabulate value of chi –square = 12.592, we accept null hypothesis and we conclude that the given frequencies are consistent with the belief that number of failed students were the same during the past 7-semester

Example 4

In an experiment on pea-breeding, Mendel obtained the following frequencies of seeds: 315 round and yellow, 101 wrinkled and yellow, 108 round and green, 32 wrinkled and green, Total 556 . Theory predicts that the frequencies should be in the proportion 9:3:3:1respectively. Set up proper hypothesis and test it at 10% level of significance

Solution

i. **Null hypothesis H_o:** The observed figure don't differ significantly from the hypothetical frequencies which are in the ratio 9:3:3:1.

ii. **Alternative hypothesis H_1:** The observed figure differ significantly from the hypothetical frequencies which are in the ration 9:3:3:1.

iii. Level of significance = 10% = 0.10

iv. **Test statistics:** Under null hypothesis Ho, the test statistics is

$\chi^2 = \Sigma\frac{(O - e)^2}{e}$, follows distribution with (n-1) degrees of freedom, where O- observed frequency and e-expected frequency. The expected

frequencies for round and yellow = 556x9/16 = 313 and so on since under null hypothesis the numbers are in the ratio 9:3:3:1.

Calculation of chi-square

Observed frequency (O)	Expected frequency(e)	(O-e)	$(O-e)^2$	$\frac{(o-e)^2}{e}$
315	313	2	4	0.013
101	104	-3	9	0.087
108	104	4	16	0.154
32	35	-3	9	0.257
556	**556**			**0.510**

Therefore, $\chi^2 = \Sigma\frac{(O-e)^2}{e} = 0.51$, degree of freedom = (n –1) = 4 – 1 = 3 degrees of freedom

v. **Critical value:** for 10% level of significance and 3 degrees of freedom the tabulated value of chi-square is = 6.251

vi. **Conclusion:** here calculated value of chi- square = 0.51 is less than tabulate value of chi –square = 6.251, we accept null hypothesis and we conclude that the observed figures are in the ration 9:3:3:1

Example 5

A survey of 800 families with four children each revealed the following distribution.

No. of boys	0	1	2	3	4
No. of girls	4	3	2	1	0
No. of families	32	178	290	236	64

Is this result consistent with the hypothesis that male and female births are equally probable?

Solution

i. **Null hypothesis H_o:** the data are consistent with the hypothesis of equal probability for male and female births.

ii. **Alternative hypothesis H_1:** the data are not consistent with the hypothesis of equal probability for male and female births

iii. Level of significance = 5% = 0.05

iv. **Test statistics:** Under null hypothesis Ho, the test statistics is

$\chi^2 = \Sigma\frac{(O-e)^2}{e}$, follows distribution with (n-1) degrees of freedom, where

O- observed frequency and e-expected frequency. Under null hypothesis p = probability of male birth = ½ = q.

let x be the number of male birth then by binomial distribution.

P(x) = Probability of male birth in a family of 4 children=

$= {}^nC_x\, p^x\, q^{n-x} = {}^4C_x\,(1/2)^x\,(1/2)^{4-x} = {}^4C_x\,(1/2)^4$,x = 0, 1, ...4

Calculation of expected frequencies and chi-square

No. of boys	(O)	(e)	(O-e)	$(o-e)^2$	$\frac{(O-e)^2}{e}$
0	32	50	18	324	6.48
1	178	200	22	484	2.42
2	290	300	10	100	0.33
3	236	200	-36	1296	6.48
4	64	50	-14	196	3.92
Total	800	800	0	2400	19.63

Therefore $\chi^2 = \sum\frac{(O-e)^2}{e}$, = 19.63, degree of freedom = (n – 1) = 5 – 1 = 4 degrees of freedom

v. **Critical value:** for 5% level of significance and 4 degrees of freedom the tabulated value of chi-square is = 9.48

vi **Conclusion:** here calculated value of chi- square = 19.63 is greater than tabulate value of chi –square = 9.48, we reject null hypothesis and accept out alternative hypothesis and conclude that male and female birth are not equal probable.

15.4 TWO WAYS CONTINGENCY TABLE

A contingency table is a two-way table in which the columns are classified according to one criterion or attribute and rows are classified according to the other criteria or attribute. Each cell contains that the number of items O_{ij} possessing the quality of i^{th} row and j^{th} column,

Where i = 1, 2, ... n and j = 1, 2, ...m. In such a case the contingency table is said to be of order (n x m). Each row or column total is known as marginal total. Also the sum of row total $\sum_i R_i$ is equal to the sum of column total $\sum_j C_j$ i.e. $\sum_i R_i = \sum_j C_j = N$ the total sample size.

Contingency table helps to test the independence of two attribute. A two ways contingency table is as

Two ways Contingency Table

Rows	Columns						
	B_1	B_2	...	B_j	...	B_m	Total
A_1	O_{11}	O_{12}	...	O_{1j}	...	O_{1m}	R_1
A_2	O_{21}	O_{22}	...	O_{2j}	...	O_{2m}	R_2
...	...	...	...	.....	...	...	...
A_i	O_{i1}	O_{i2}	...	O_{ij}	...	O_{im}	R_i
...	...	...	...	...	...	...	...
A_n	O_{n1}	O_{n2}	...	O_{nj}	...	O_{nm}	R_n
Total	C_1	C_2	...	C_j	...	C_m	N

15.5 CHI-SQUARE TEST FOR INDEPENDENCE OF ATTRIBUTES

In order to test whether attribute or characteristic A and B are independent i.e. there is no association or relationship between the two characteristics or attribute A and B, we use the following step for of independent of attributes. Similarly to the **TEST OF HOMOGENEITY** of two or more then two population we use the same process with slightly different hypothesis.

i. **Set up null hypothesis H_o:** two attributes (two categorical variable) A and B are independence i.e. there is no relationship or association between them.

ii. **Set up alternative hypothesis H_1:** two attributes (two categorical variable) A and B are dependence i.e. there is some relationship or association between them.

iii. Fix the level of significance α

iv. **Test statistics:** Under null hypothesis H_o, the test statistics is

$\chi^2 = \sum\frac{(O - e)^2}{e}$, follows chi-square distribution with (r-1) (c-1) .where O- observed frequency, e- expected frequency r- number of rows, c- numbers of columns.

The expected frequency in a cell is the product of its row total and column total divided by the total frequency, thus $e = \frac{RT \times CT}{N}$ where RT- row total, CT- column total and N is the grand total.

v. **Critical Value:** For α–level of significance and (r-1) (c-1) degrees of freedom, find the critical value of χ^2 from table as χ_α^2

vi. **Conclusion:** if calculated value of χ^2 is less than tabulated value of χ^2 we accept our null hypothesis and otherwise reject it.

15.6 χ^2 - TEST FOR 2 X 2 CONTINGENCY TABLE

$\chi 2$ –test for 2 x 2 Contingency Table can be computed alternatively as follows

i. **Set up null hypothesis H_o:** two attributes (two categorical variable) A and B are independence i.e. there is no relationship or association between them.

ii. **Set up alternative hypothesis H_1:** two attributes (two categorical variable) A and B are dependence i.e. there is some relationship or association between them.

iii. Fix the level of significance α

iv. **Test statistics:** Under null hypothesis H_o, the test statistics is

$\chi^2 = \frac{N(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$, follows chi-square distribution with (2-1) (2-1) = 1 degree of freedom.

Where a, b, c and d are cell frequencies and (a + b + c + d) = N is total frequency in the 2 x 2 Contingency Table.

2 x 2 Contingency Table

Attribute I	Attribute II		**Total**
	Column I	Column II	
Row I	a	b	a + b
Row II	c	d	c + d
Total	a + c	b + d	a + b + c + d = N

v. **Critical Value:** For α–level of significance and one degree of freedom, find the critical value of χ^2 from table as χ_α^2

vi. **Conclusion:** if calculated value of χ^2 is less than tabulated value of χ^2 we accept our null hypothesis and otherwise reject it.

Example 6

Following table provides data with regard to stature of the fathers and their first sons at the age of 25 years

Stature of son	Stature of father		
	Tall	Short	Total
Tall	8	12	20
Short	14	6	20
Total	22	18	40

Test that the stature of sons is independence of the stature of the fathers.

Solution

i. **Null hypothesis H_o:** the stature of sons is independence of the stature of the fathers.

ii. **Set up alternative hypothesis H_1:** the stature of son is dependence of the stature of the fathers.

iii. The level of significance α = 5% = 0.05

iv. **Test statistics:** Under null hypothesis H_o, the test statistics is

$\chi^2 = \Sigma\frac{(O-e)^2}{e}$, follows chi-square distribution with (r-1) (c-1) degrees of freedom,

Where O- observed frequency, e- expected frequency r- number of rows, c- numbers of columns.

The expected frequencies are calculated by using $e = \frac{RT \times CT}{N}$

Where RT- row total, CT-column total and N is the grand total.

$E(8) = \frac{20 \times 22}{40} = 11$, $E(12) = \frac{20 \times 18}{40} = 9$,

$E(14) = \frac{20 \times 22}{40} = 11$, $E(6) = \frac{20 \times 18}{40} = 9$

Calculation of value of χ^2

Cell (r, c)	Observed frequency(O)	Expected frequency(e)	(O-e)	$(O-e)^2$	$\frac{(o-e)^2}{e}$
(1,1)	8	11	-3	9	0.818
(1,2)	12	9	3	9	1.000
(2,1)	14	11	3	9	0.818
(2,2)	6	9	-3	9	1.000
Total	40	40			3.636

Thus, the calculated value of chi square is

$\chi^2 = \Sigma\frac{(O-e)^2}{e} = 3.636$

Alternatively the value of χ^2 is calculated as

$$\chi^2 = \frac{N(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$= \frac{40(8 \times 6 - 12 \times 14)^2}{(8+14)(12+6)(8+12)(14+6)}$$

= 3.636, which is same as earlier

v. **Critical Value:** For α = 5% level of significance and 1 degrees of freedom, the critical value of χ^2 from table is $\chi^2_{Tab} = 3.841$

vi. **Conclusion:** here calculated value of $\chi^2 = 3.636$ is less than tabulated value of $\chi^2_{Tab} = 3.841$ we accept our null hypothesis and we conclude that the stature of sons is independence of the stature of fathers.

15.7 YATE'S CORRECTION FOR CONTINUITY FOR 2 X 2 CONTINGENCY TABLE

When the expected frequency in a (2 x 2) contingency table is small, say less than 5 the continuity of χ^2 distribution is disturbed. To remove this deficiency Yates in 1934 suggested that 0.5 be added to the **small frequency** for which the expected frequency is less than 5 by adding and subtracting 0.5 in such the way that the marginal totals remains the same. After adjustment, the value of χ^2 is calculated in usual manner

An alternative formula is suggested which automatically takes care of adjustments in cell frequencies which is as follows

$$\chi^2 = \frac{N\left\{|ad-bc| - \frac{N}{2}\right\}^2}{(a+c)(b+d)(a+b)(c+d)}$$

Note that Yate's correction is applicable only n case of contingency table of (2 x 2)

Example 7

In an experiment on the immunization of dogs from certain diseases, the following results were obtained; derive your inference on the efficacy of the vaccine.

Immunization	Status of dog		
	Died	Survived	Total
Inoculated with vaccine	2	10	12
Not Inoculated	5	6	12
Total	8	16	24

Solution

i. **Null hypothesis H_o:** vaccine is not effective in the immunization of dogs from disease.

ii. **Set up alternative hypothesis H_1:** vaccine is effective in the immunization of dogs from disease.

iii. The level of significance α = 5% = 0.05

iv. **Test statistics:** Under null hypothesis H_o, the test statistics is

$\chi^2 = \Sigma\frac{(O-e)^2}{e}$, follows chi-square distribution with (r-1) (c-1) degrees of freedom,

The expected frequencies are calculated by using the following formula as $(e) = \frac{RT \times CT}{N}$

Where RT- row total, CT-column total and N is the grand total.

$E(2) = \frac{8 \times 12}{24} = 4$, $E(10) = \frac{12 \times 16}{24} = 8$,

$E(6) = \frac{8 \times 12}{24} = 4$, $E(6) = \frac{16 \times 12}{24} = 8$

By Applying Yate's Correction

Since the cell frequency 2 is less than 5 and expected frequencies are less than 5 so we apply Yates correction by adding 0.5 to the small value 2 and adjust the total frequency as

Calculation of value of χ^2

Cell (r, c)	Observed frequency(O)	Expected frequency(e)	(O-e)	$(O-e)^2$	$\frac{(o-e)^2}{e}$
(1,1)	2.5	4	-1.5	2.25	0.5625
(1,2)	9.5	8	1.5	2.25	0.2812
(2,1)	5.5	4	1.5	2.25	0.5625
(2,2)	6.5	8	-1.5	2.25	0.2812
Total	24	24	0		1.6875

Thus, the calculated value of $\chi^2 = \Sigma\frac{(O-e)^2}{e} = 1.6875$

Alternatively we will get the same result by applying the formula as

$$\chi^2 = \frac{N\left\{|ad - bc| - \frac{N}{2}\right\}^2}{(a+c)(b+d)(a+b)(c+d)}$$

$$= \frac{24\left\{|2 \times 6 - 10 \times 6| - 12\right\}^2}{(2+6)(10+6)(2+10)(6+6)} = 1.6875$$, which is same as earlier

v. **Critical Value:** For α = 5% level of significance and 1 degrees of freedom, the critical value of χ^2 from table is $\chi^2_{Tab} = 3.841$

vi. **Conclusion:** here calculated value of $\chi^2 = 1.6875$ is less than tabulated value of $\chi^2_{Tab} = 3.841$ we accept our null hypothesis and we conclude that the vaccine is effective in the immunization of dogs from disease.

Example 8

Two researchers adopted different sampling techniques while investigating the same group of students to find the number of students failing in different intelligence levels. The results are as follows.

Researcher	No of students in each level				
	Below average	Average	Above Average	Genius	Total
X	86	60	44	1	200
Y	40	33	25	2	100
Total	126	93	69	12	300

Would you say that the sampling techniques adopted by the two researchers are significantly different?

Solution

i. **Null hypothesis H_o:** there is no significant difference between the sampling techniques used by the two researchers for collecting the required data.

ii. **Alternative hypothesis H_1:** there is significant difference between the sampling techniques used by the two researchers for collecting the required data

iii. The level of significance α = 5% = 0.05

iv. **Test statistics:** Under null hypothesis H_o, the test statistics is $\chi^2 = \Sigma\frac{(O-e)^2}{e}$, follows chi-square distribution with (r-1) (c-1) d.f, where O- observed frequency, e- expected frequency r- number of rows, c- numbers of columns.

The expected frequencies are calculated by using $e = \frac{RT \times CT}{N}$ .where RT- row total, CT-column total and N is the grand total.

$E(86) = \frac{126 \times 200}{300} = 84$, $E(60) = \frac{93 \times 200}{300} = 62$

$E(44) = \frac{69 \times 200}{300} = 46$, $E(10) = \frac{200 \times 12}{300} = 8$

And E(40)= 126 – 84 = 42, E(33) = 93 – 62 = 31, E(25) = 69 – 46 =23, E(2) = 12 – 8 =4 Since $\Sigma O = \Sigma e = N$

Calculation of value of chi square

Types of students		observed frequency (O)	expected frequency (e)	(O-e)	$(O-e)^2$	$\frac{(O-e)^2}{e}$
X	Below average	86	84	2	4	0.048
	Average	60	62	-2	4	0.065
	Above Average	44	46	-2	4	0.087
	Genius	10	8	2	4	0.500
Y	Below average	40	42	-2	4	0.095
	Average	33	31	2	4	0.129
	Above Average	25 2 | 27	23 2 | 27	0	0	0.000
	Genius					
Total		300	300			0.923

Thus, the calculated value of $\chi^2 = \sum\frac{(O-e)^2}{e} = 0.923$

v. **Critical Value:** For $\alpha = 5\%$ level of significance and {(r-1) (c-1)} –1= {(4 –1) (2 –1)}-1 = 2 degrees of freedom, the critical value of χ^2 from table is $\chi^2_{Tab} = 5.991$ [since 1 d.f is lost due to pooling]

vi. **Conclusion:** here calculated value of $\chi^2 = 0.923$ is less than tabulated value of $\chi^2_{Tab} = 5.991$ we accept our null hypothesis and we conclude that there in no significant difference between the sampling techniques used by the two researchers for collecting the required data.

EXERCISE - 15

1. Define chi-square test. What are the applications of chi-square test?
2. What are the tests of significance of single population variance?
3. What is goodness of fit? What are the steps of test of significance of goodness of fit?
4. What is contingency table? What are the steps of test of significance of independence of attributes?

Numerical for test of goodness of fit

5. If random sample of size 15 yielded sample variance 0.00011.Can we concluded that these sample bolts are drawn from the normal population whose standard deviation diameter is 0.015 at 1% Level of significance.
6. Playing 10 round of golf on his home course. A golf professional averaged 71.3 with a standard deviation of 1.32. Test the hypothesis that the consistency of his game on his home course is actually measured by standard deviation $\sigma = 1.20$ at 5% level of significance.
7. Weights in kilograms of 10 students are given as:

   38, 40, 45, 53, 47, 43, 55, 48, 52 and 49. Can we say that variance of distribution of weights of all students from which the above sample of 10 students was drawn is equal to 20 square kilograms.
8. A random sample of size 20 from a population gives a sample of standard deviation of 6. Test the hypothesis that the population standard deviation is 9.
9. 200 digits are chosen at random from a set of tables. The frequencies of the digits were as.

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	13	25	22	20	21	18

ii. **Alternative hypothesis H_1:** there is significant difference between the sampling techniques used by the two researchers for collecting the required data

iii. The level of significance $\alpha = 5\% = 0.05$

iv. **Test statistics:** Under null hypothesis H_o, the test statistics is $\chi^2 = \Sigma\frac{(O-e)^2}{e}$, follows chi-square distribution with (r-1) (c-1) d.f, where O- observed frequency, e- expected frequency r- number of rows, c- numbers of columns.

The expected frequencies are calculated by using $e = \frac{RT \times CT}{N}$ .where RT- row total, CT-column total and N is the grand total.

$E(86) = \frac{126 \times 200}{300} = 84$, $E(60) = \frac{93 \times 200}{300} = 62$

$E(44) = \frac{69 \times 200}{300} = 46$, $E(10) = \frac{200 \times 12}{300} = 8$

And E(40)= 126 – 84 = 42, E(33) = 93 – 62 = 31, E(25) = 69 – 46 =23, E(2) = 12 – 8 =4 Since $\Sigma O = \Sigma e = N$

Calculation of value of χ^2

	Types of students	observed frequency(O)	expected frequency(e)	(O-e)	$(O-e)^2$	$\frac{(O-e)^2}{e}$
X	Below average	86	84	2	4	0.048
	Average	60	62	-2	4	0.065
	Above Average	44	46	-2	4	0.087
	Genius	10	8	2	4	0.500
Y	Below average	40	42	-2	4	0.095
	Average	33	31	2	4	0.129
	Above Average	25 / 2 | 27	23 / 2 | 27	0	0	0.000
	Genius					
Total		300	300			0.923

Thus, the calculated value of $\chi^2 = \Sigma\frac{(O-e)^2}{e} = 0.923$

v. **Critical Value:** For $\alpha = 5\%$ level of significance and {(r-1) (c-1)} –1= {(4 –1) (2 –1)}-1 = 2 degrees of freedom, the critical value of χ^2 from table is $\chi^2_{Tab} = 5.991$ [since 1 d.f is lost due to pooling]

vi. **Conclusion:** here calculated value of $\chi^2 = 0.923$ is less than tabulated value of $\chi^2_{Tab} = 5.991$ we accept our null hypothesis and we conclude that there in no significant difference between the sampling techniques used by the two researchers for collecting the required data.

Exercise - 15

1. Define chi-square test. What are the applications of chi-square test?
2. What are the tests of significance of single population variance?
3. What is goodness of fit? What are the steps of test of significance of goodness of fit?
4. What is contingency table? What are the steps of test of significance of independence of attributes?
5. If random sample of size 15 yielded sample variance 0.00011.Can we concluded that these sample bolts are drawn from the normal population whose standard deviation diameter is 0.015 at 1% Level of significance.
6. Playing 10 round of golf on his home course. A golf professional averaged 71.3 with a standard deviation of 1.32. Test the hypothesis that the consistency of his game on his home course is actually measured by standard deviation $\sigma = 1.20$ at 5% level of significance.
7. Weights in kilograms of 10 students are given as:

   38, 40, 45, 53, 47, 43, 55, 48, 52 and 49. Can we say that variance of distribution of weights of all students from which the above sample of 10 students was drawn is equal to 20 square kilograms.
8. A random sample of size 20 from a population gives a sample of standard deviation of 6. Test the hypothesis that the population standard deviation is 9.
9. 200 digits are chosen at random from a set of tables. The frequencies of the digits were as.

Digits	0	1	2	3	4	5	6	7	8	9
Frequency	18	19	23	21	12	25	22	20	21	15

   Test the hypothesis that the digits were distributed in equal numbers in the table.
10. The following table gives the number of aircraft accidents that occurred during the six days of the week. Find whether the accidents are uniformly distributed over the week.

Days	Mon	Tue	Wed	Thru	Fri	Sat
No. of accidents	14	18	12	11	15	14

11. A personnel manager is interested in trying to determine whether absenteeism is greater on one day of the week than on another. His records for the part year yield this sample distribution.

Day of Week	Mon	Tues	Wed	Thu	Fri
No. of absentees	66	57	48	54	75

Test whether the absence is uniformly distributed over the week.

12. Among 64 offspring of a certain cross between guinea pigs, 34 were red, 10 were black and 20 were white, According to genetic model these numbers should be in the ration 9:3:4. Are the data consistent with the model at 5 percent level of significance ?

13. A chemical extraction plant processed sea water to collect sodium chloride and magnesium, It is known that sea water contains sodium chloride, magnesium and other elements in the ration of 62:4:34. A sample of 200 tones of sea water has resulted in 130 tones of sodium chloride and 6 tones of magnesium. Are these consistent with the known composition of sea water at 5% level of significance ?

14. Genetic theory state that children having one parent of blood type A and the other of blood type B will always be of one of the three types A, AB and B are that the proportion of three types will be on average be as 1:2:1. A report states that out of 300 children having one A parent and one B parent 30% were found to be type A, 45% type AB and remaining type B. Test the hypothesis at 5% level of significance.

15. The theory predicts the proportion of beans, in the four groups A, B, C, and D should b e 9:3:3:1. In an experiment among 1600 beans, the numbers in the four groups were 882, 313, 287 and 118. Does the experimental result support the theory?

16. A dice is thrown 60 times with the following results. Test at 5% level of significance if the dice is unbiased or not.

Face	1	2	3	4	5	6
Frequency	8	7	12	8	14	11

17. Given below is the number of male and female birth in 1,000 families' having five children?

No. of male birth	No. of female birth	No of families
0	5	58
1	4	252
2	3	304
3	2	248
4	1	96
5	0	42

Test whether the given data is consistent with the hypothesis that the binomial law holds if the chance of a male birth is equal to that of female.

18. A set of 5 coins is tossed 3,200 times and the number of heads appearing each time is noted. The results are given below.

No. of heads	0	1	2	3	4	5
Frequency	80	570	1100	900	500	50

Test the hypothesis that the coins are unbiased

19. 192 families for each of which the probability of an albino child being born is otherwise established] have the following distribution of albinos among the first children.

No. Of albinos	0	1	2	3	Total
No. Of families	77	90	20	5	192

Find the expected frequencies on the basis of theoretical probability of a child being an albino and test the goodness of fit of the above data at 5% level of significance.

20. The following is the distribution of the hourly number of truck arriving at a company's warehouse. Fit the Poisson distribution at 5% level of significance.

Trucks arriving per hour	Frequency
0	52
1	151
2	130
3	102
4	45
5	12

6	5
7	1
8	2

21. The following is the distribution of the daily number of power failures reported in a Kathmandu valley on 300 days.

Number of power failure	Number of days
0	9
1	43
2	64
3	62
4	42
5	36
6	22
7	14
8	6
9	2

Test at the 0.05 level of significance whether the daily number of power failures in Kathmandu valley is a random variable having Poisson distribution with parameter $\lambda = 3.2$.

22. Two sample pools of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the following table. Examine whether the nature of the area is relate to voting preference in this election.

Area	Vote for	
	Candidate A	Candidate B
Rural	620	380
Urban	550	450

23. Out of 8,000 graduated in a town 800 are females. Out of 1600 graduate employer 120 are female. Use chi-square test determines if any distinction is made in appointment on the basis of sex.

24. 1,072 students were classified according to their intelligence and economic conditions. Test whether there is any association between intelligence and economic condition.

Economic Condition	Intelligence			
	Excellent	Good	Mediocre	Dull
Good	48	199	181	82
Not good	81	185	190	106

25. Brick made in two kilns have been graded as facings (high quality), seconds (medium quality) and common (rather poor quality). The production of bricks in a particular period was as follows.

Kiln	Facings '000	Seconds '000	Common '000
A	24	43	13
B	31	57	32

Test whether Kiln and quality of bricks are independent or not at 10% level of significance.

26. The results of a survey regarding 'radio listeners' preference for different types of music are given in the following table, with listeners classified by age group & preference of type of music influenced by age?

Types of music	Age group		
	19 – 25	26 – 35	above 36
Natural music	80	60	9
Foreign music	210	325	44
Indifferent	16	45	32

27. In a certain sample of 2,000 families, 1,400 families are consumers of tea out of 1,800 Hindu families, 1,236 families consume tea , Use chi-square test and state whether there is any significant difference between consumption of tea among Hindu and non – Hindu families.

28. Two groups A and B of certain type of patients each consisting of 200 people are used to test the effectiveness of a new serum. Both groups are treated identically except that group A is given the serum while group B is not. It is found that 140 and 120 of group A and B, respectively, recover from the disease. Does this observed result give sufficient evidence for the conclusion that the new serum helps to cure the disease if we were wiling to assume at 1% level of significance

29. A manufacturer of TV sets was trying to find out what variable influenced the purchased of TV set. A sample of 500 households was selected and the information obtained is classified as shown below:

Households	Purchase of TV	
	Have TV set	Don't have TV set
Low income group	0	250
Middle Income Group	50	100
High Income Group	80	20

Is there evidence from the above data of a relation between ownership of TV sets and level of income?

30. A large electronic firm that hires many handicapped workers wants to determine whether their handicaps affect such worker's' performance. Use the 5% level of significance to decide on the basis of the sample data shown in the following table whether it is reasonable to maintain that the handicaps have no effect on the workers performance

Types of Disability	Performance		
	Above average	Average	Below average
Blind	21	64	17
Deaf	10	49	14
No handicap	29	93	28

31. Test of the fidelity and the selectivity of 190 radio receivers produced the results shown in the following table. Use the 1% level of significance to test whether there is a relationship (dependence) between fidelity and selectivity

Selectivity	Fidelity		
	Low	Average	High
Low	6	12	32
Average	33	61	18
High	13	15	0

32. Following date show the durability of four kinds of tires. Use the 1% level of significance to test the null hypothesis that there is no difference in the quality of the four kinds of tire with regard to their durability.

Durability of Tires	Brand A	Brand B	Brand C	Brand D
Failed to last 20,000 miles	26	23	15	32
Lasted from 20,000 to 30, 000	118	93	116	121
Lasted more than 30, 000	56	84	69	47

33. Mechanical engineers testing a new arc welding technique, classified welds both with respect to appearance and an X-ray inspection. Test for independence using $\alpha = 5\%$.

X-ray	Appearance		
	Bad	Normal	Good
Bad	20	7	3
Normal	13	51	16
Good	7	12	21

34. Nepal Bank Ltd. has a new retirement plan. However, before implementing the plan, the bank wants to find out the reaction of the employees towards this plan. 200 employees of various stages were selected at random and their age groups and the corresponding opinions are tabulated as follows.

Opinion	Age group			
	20 -29	30-39	40 -50	Over 50
Good	18	36	14	12
Not good	32	24	36	28

At 5% level of significance, test if there is any relationship between age of the employees and their opinion about the retirement plan.

35. Samples of 400 contracts were classified according to two characteristics duration of contract and type of industry as in the following two way table

Types of Industry	Duration of Contact		
	2 Years or less	3 Years	4 Year or more
Manufacturing	10	187	48
Non manufacturing	13	107	35

Does the duration of a contract depend on the type of industry ? Answer by conduction a test for independence with $\alpha = 5\%$ level of significance.

36. A Campus newsletter polls a random sample of 350 undergraduate students and obtains the following frequency counts regarding a proposed change in library regulations.

Students	Responses		
	Favour	Indifference	Oppose
Male	93	72	21
Female	55	79	30

Does the proposal seem to appeal differently to male and female students test with 5% level of significance?