ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA

(Established by Act No. 19 of 2019 of the Legislature of State of Punjab)

The Motto of the University
(SEWA)

SKILL ENHANCEMENT EMPLOYABILITY WISDOM
ACCESSIBILITY

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ, ਪਟਿਆਲਾ
JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA

M.SC. (COMPUTER SCIENCE)
SEMESTER-I
Course: PROBABILITY & STATISTICAL ANALYSIS (MSCS-1-03T)

ADDRESS: C/28, THE LOWER MALL, PATIALA-147001
WEBSITE: www.psou.ac.in

JAGAT GURU NANAK DEV

PUNJAB STATE OPEN UNIVERSITY PATIALA

(Established by Act No.19 of 2019 of Legislature of the State of Punjab)

M.Sc. (Computer Science) Programme Coordinator:
Dr. Karan Sukhija (Assistant Professor)
School of Sciences and Emerging Technologies
JGND PSOU, Patiala

Faculty of School of Science and Emerging Technologies:
Dr. Baljit Singh Khera (Head)
Professor, School of Sciences and Emerging Technologies
Jagat Guru Nanak Dev Punjab State Open University, Patiala
Dr. Kanwalvir Singh Dhindsa
Professor, School of Sciences and Emerging Technologies
Jagat Guru Nanak Dev Punjab State Open University, Patiala
Dr. Amitoj Singh
Associate Professor, School of Sciences and Emerging Technologies
Jagat Guru Nanak Dev Punjab State Open University, Patiala
Dr. Monika Pathak
Assistant Professor, School of Sciences and Emerging Technologies
Jagat Guru Nanak Dev Punjab State Open University, Patiala
Faculty of School of Business Management & Commerce:
Dr. Pooja Aggarwal
Assistant Professor, School of Business & Commerce
Jagat Guru Nanak Dev Punjab State Open University, Patiala
Faculty of School of Social Sciences and Liberal Arts:
Dr. Pinky Sra
Assistant Professor, School of Social Sciences and Liberal Arts
Jagat Guru Nanak Dev Punjab State Open University, Patiala

\

ਜਗਤ ਗੁਰੂ ਨਾਨਕ ਦੇਵ
ਪੰਜਾਬ ਸਟੇਟ ਓਪਨ ਯੂਨੀਵਰਸਿਟੀ
ਪਟਿਆਲਾ

JAGAT GURU NANAK DEV
PUNJAB STATE OPEN UNIVERSITY, PATIALA
A State University Established by Govt. of Punjab vide Act No. 19
of 2019 and Approved Under section 2(f) of UGC

PROGRAMME COORDINATOR

Dr. Karan Sukhija (Assistant Professor)
School of Sciences and Emerging Technologies
JGND PSOU, Patiala

COURSE COORDINATOR & EDITOR

Dr. Pinky Sra
Assistant Professor, School of Social Sciences and Liberal Arts
Jagat Guru Nanak Dev Punjab State Open University, Patiala

Course: Probability & Statistical Analysis	
Course Code: MSCS-1-03T	
Course Outcomes (COs) After the completion of this course, the students will be able to:	
CO1	Apply measures of central tendency for analysis of data.
CO2	Learn tabulated and graphical representation techniques for discrete and continuous data.
CO3	Infer the concept of correlation and regression for two or more related variables.
CO4	Understand the fundamentals of statistics to apply descriptive measures and probability for data analysis.
CO5	Understand the concepts of Random Variable, Probability Mass Function and Density Function.

JAGAT GURU NANAK DEV

PUNJAB STATE OPEN UNIVERSITY PATIALA

(Established by Act No.19 of 2019 of Legislature of the State of Punjab)

PREFACE

Jagat Guru Nanak Dev Punjab State Open University, Patiala was established in Decembas 2019 by Act 19 of the Legislature of State of Punjab. It is the first and only Open Universit of the State, entrusted with the responsibility of making higher education accessible to all especially to those sections of society who do not have the means, time or opportunity to pursue regular education.

In keeping with the nature of an Open University, this University provides a flexible education system to suit every need. The time given to complete a programme is double the duration of a regular mode programme. Well-designed study material has been prepared in consultation with experts in their respective fields.

The University offers programmes which have been designed to provide relevant, skill-based and employability-enhancing education. The study material provided in this booklet is self instructional, with self-assessment exercises, and recommendations for further readings. The syllabus has been divided in sections, and provided as units for simplification.

The Learner Support Centres/Study Centres are located in the Government and Government aided colleges of Punjab, to enable students to make use of reading facilities, and for curriculum-based counselling and practicals. We, at the University, welcome you to be a part of this institution of knowledge.

Prof. G. S. Batra,
Dean Academic Affairs

Name of Programme: M.Sc. (Computer Science)
Name of Course: Probability & Statistical Analysis
Course Code: MSCS-1-03T
Semester 1

Total Marks: 100
External Marks: 70
Internal Marks: 30
Credits: 4
Pass Percentage: 40%

INSTRUCTIONS FOR THE PAPER SETTER/EXAMINER

1. The syllabus prescribed should be strictly adhered to.
2. The question paper will consist of three sections: A, B, and C. Sections A and B will have four questions from the respective sections of the syllabus and will carry 10 marks each. The candidates will attempt two questions from each section.
3. Section C will have fifteen short answer questions covering the entire syllabus. Each question will carry 3 marks. Candidates will attempt any ten questions from this section.
4. The examiner shall give a clear instruction to the candidates to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.
5. The duration of each paper will be three hours.

INSTRUCTIONS FOR THE CANDIDATES

Candidates are required to attempt any two questions each from the sections A and B of the question paper and any ten short q questions from Section C. They have to attempt questions only at one place and only once. Second or subsequent attempts, unless the earlier ones have been crossed out, shall not be evaluated.

SECTION-A

Unit I: Origin and Development of Statistics: Scope, limitation and misuse of statistics. Types of data: primary, secondary, quantitative and qualitative data. Types of Measurements: nominal, ordinal, discrete and continuous data.

Unit II: Presentation of Data by Tables: construction of frequency distributions for discrete and continuous data, graphical representation of a frequency distribution by histogram and frequency polygon, cumulative frequency distributions. Classification and Graphical representation of data (Pie Chart, Bar Diagram, Histogram, Frequency Polygon, Ogive Curve, etc.).

Unit III: Measures of Central Tendency: Arithmetic Mean, Median and Mode and its Graphical representation, Measures of dispersion – range, variance, mean deviation, standard deviation and Coefficient of variation, Concepts and Measures of Skewness and Kurtosis.

Unit IV: Descriptive Statistics: Exploratory data analysis, Coefficient of variation, Data visualization, Scatter diagram, Grouped data. 27

SECTION-B

Unit V: Correlation: Scatter plot, Karl Pearson coefficient of correlation, Spearman's rank correlation coefficient, multiple and partial correlations (for 3 variates only). Regression: Introduction to regression analysis: Modelling a response, overview and applications of regression analysis, Simple linear regression (Two variables)

Unit VI: Mathematical and Statistical probability: Elementary events, Sample space, Compound events, Types of events, Random experiment, sample point and sample space, event, algebra of events.

Unit VII: Definition of Probability: classical, empirical and axiomatic approaches to probability, properties of probability. Theorems on probability, conditional probability and independent events

Unit VIII: Statistical inference: Concept of Random Variable, Probability Mass Function & Density Function, Mathematical Expectation (meaning and properties), Moments, Moment Generating Function and Characteristic Function.

Reference Books:

Gupta, S.C. and Kapoor, V.K., “Fundamentals of Mathematical Statistics”, Sultan & Chand & Sons, New Delhi, 11th Ed.

Hastie, Trevor, et al. “The elements of Statistical Learning”, Springer.

Ross, S.M., “Introduction to Probability and Statistics”, Academic Foundation.

Papoulis, A. and Pillai, S.U., “Probability, Random Variables and Stochastic Processes”, TMH.

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

UNIT I: ORIGIN AND DEVELOPMENT OF STATISTICS

STRUCTURE

1.0 OBJECTIVES

In this module, we will try to understand about the historical development of statistics. Various definitions given by different scientists will also be explored in this module. Further need of statistics along with its applications will also be investigated. Limitations and misuse of statistics will also be looked upon. In the end, various types of data used in statistics will also be explored.

1.1 INTRODUCTION

This module is designed to know about the development of statistics using historical background. Statistics is not a subject that can be studied alone, rather it proves to be basis for almost all other subjects as data handling is essential in almost all fields of life. Due to this reason, a number of definitions are given to statistics. Some of the major fields where statistics is prominently used are planning, finance, business, agriculture, biology, economics, industry, education, etc. Actually a country"s growth is very much dependent on statistics as without statistics it would not be possible to estimate the requirements of the country.However, statistics is based on probabilistic estimations and therefore not actual (in some cases), therefore can"t be believed with 100% guarantee. Also some people may misuse statistics for their own benefits. But, still statistics is very essential and very much need of thelife. There are various classifications of the data used in statistics viz., continuous, discrete, nominal, ordinal, etc. The data can be used as per requirement for a particular application.

1.2 MAIN CONTENT

1.2.1 Origin and Development of Statistics

The term Statistics, is not a new term rather it is as old as being human society. However, it has come up leaps and bounds with the emergence of data analytics and machine learning and the role of statistics in those areas. Directly and indirectly, statistics has been used right from the ancient days, although as a subject it was introduced much later.

In the earlier days it was regarded more or less as a by-product of the administrative activity for analysing the various activities. The word Statistics is said to be derived from a Latin word „status" or an Italian word „statista" or may be the German word „statistik" or the even from the French word „statistique", each of which actually means a political state. The term was used to collect information about the population in a country so that various schemes could be implemented depending on the requirement and size of population.

If we talk specifically about India, various efficient mechanisms for collecting various type of official statistics existed even 2000 years ago, during the tenure of Chandragupta Maurya. In history, evidences have been found for the existence of an excellent system of collecting important statistics and registration of births and deaths, even before 300 B.C. are available in Kautilya"s „Arthashastra". Further, the records of land, agriculture and wealth statistics were maintained by Todermal, a well-known land and revenue minister in the era of Akbar (1556-1605 A.D). A detailed information about the administrative and statistical surveys conducted during Akbar"s era is very much available in the book “Ain-e- Akbari” written by Abul Fazl (1596-97), one of the nine gems of Akbar.

The Statistics was applied for collecting the data related to the movements of heavenly bodies in the 16th century, viz., stars and planets so that their position may be known and various Eclipses may be predicted. Further, in the seventeenth century, the vital statistics was originated. Captain John Graunt of London (1620-1674) is known as the Father of vital statistics. He was the person behind a systematic study of the birth and death statistics.

Modern stalwarts involved in the development of the subject of Statistics, are various Englishmen, who did revolutionary work in applying the Statistics to different disciplines. Francis Galton (1822-1921) pioneered the study of „Regression Analysis" in Biometry; Karl Pearson (1857-1936), founder of the greatest statistical laboratory in England pioneered the study of „Correlation Analysis". His Chi-Square test (X^2-test) of Goodness of Fit is one of the most important tests of significance in Statistics; W.S. Gosset introduced t-test, which escorted the era of exact (small) sample tests.

However most of the work in the statistical theory during the past few decades can be attributed to a single person Sir Ronald A. Fisher (1890-1962), who applied statistics to a variety of diversified fields such as genetics, biometry, psychology and education, agriculture, etc., and who is rightly termed as the Father of Statistics. He not only enhanced the existing statistical theory, but also he is the pioneer in Estimation Theory; Exact (small) Sampling Distributions; Analysis of Variance and Design of Experiments. One can easily saythat R.A. Fisher is the real giant in the development of the theory of Statistics. It is due to the outstanding contributions of R. A. Fisher that put the subject of Statistics on a very firm footing and earned for it the status of a full-fledged science.

1.2.2 Definitions of Statistics

Statistics has been defined by number of authors in different ways. The main reason for the various definitions are the changes that has taken place in statistics from time to time. Statistics in general is defined in two different ways viz., as „statistical data", i.e., based on numerical statement of data and facts, and as 'statistical methods', i.e., based on the principles and techniques used in collecting and analysing such data. Some of the important definitions under these two categories are given below.

Statistics as Statistical data

Webster defines Statistics as "classified facts representing the conditions of the people in a State, especially those facts which can be stated in numbers or in any other tabular or classified arrangement." Bowley defines Statistics as "numerical statements of facts in any department of enquiry placed in relation to each other."

A more exhaustive definition is given by Prof. Horace Secrist as follows: "By statistics we mean aggregation of facts affected to a marked extent by multiplicity of causes numerically expressed, enumerated or estimated according to reasonable standards of accuracy, collected in a systematic manner for a predetermined purpose and placed in relation to each other."

Statistics as Statistical Methods

Bowley himself has defined Statistics in a number of ways:

(i) Statistics may be called the science of counting.

(ii) Statistics may rightly be called the science of averages.

(iii) Statistics is the science of the measurement of social organism, regarded as a whole in all its manifestations.

However, these definitions are not complete in any sense as they don‟t provide the complete view of statistics. According to Boddington, "Statistics is the science of estimates and probabilities." Again this definition is not complete as statistics is not just probabilities and estimates but more than that.

Some other definitions are: "The science of Statistics is the method of judging collective, natural or social phenomenon from the results obtained from the analysis or enumeration or collection of estimates."- as provided by King.

"Statistics is the science which deals with collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon." as given by Lovitt.

But the best definition is the one given by Croxton and Cowden, according to whom Statistics may be defined as "the science which deals with the collection, analysis and interpretation of numerical data."

1.2.3 Importance and Scope of Statistics

Statistics is primarily used either to make predictions based on the data available or to make conclusions about a population of interest when only sample data is available. In both cases statistics tries to make sense of the uncertainty in the available data. When making predictions statisticians determine if the difference in the data points are due to chance or if there is a systematic relationship. The more the systematic relationship that is observed the better the prediction a statistician can make. The more random error that is observed the moreuncertain the prediction.

Statisticians can provide a measure of the uncertainty to the prediction. When making inference about a population, the statistician is trying to estimate how good a summary statistic of a sample really is at estimating a population statistic.

For computer students, knowing the basic principles and methods in statistics could help them in doing their research work like comparing the speed of internet connection in different countries and the probability of how many times does each experience the same level of internet connection speed in a week, month or year. It could also be helpful in determining the best operating system to use. Whenever there is the need to compare data and know the best option that we should take statistics can give the answer.

Statistics is having applications in almost all sciences - social as well as physical such as biology, psychology, education, economics, business management, etc. It is hardly possible to think of even a single department of human activity where statistics is not involved. It has rather become indispensable in all phases of human endeavour.

Statistics and Planning

Statistics is mother of planning. In the modern age which is termed as 'the age of planning', almost all over the world, particularly of the upcoming economies, are resorting to planning for the economic development. In order that planning is successful, it must be based soundly on the correct analysis of complex statistical data.

Statistics and Economics

Statistical data and technique of statistical analysis have proved immensely useful in solving various economic problems, such as wages, prices, analysis of time series and demandanalysis. A number of applications of statistics in the study of economics have led to the development of new disciplines called Economic Statistics and Econometrics.

Statistics and Business

Statistics is an essential tool for production control. Statistics not only helps the business executives to know the requirements of the consumers, but also for many other purposes. The success of a business actually depends upon the accuracy and precision of his statistical forecasting. Wrong analysis, due to faulty and inaccurate analysis of various causes affecting a particular phenomenon, might prove to be a disaster. Consider an examples of manufacturing readymade garments. Before starting one must have an overall idea as to „howmany garments are to be manufactured', 'how much raw material and labour is needed for that', and 'what is the quality, shape, color, size, etc., of the garments to be manufactured'. If these questions are not analysed statistically in a proper manner, the business is bound to be failed. Therefore, most of the large industrial and commercial enterprises are employing trained and efficient statisticians.

Statistics and Industry

In industry, statistics is very widely used in 'Quality Control'. In production engineering, to find whether the product is conforming to specifications or not, statistical tools, viz. inspection plans, control charts, etc., are of extreme importance.

Statistics and Mathematics

Statistics and mathematics arc very intimately related. Recent advancements in statistical techniques are the outcome of wide applications of advanced mathematics. Main contributors to statistics, namely, Bernouli, Pascal, Laplace, De-Moirve, Gauss, R. A. Fisher, to mention only a few, were primarily talented and skilled mathematicians. Statistics may be regarded as that branch of mathematics which provided us with systematic methods of analysing a large number of related numerical facts. According to Connor, " Statistics is a branch of Applied Mathematics which specialises in data."

Statistics and Biology, Astronomy and Medical Science

The association between statistical methods and biological theories was first studied byFrancis Galton in his work in Regression. According to Prof. Karl Pearson, the whole 'theory of heredity' rests on statistical basis. He said, "The whole problem of evolution is a problemof vital statistics, a problem of longevity, of fertility, of health, of disease and it is impossible to discuss the national mortality without an enumeration of the population, a classification of deaths and knowledge of statistical theory." In astronomy, the theory of Gaussian 'Normal Law of Errors' for the study of the movement of stars and planets is developed by using the 'Principle of Least Squares'. In medical science also, the statistical tools for the collection, presentation and analysis of observed facts relating to the causes and diseases and the results obtained from the use of various drugs and medicines, are of great importance. Moreover, theefficacy of a manufactured drug or injection or medicine is tested by analysing the 'tests of significance'.

Statistics and Psychology and Education

In education and psychology, too, statistics has found wide applications, e.g., to determine the reliability and validity of a test, 'Factor Analysis', etc., so much so that a new subject called 'Psychometry' has come into existence.

Statistics and War

In war, the theory of 'Decision Functions' can be of great assistance to military and technical personnel to plan 'maximum destruction with minimum effort'. Thus, we see that the science of Statistics is associated with almost all the sciences - social as well as physical. Bowley has rightly said, "A knowledge of Statistics is like a knowledge of foreign language or algebra; it may prove of use at any time under any circumstance."

1.2.4 Limitations of Statistics

Statistics, with its wide applications in almost every sphere of human activity; is not without limitations. The following are some of its important limitations:

(i) **Statistics is not suited to the study of qualitative phenomenon.** Statistics, being a science dealing with a set of numerical data, is applicable to the study of only those subjects of enquiry which are capable of quantitative measurement. As such; qualitative phenomena like honesty, poverty, culture, etc., which cannot be expressed numerically, are not capable of direct statistical analysis. However, statistical techniques may be applied indirectly by first reducing the qualitative expressions to precise quantitative terms. For example, the intelligence of a group of candidates can be studied on the basis of their scores in a certain test.

(ii) **Statistics does not study individuals.** Statistics deals with an aggregate of objects and does not give any specific recognition to the individual items of a series. Individual items, taken separately, do not constitute statistical data and are meaningless for any statistical enquiry. For example, the individual figures of agricultural production, industrial output or national income of any country for a particular year are meaningless unless, to facilitate comparison, similar figures of other countries or of the same country for different years are given. Hence,statistical analysis is suited to only those problems where group characteristics areto be studied.

(iii) **Statistical laws are not exact.** Unlike the laws of physical and natural sciences, statistica1laws are only approximations and not exact. On the basis of statistical analysis, we can talk only in terms of probability and chance and not in terms of certainty. Statistical conclusions are not universally true, rather they are true only on an average.

1.2.5 Misuse of Statistics

Statistics is liable to be misused. As they say, "Statistical methods are the most dangerous tools in the hands of the in experts. Statistics is one of those sciences whose adepts must exercise the self-restraint of an artist." The use of statistical tools by inexperienced and untrained persons might lead to very fallacious conclusions. One of the greatest shortcomingsof statistics is that by just looking at them one can‟t comment about their quality and as such can be represented in any manner to support one's way of argument and reasoning. As King said, "Statistics are like clay of which one can make a god or devil as one place." The requirement of experience and judicious use of statistical methods restricts their use toexperts only and limits the chances of the mass popularity of this useful and important science.

It may be pointed out that Statistics neither proves anything nor disproves anything. It is only a tool which if rightly used may prove extremely useful and if misused might be disastrous. According to Bowley, "Statistics only furnishes a tool necessary though imperfect, which is dangerous in the hands of those who do not know its use and its deficiencies." It is not the statistics which can be blamed but those persons who twist the numerical data and misuse them either due to ignorance or deliberately for personal selfish motives. As King pointed out, "Science of Statistics is the most useful servant but only of great value to those who understand its proper use."

A few interesting examples showing the impact of misrepresentation of statistical data are:

(i) A statistical report, "The number of accidents taking place in the middle of the road is much less than the number of accidents taking place on its side. Hence it issafer to walk in the middle of the road." This conclusion is obviously wrong since we are not given the proportion of the number of accidents to the number of persons walking in the two cases.

(ii) Another saying that, "The number of students taking up Computer Science in a University has increased 5 times during the last 3 years. Thus, Computer Science is gaining popularity among the students of the university." Again, the conclusion is faulty since we are not given any such details about the other subjects and hence comparative study is not possible.

(iii) One more interesting examples says that, "99% of the people who drink alcohol die before attaining the age of 100 years. Hence drinking is harmful for longevity of life." This statement, too, is incorrect since nothing is mentioned about the number of persons who do not alcohol and die before attaining the age of 100 years. Thus, statistical arguments based on incomplete data often lead to fallaciousconclusions.

1.2.6 Types of Data

In statistics, the data are the individual pieces of factual information recorded, and it is used for the purpose of the analysis process. The two processes of data analysis are interpretation and presentation. Statistics are the result of data analysis. Data classification and data handling are an important process as it involves a multitude of tags and labels to define the data, its integrity and confidentiality. The data can be classified as shown in figure 1.1 and has been described as follows:

Qualitative or Categorical Data

Qualitative data, also known as the categorical data, describes the data that fits into the categories. Qualitative data are not numerical. The categorical information involves categorical variables that describe the features such as a person‟s gender, home town etc. Categorical measures are defined in terms of natural language specifications, but not in terms of numbers.

Sometimes categorical data can hold numerical values (quantitative value), but those values do not have mathematical sense. Examples of the categorical data are birthdate, favourite sport, school postcode. Here, the birthdate and school postcode hold the quantitative value, but it does not give numerical meaning. It can be further classified as nominal and ordinal data.

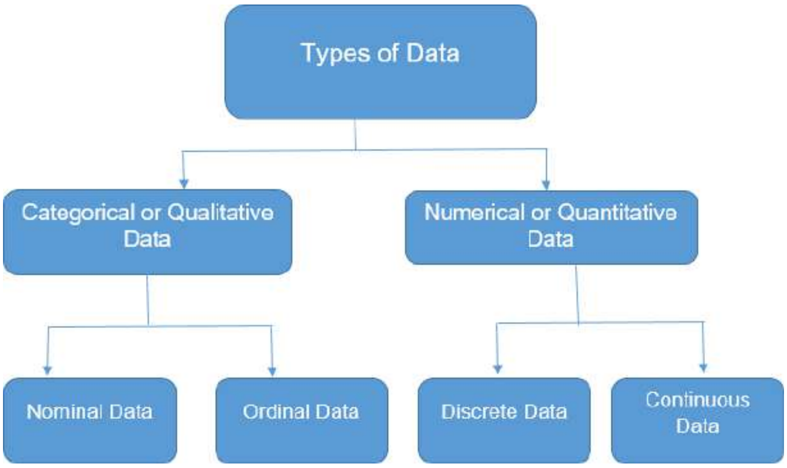


Figure 1.1: Classification of Data used in Statistics

Nominal Data: Nominal data is one of the types of qualitative information which helps to label the variables without providing the numerical value. Nominal data is also called the nominal scale. It cannot be ordered and measured. But sometimes, the data can be qualitative and quantitative. Examples of nominal data are letters, symbols, words, gender etc.

The nominal data are examined using the grouping method. In this method, the data are grouped into categories, and then the frequency or the percentage of the data can be calculated. These data are visually represented using the pie charts.

Ordinal Data: Ordinal data is a type of data which follows a natural order. The significant feature of the nominal data is that the difference between the data values is not determined. This variable is mostly found in surveys, finance, economics, questionnaires, and so on.

The ordinal data is commonly represented using a bar chart. These data are investigated and interpreted through many visualisation tools. The information may be expressed using tables in which each row in the table shows the distinct category.

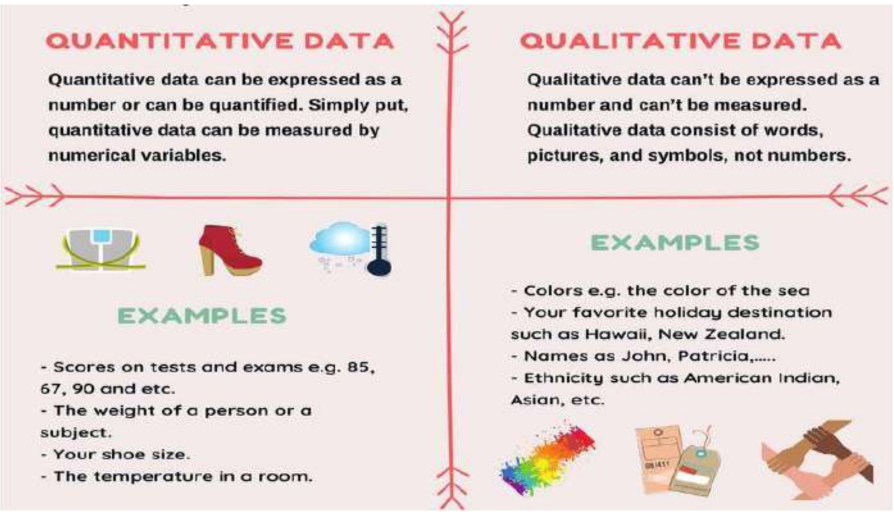
Quantitative or Numerical Data

Quantitative data is also known as numerical data which represents the numerical value (i.e., how much, how often, how many). Numerical data gives information about the quantities of a specific thing. Some examples of numerical data are height, length, size, weight, and so on. The quantitative data can be classified into two different types based on the data sets. The two different classifications of numerical data are discrete data and continuous data.

Discrete Data: Discrete data can take only discrete values. Discrete information contains only a finite number of possible values. Those values cannot be subdivided meaningfully. Here, things can be counted in the whole numbers e.g. Number of students in the class

Continuous Data: Continuous data is data that can be calculated. It has an infinite number of probable values that can be selected within a given specific range e.g. Temperature range.

The quantitative and qualitative data can be represented as in figure 1.2.



QUANTITATIVE DATA

Quantitative data can be expressed as a number or can be quantified. Simply put, quantitative data can be measured by numerical variables.

EXAMPLES

- Scores on tests and exams e.g. 85, 67, 90 and etc.
- The weight of a person or a subject.
- Your shoe size.
- The temperature in a room.

QUALITATIVE DATA

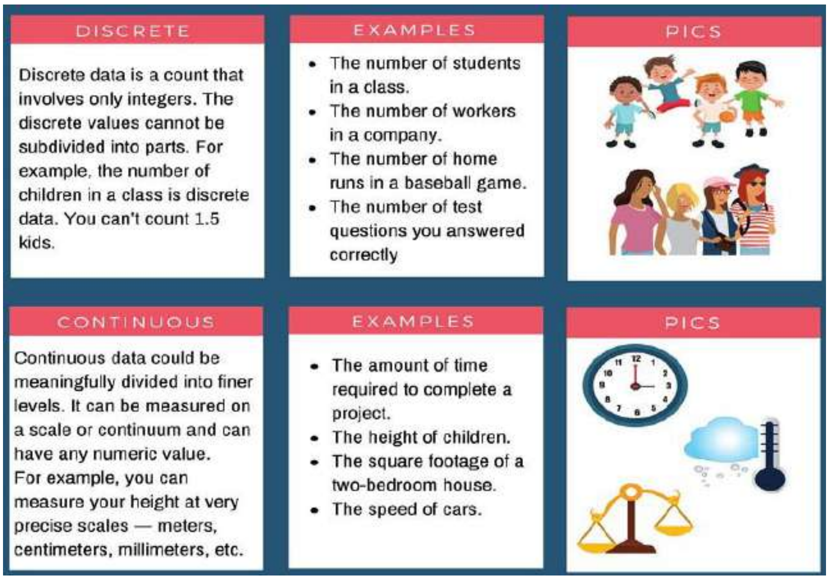
Qualitative data can't be expressed as a number and can't be measured. Qualitative data consist of words, pictures, and symbols, not numbers.

EXAMPLES

- Colors e.g. the color of the sea
- Your favorite holiday destination such as Hawaii, New Zealand.
- Names as John, Patricia,.....
- Ethnicity such as American Indian, Asian, etc.

Figure 1.2: Quantitative and Qualitative Data

Figure 1.3 shows the types of qualitative data i.e. discrete and continuous data.



DISCRETE

Discrete data is a count that involves only integers. The discrete values cannot be subdivided into parts. For example, the number of children in a class is discrete data. You can't count 1.5 kids.

EXAMPLES

- The number of students in a class.
- The number of workers in a company.
- The number of home runs in a baseball game.
- The number of test questions you answered correctly

PICS

CONTINUOUS

Continuous data could be meaningfully divided into finer levels. It can be measured on a scale or continuum and can have any numeric value. For example, you can measure your height at very precise scales — meters, centimeters, millimeters, etc.

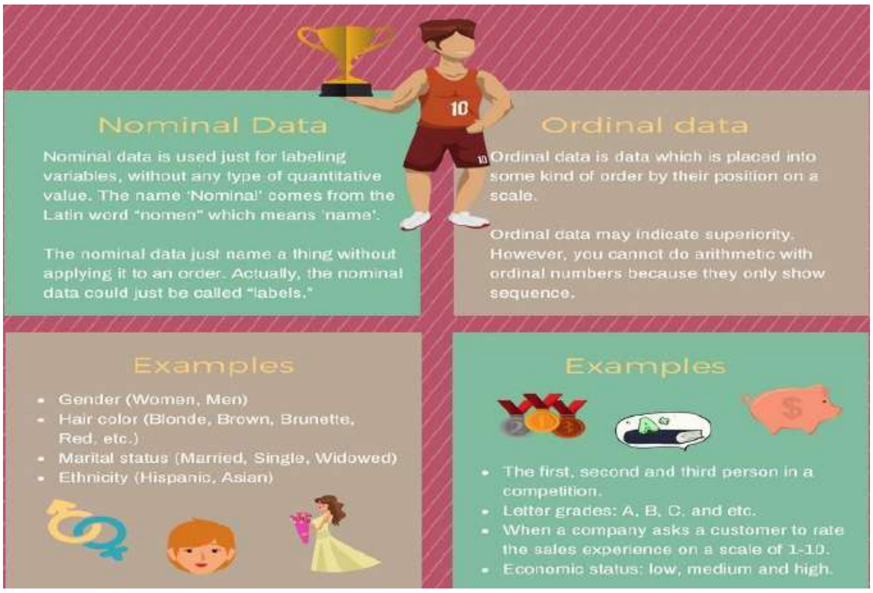
EXAMPLES

- The amount of time required to complete a project.
- The height of children.
- The square footage of a two-bedroom house.
- The speed of cars.

PICS

Figure 1.3: Types of Qualitative Data viz., Discrete and Continuous

Figure 1.4 shows the types of quantitative data i.e. nominal and ordinal data.



Nominal Data

Nominal data is used just for labeling variables, without any type of quantitative value. The name 'Nominal' comes from the Latin word "nomen" which means 'name'.

The nominal data just name a thing without applying it to an order. Actually, the nominal data could just be called "labels."

Ordinal data

Ordinal data is data which is placed into some kind of order by their position on a scale.

Ordinal data may indicate superiority. However, you cannot do arithmetic with ordinal numbers because they only show sequence.

Examples

- Gender (Women, Men)
- Hair color (Blonde, Brown, Brunette, Red, etc.)
- Marital status (Married, Single, Widowed)
- Ethnicity (Hispanic, Asian)

Examples

- The first, second and third person in a competition.
- Letter grades: A, B, C, and etc.
- When a company asks a customer to rate the sales experience on a scale of 1-10.
- Economic status: low, medium and high.

Figure 1.4: Types of Quantitative Data viz., Nominal and Ordinal

1.2.7 Data Collection

Depending on the source, it can be classified as primary data or secondary data. Let us take a look at them both.

Primary Data

These are the data that are collected directly by an investigator for a specific purpose. Primary data are „pure" in the sense that no statistical operations have been performed on them and they are original. An example of primary data is the Census of India.

Secondary Data

They are the data that are sourced from someplace that has originally collected it.This means that this kind of data has already been collected by some researchers or investigators in the past and is available either in published or unpublished form. Thisinformation is impure as statistical operations may have been performed on them already. An example is an information available on the Government of India, the Department of Finance‟s website or in other repositories, books, journals, etc.

1.3 SUMMARY

In this module, the overall development and history of statistics has been discussed. Various definitions given be various authors have been provided and discussed. Following this, the applications along with advantages and limitations of statistics have also been discussed in detail. One of the very important aspect, misuse of statistics has also been explored along with few examples for misuse of the statistics. Then the various classifications of statistical data have also been discussed in detail. In the end, the types of data collection methods have been discussed in brief. Overall, this module provides an overview of what is statistics along with its applications in depth.

1.4 QUESTIONS FOR PRACTICE

- Give a historical background of statistics.
- Write various definitions of statistics and discuss these definitions in brief.
- State and explain various applications of statistics.
- What are the various limitations of statistics?
- "Statistics don‟t lie". Comment on this statement.
- Provide a few examples which can lead to incorrect conclusion due to wrong analysis of statistics.
- Give any two examples of collecting data from day-to-day life.
- How can you classify the statistical data?
- Categorize the following data in various types: (i) Speed (ii) Gender (iii) Height (iv) Grades (v) No. of Employees (vi) Time (vii) Colour (viii) Score (ix) Weight.
- The word statistics seems to have been derived from which word?
- From__________it is known that even before 300 B.C. a very good system of collecting "Vital Statistics" and registration of births and deaths was in vogue____ .
- Who is known as the father of "Vital Statistics"?

REFERENCES

- A. Abebe, J. Daniels, J.W.Mckean, "Statistics and Data Analysis".
- Clarke, G.M. & Cooke, D., "A Basic course in Statistics", Arnold.
- David M. Lane, "Introduction to Statistics".
- S.C.Gupta and V.K.Kapoor, "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, New Delhi.

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

UNIT II: PRESENTATION OF DATA BY TABLES

STRUCTURE

2.0 OBJECTIVES

In this module, we will try to understand about the representation of data through tables. The various types of distribution of data and tabular presentation will be discussed in this module. Further, graphical representation of data through some popular methods such as histogram and frequency polygons will also be discussed in detail. Thereafter, the classification and graphical representation of statistical data through various methods will be explained. Further, we shall also learn the various types of representations (both tabular and graphical) of data in Python.

2.1 INTRODUCTION

This module is designed to know about the representation of data in tabular and graphical forms. For analysing the statistical data, it must be represented in a tabular form and this module does the same, i.e., describe the techniques to convert the data in tabular forms. For the purpose of planning and interpreting the data, visual effects are very useful and necessary.The visual effects in statistics can be obtained by representing the data through graphs. In thismodule, various types of graphs, viz., histogram, frequency polygons, bar graphs, pie charts, ogives, etc. have been discussed. Further, an automated tool for representing such type ofdata must be used for quick analysis and better representation. One such tool Python is usedin this chapter to represent the data.

2.2 MAIN CONTENT

2.2.1 Presentation of Data by Tables

Whenever we want to analyse and interpret the data, it can be done in an effective manner only when it is represented in tabular and/or graphical manner. The data in tabular form can be represented by using frequency distributions as explained in the following section.

2.2.2 Construction of Frequency Distribution

Frequency Distributions

When observations, discrete or continuous, are available on a single characteristic of a large number of individuals, often it becomes necessary to condense the data as far as possible without losing any information of interest. For condensing the data, it is represented using either discrete or continuous frequency distribution tables.

Discrete frequency distribution: In discrete frequency distribution, values of the variable is arranged individually. The frequencies of the various values are the number of times each value occurs. For examples, the weekly wages paid to the workers are given below.

300, 240, 240, 150, 120, 240, 120, 120, 150, 150, 150, 240, 150, 150, 120, 300, 120, 150, 240, 150, 150, 120, 240, 150, 240, 150, 120, 120, 240, 150.

There are various ways to form a frequency distribution for this data. In the first case, let us assume that data is represented in terms of tally marks in a tabular manner as shown below in the table 2.1:

Table 2.1: Representation of Data using Tally Marks

Weekly Wages	Tally Marks	No. of Workers
120	~~IIII~~ III	8
150	~~IIII~~ ~~IIII~~ II	12
240	~~IIII~~ III	8
300	II	2

This data can also be represented without using tally marks i.e. using frequency only as shown in table 2.2 and is known as frequency table.

Table 2.2: Frequency Table for the Data in Table 2.1

Weekly Wages (x)	120	150	240	300	Total
No. of Workers (f)	8	12	8	2	30

The frequency table 2.1 is ungrouped frequency table. We can also draw a grouped frequency table depending on the data we are having. For designing a grouped frequency table, let us consider the following example regarding daily maximum temperatures in in a city for 50 days.

28, 28, 31, 29, 35, 33, 28, 31, 34, 29, 25, 27, 29, 33, 30, 31, 32, 26, 26, 21, 21, 20, 22, 24, 28, 30, 34, 33, 35, 29, 23, 21, 20, 19, 19, 18, 19, 17, 20, 19, 18, 18, 19, 27, 17, 18, 20, 21, 18, 19.

Table 2.3: Grouped Frequency Table

Temperature	Frequency
17-21	17
22-26	9
27-31	13
32-36	11
Total	50

The classes of type 17-21 and 22-26 are inclusive in nature i.e. both the lower bound and upper bound are included in the limit.

Although there are no hard and fast rules that have been laid down for it The following points may be kept in mind for classification:

(i) The classes should be clearly defined and should not lead to ambiguity.

(ii) The classes should be exhaustive, i.e., each of the given values should be included in one of the classes.

(iii) The classes should be mutually exclusive and non-overlapping.

(iv) The classes should be of equal width. The principle, however, cannot be rigidly followed.

(v) Indeterminate classes, e.g., the open-end classes such as less than 'a' or greater than 'b' should be avoided as far as possible since they create difficulty in analysis and interpretation.

(vi) The number of classes should neither be too large nor too small. It should preferably lie between 5 and 15. However. the number of classes may be more than 15 depending upon the total frequency and the details required. But it is desirable that it is not less than 5 since in that case the classification may not reveal the essential characteristics of the population.

Terms used in frequency distribution

Class Interval: The whole range of variable values is classified in some groups in the form of intervals. Each interval is called a class interval.

Class Frequency: The number of observations in a class is termed as the frequency of the class or class frequency.

Class limits and Class boundaries: Class limits are the two endpoints of a class interval which are used for the construction of a frequency distribution. The lowest value of the variable that can be included in a class interval is called the lower class limit of that class interval. The highest value of the variable that can be included in a class interval is called the upper-class limit of that class interval. In the table 2.3, the class intervals are 17-21, 22-26, 27-31 and 32-36. Here, say for the class 17-21, the lower-class limit is 17 and the upper-class limit is 21. Both 17 and 21 are part of this class. This is called inclusive class. Another typeof class is exclusive class as shown below in table 2.4:

Table 2.4: Exclusive Class Grouped Frequency Table

Temperature	Frequency
17-21	17
21-25	7
25-29	10
29-33	9
33-37	7
Total	50

In table 2.4 upper values are excluded from the class i.e., in the class 17-21 only values from 17 to 20 are taken and the values of 21 in considered in the next class. Such type of distribution is known as exclusive class.

Open-end classes: It may be the case that some values in the data set are extremely small compared to the other values of the data set and similarly some values are extremely large in comparison. Then what we do is we do not use the lower limit of the first class and the upper limit of the last class. Such classes are called open end classes.

Table 2.5: Open-end Class Grouped Frequency Table

Temperature	Frequency
Below 21	17
21-25	7
25-29	10
29-33	9
Above 33	7
Total	50

Size of the Class: The length of the class is called the class width. It is also known as class size.

Class interval or size of the class = Upper Limit – Lower Limit

Mid-point of the Class: The midpoint of a class interval is called Mid-point of the Class. It is the representative value of the entire class.

Mid-point of the class = (Upper Limit + Lower Limit) / 2

Continuous Frequency Distribution: If we deal with a continuous variable, it is not possible to arrange the data in the class intervals of above type. Let us consider the distribution of age in years. If class intervals are 15-19, 20-24 then the persons with ages between 19 and 20 years are not taken into consideration. In such a case we form the class intervals as shown below in table 2.6.

Table 2.6: Continuous Data

Age(in Years)
Below 5
5 or more but less than 10
10 or more but less than 15
15 or more but less than 20
20 or more but less than 25
…

As all cases have been covered in this table. But it is difficult to perform calculations using this table, therefore data is represented as in the table 2.7.

Table 2.7: Continuous Data using Classes

Age(in Years)
0-5
5-10
10-15
15-20
20-25
25-30

This form of frequency distribution is known as continuous frequency distribution. It should be clearly understood that in the above classes, the upper limits of each class are excluded from the respective classes. Such classes in which the upper limits are excluded from the respective classes and are included in the immediate next class are known as 'exclusive classes' and the classification is termed as 'exclusive type classification'.

2.2.3 Graphical Representation of a Frequency Distribution

It is often useful to represent a frequency distribution by means of a diagram which makes the data easily understandable and conveys the general information about the data. Diagrammatic representation also facilitates the comparison of two or more frequency distributions.

Graphs are charts consisting of points, lines and curves. Charts are drawn on graph sheets. Scales are to be chosen suitably in both X and Y axes so that entire data can be presented in the graph sheet. Statistical measures such as quartiles, median and mode can be found from the appropriate graph. Graphs are useful for analysis of time series, regression analysis, business forecasting, interpolation, extrapolation, etc.

Types of graphs

Graphs in statistics are broadly divided into two categories.

i) Graphs of time series or Historigrams
ii) Graphs of frequency distribution

Graphs of time series or Historigrams: A historigram is a graph to show a time series. It shows the fluctuation of a variable over a given period. X axis is used to denote the time and Y axis the value of the variable. Each pair of (time, variable) is denoted by a point on the graph. After plotting all such points, successive points are joined by straight lines. The resulting curve is historigram.

For example, let us draw a historigram (as in figure 2.1) to show the population in various census years with the given data as in table 2.8.

Table 2.8: Population in Various Census Years

Census Year	1951	1961	1972	1981	1998	2017
Population(in Million)	33.44	42.88	65.31	83.78	130.58	200.17

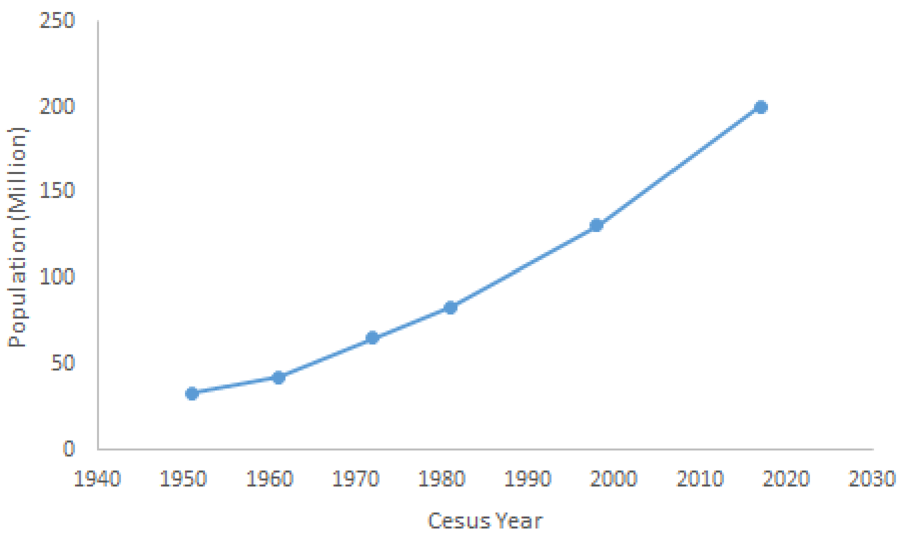


Figure 2.1: Historigram for data in table 2.8

Graphs of frequency distribution: There are various types of graphs of frequency distribution such as:

a) Histogram

b) Frequency polygon used to present continuous frequency distribution

c) Frequency curve

d) Ogive curve used to represent cumulative frequency distribution

e) Pie chart used to represent relative frequency.

f) Bar Diagram used to compare the frequencies.

HISTOGRAM: In drawing the histogram of a given continuous frequency distribution we first mark off along the x-axis all the class intervals on a suitable scale. On each class interval rectangles are drawn with heights proportional to the frequency of the corresponding class interval. The diagram of continuous rectangles so obtained is called histogram.

For examples, the table 2.9 gives the life times of 400 bulbs.

Table 2.9: Lifetime of Bulbs

Lifetime (in hours)	Number of bulbs
300 – 400	14
400 – 500	56
500 – 600	60
600 – 700	86
700 – 800	74
800 – 900	62
900 – 1000	48

The histogram for table 2.9 is:

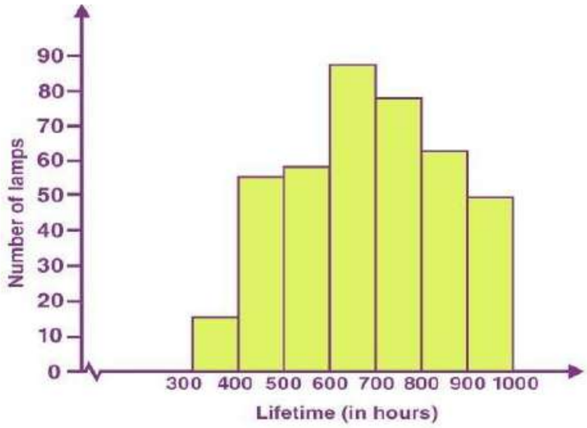


Figure 2.2: Histogram for data in table 2.9

Histogram in Python: It is important to represent the given data using histogram. But it is more important to represent it using a tool. In this course, we will be using Python as a tool for representing any data. From this point onwards, we will learn the various representations using Python too.

Creating Numpy Histogram: Numpy has a built-in numpy.histogram() function which represents the frequency of data distribution in the graphical form. The rectangles having equal horizontal size corresponds to class interval called bin and variable heightcorresponding to the frequency. It can be created using following statement:

numpy.histogram(data, bins=10, range=None, normed=None, weights=None,density=None).

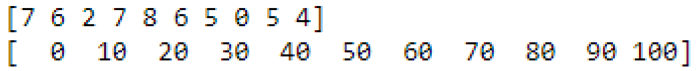
Where,

Attribute	*Parameter*
data	*array or sequence of array to be plotted*
bins	*int or sequence of str defines number of equal width bins in a range, default is 10*
range	*optional parameter sets lower and upper range of bins*
normed	*optional parameter same as density attribute, gives incorrect result for unequal bin width*
weights	*optional parameter defines array of weights having same dimensions as data*
density	*optional parameter if False result contain number of sample in each bin, if True result contain probability density function at bin*
data	*array or sequence of array to be plotted*
bins	*int or sequence of str defines number of equal width bins in a range, default is 10*
range	*optional parameter sets lower and upper range of bins*
normed	*optional parameter same as density attribute, gives incorrect result for unequal bin width*
weights	*optional parameter defines array of weights having same dimensions as data*
density	*optional parameter if False result contain number of sample in each bin, if True result contain probability density function at bin*

The creation of Numpy histogram can be better understood by the following programs:

```
# Program 2.1: Histogram Numeric Representation
# Import libraries
import numpy as np
# Creating dataset
a = np.random.randint(100, size =(50))
```

```
# Creating histogram
np.histogram(a, bins = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100])
hist, bins = np.histogram(a, bins = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100])
# printing histogram
print  (hist)
print  (bins)
Output:
```



```
[7 6 2 7 8 6 5 0 5 4]
[  0  10  20  30  40  50  60  70  80  90 100]
```

The above numeric representation of histogram can be converted into a graphical form. The plt() function present in pyplot submodule of Matplotlib takes the array of dataset and array of bin as parameter and creates a histogram of the corresponding data values.

```
# Program 2.2: Histogram
# import libraries
from matplotlib import pyplot as plt
import numpy as np
# Creating dataset
a = np.random.randint(100, size =(50))
# Creating plot
fig = plt.figure(figsize =(10, 7))
plt.hist(a, bins = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100])
plt.title("Numpy  Histogram")
# show plot
plt.show()
```

Output:

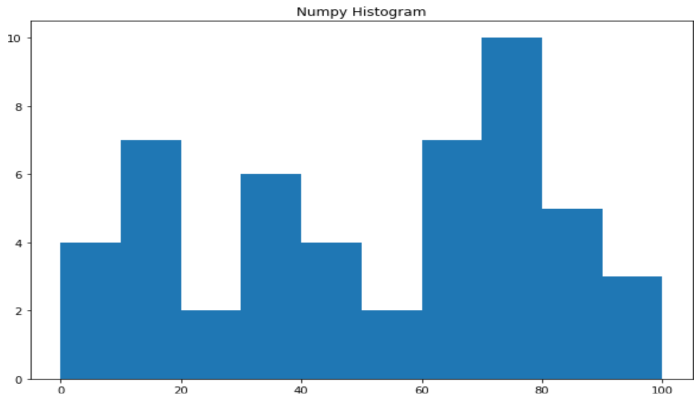


Figure 2.3: Histogram using Program 2.2

Creating Histogram using Matplotlib: To create a histogram the first step is to create bin of the ranges, then distribute the whole range of the values into a series of intervals, and the count the values which fall into each of the intervals. Bins are clearly identified as consecutive, non-overlapping intervals of variables. The matplotlib.pyplot.hist() function is used to compute and create histogram of x.

The following table shows the parameters accepted by matplotlib.pyplot.hist() function :

Attribute	*parameter*
X	*array or sequence of array*
bins	*optional parameter contains integer or sequence or strings*
density	*optional parameter contains boolean values*
range	*optional parameter represents upper and lower range of bins*
histtype	*optional parameter used to creae type of histogram [bar, barstacked, step, stepfilled], default is "bar"*
align	*optional parameter controls the plotting of histogram [left, right, mid]*
weights	*optional parameter contains array of weights having same dimensions as x*
bottom	*location of the baseline of each bin*
rwidth	*optional parameter which is relative width of the bars with respect to bin width*

Attribute	*parameter*
color	*optional parameter used to set color or sequence of color specs*
label	*optional parameter string or sequence of string to match with multiple datasets*
Log	*optional parameter used to set histogram axis on log scale*

Frequency Polygon and Curves: For an ungrouped distribution, the frequency polygon is obtained by plotting points with corresponding frequencies and joining the plotted points by means of straight lines. For a grouped frequency distribution, the points are mid-values of the class intervals. For equal class intervals the frequency polygon can be obtained by joining the middle Points of the upper sides of the adjacent rectangles of the histogram by means of straight lines. If the class intervals are of small width the polygon can be approximated by a smooth curve. The frequency curve can be obtained by drawing a smooth freehand curve through the vertices of the frequency polygon.

For example, let us present the following data given for **a batch of 400 students, the height of students is provided in the table 2.10, using frequency polygon.**

Table 2.10: Heights of Students Data

Height (In cm)	Number of Students
140-150	74
150-160	163
160-170	135
170-180	28
Total	400

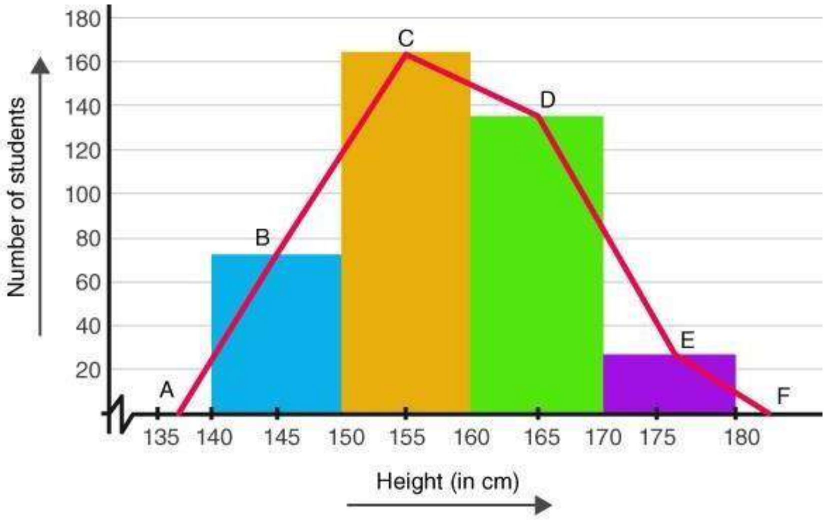


Figure 2.4: Frequency Polygon for the data in table 2.10

ABCDEF represents the given data graphically in form of frequency polygon as shown above in figure 2.4.

Now let us consider an example to draw a frequency curve.

Table 2.11: Data for Frequency Curve

Seed Yield (gms)	No. of Plants
2.5-3.5	4
3.5-4.5	6
4.5-5.5	10
5.5-6.5	26
6.5-7.5	24
7.5-8.5	15
8.5-9.5	10
9.5-10.5	5

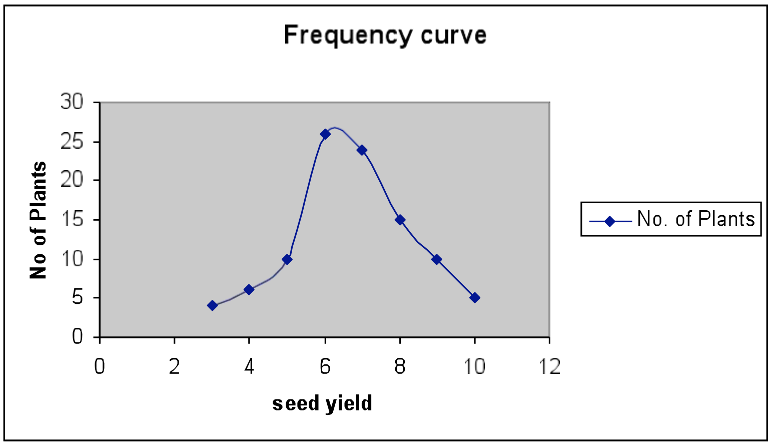


Figure 2.5: Frequency Curve for data in table 2.11

Drawing a frequency polygon in Python may be understood from the following example.

Suppose you have only the angle values for a set of data. Now you need to plot an angle distribution curve i.e., angle on the x axis v/s no. of times/frequency of angle occurring on the y axis. These are the angles sorted out for a set of data: -

[98.1706427, 99.09896751, 99.10879006, 100.47518838, 101.22770381, 101.70374296, 103.15715294, 104.4653976,105.50441485, 106.82885361, 107.4605319, 108.93228646, 111.22463712, 112.23658018, 113.31223886, 113.4000603, 114.14565594, 114.79809084, 115.15788861, 115.42991416, 115.66216071, 115.69821092, 116.56319054, 117.09232139, 119.30835385, 119.31377834, 125.88278338, 127.80937901, 132.16187185, 132.61262906, 136.6751744, 138.34164387,]

The data can easily be represented using Python with the help of following code:

```
# Program 2.3: Frequency Polygon
from matplotlib import pyplot as plt
import numpy as np
angles = [98.1706427, 99.09896751, 99.10879006, 100.47518838, 101.22770381,
101.70374296, 103.15715294, 104.4653976, 105.50441485, 106.82885361, 107.4605319,
108.93228646, 111.22463712, 112.23658018, 113.31223886, 113.4000603, 114.14565594,
114.79809084, 115.15788861, 115.42991416, 115.66216071, 115.69821092, 116.56319054,
117.09232139, 119.30835385, 119.31377834, 125.88278338, 127.80937901, 132.16187185,
132.61262906, 136.6751744, 138.34164387, ]
hist,edges = np.histogram(angles, bins=20)
bin_centers = 0.5*(edges[:-1] + edges[1:])
bin_widths = (edges[1:]-edges[:-1])
plt.bar(bin_centers,hist,width=bin_widths)
plt.plot(bin_centers, hist,'r')
plt.xlabel('angle [$^\circ$]')
plt.ylabel('frequency')
plt.show()
```

Output:

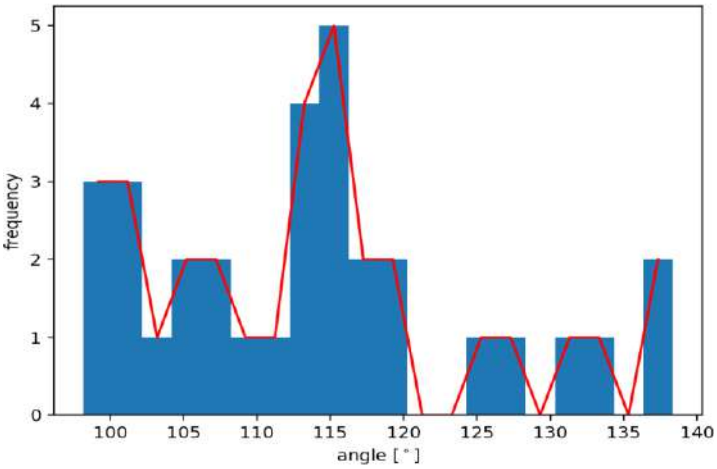


Figure 2.6: Frequency Polygon

2.2.4 Cumulative Frequency Distribution

Cumulative frequency is defined as a running total of frequencies. The frequency of an element in a set refers to how many of that element there are in the set. Cumulative frequencycan also be defined as the sum of all previous frequencies up to the current point.

Consider an example which shows the ages of participants in a certain class. We need to draw a cumulative frequency table for the data given in table 2.12.

Table 2.12: Frequency Table

Age	Frequency
10	3
11	18
12	13
13	12
14	7
15	27

The cumulative frequency table for the above data can be drawn as table 2.13. In this frequencies are the sum of the current frequency and previous frequencies. In other words, we can say that cumulative frequency shows the number of participants under or equal to the age of 10, 11, 12, 13, 14 and 15 respectively.

Table 2.13: Cumulative Frequency Table

Age	Frequency	Cumulative Frequency
10	3	3
11	18	18+3=21
12	13	21+13=34
13	12	34+12=46
14	7	46+7=53
15	27	53+27=80

However, there are two kinds of cumulative frequency distribution.

i) Less than cumulative frequency distribution

ii) More than cumulative frequency distribution

Less than cumulative frequency distribution: Frequency distribution both discrete and continuous are to be taken in ascending order. The total of the frequencies from the beginning up to and including each frequency is found. That cumulative frequency shows how many items are less than or equal to the corresponding value of the class interval.

More than cumulative frequency distribution: Frequency distribution both discrete and continuous are to be taken in ascending order. The total of the frequencies from the end up to and including each frequency is found. That cumulative frequency shows how many items are more than or equal to the corresponding value of the class interval.

Consider an example for both these types of cumulative frequencies for ungrouped data using the following table 2.14.

Table 2.14: Less than and More than Cumulative Frequency Curve

Weekly Wages (X)	Number of Workers (F)	Less than Cumulative Frequency	More Than Cumulative Frequency
120	8	8	30
150	12	20	22
240	8	28	10
300	2	30	2

Consider another example for cumulative frequencies using grouped data as shown in the following table 2.15.

Table 2.15: Less than and More than Cumulative Frequency Curve for Grouped Data

Marks (x)	No. of Students (f)	Marks below	No. of students	Marks above	No. of students
		Upper limit	Less than Cumulative Frequency (C.F.)	Lower limit	More than Cumulative Frequency (C.F.)
0-20	2	20	2	0	40
20-40	7	40	9	20	38
40-60	15	60	24	40	31
60-80	9	80	33	60	16
80-100	7	100	40	80	7
Total	40				

Ogive Curve for Cumulative Frequency

Let us now draw Ogive curve for both less than and greater(more) than using an example. Suppose we are given with weekly wages of various workers as shown in the table 2.16:

Table 2.16: Grouped Data

Weekly Wages (x)	No. of Workers (f)
0-20	41
20-40	51
40-60	64
60-80	38
80-100	7

First let us convert this table into less than c.f. and more than c.f.. **Table 2.17: Less than and More than Cumulative Frequency Curve for data in table 2.16**

Weekly Wages (x)	No. of Workers (f)	C.F.(Less than)	C.F.(More than)
0-20	41	41	201
20-40	51	92	160
40-60	64	156	109
60-80	38	194	45
80-100	7	201	7

Less than ogive: Upper limits of class intervals are marked on the x-axis and less than type cumulative frequencies are taken on y-axis. For drawing less than type curve, points (20, 41), (40, 92), (60, 156), (80, 194), (100, 201) are plotted on the graph paper and these are joined by free hand to obtain the less than ogive.

Greater than ogive: Lower limits of class interval are marked on x-axis and greater than type cumulative frequencies are taken on y-axis. For drawing greater than type curve, points (0, 201), (20, 160), (40, 109), (60, 45) and (80, 7) are plotted on the graph paper and these are joined by free hand to obtain the greater than type ogive.

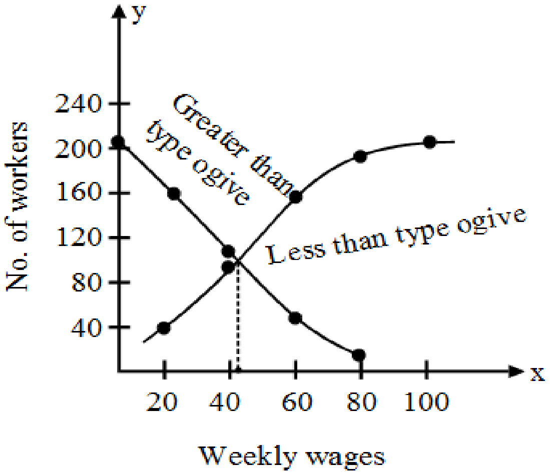


Figure 2.7: Less than and Greater than Ogive

Drawing Ogive in Python

We can draw both types of ogives in python. Let us understand by taking suitable examples. First we consider more than ogive. The more than ogive graph shows the number of values greater than the class intervals. The resultant graph shows the number of values in between the class interval, e.g., 0-10,10-20 and so on. Let us take a dataset, and we will now plot it"s more than ogive graph- [22,87,5,43,56,73, 55,54,11,20,51,5,79,31,27]. For this data the table 2.18 can be created as follows:

Table 2.18: Data for drawing more than Ogive

Class-Interval (x)	Frequency (f)	Cumulative Frequency (Less than)
0-10	2	2
10-20	1	3
20-30	3	6
30-40	1	7
40-50	1	8
50-60	4	12
60-70	0	12
70-80	2	14
80-90	1	15

Approach for drawing the ogive follows three steps:

(i) Import the modules (matplotlib and numpy)
(ii) Calculate the frequency and cumulative frequency of the data.
(iii) Plot it using the plot() function.

```
# Program 2.4: More than Ogive
# importing modules
import numpy as np
import matplotlib.pyplot as plt
# creating dataset
data = [22, 87, 5, 43, 56, 73, 55, 54, 11, 20, 51, 5, 79, 31, 27]
# creating class interval
classInterval = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90]
# calculating frequency and class interval
values, base = np.histogram(data, bins=classInterval)
# calculating cumulative sum
cumsum = np.cumsum(values)
# plotting the ogive graph
plt.plot(base[1:], cumsum, color='red', marker='o', linestyle='-')
plt.title('Ogive Graph')
plt.xlabel('Marks in End-Term')
plt.ylabel('Cumulative Frequency')
```

Output:

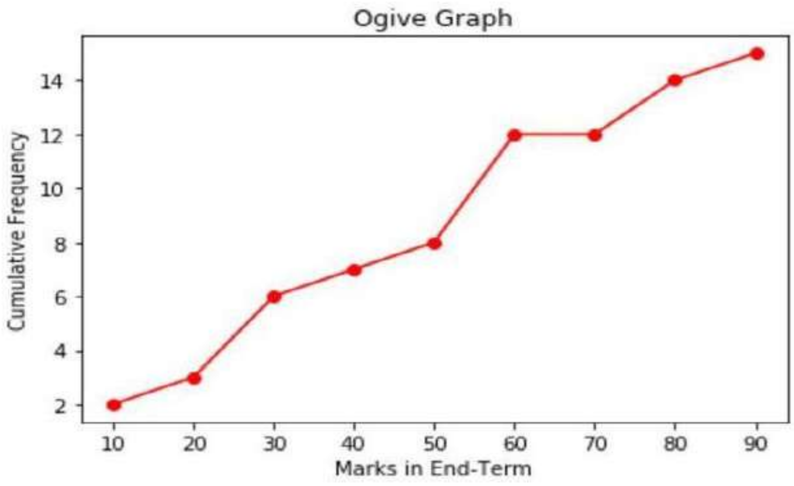


Figure 2.8: More than Ogive

Similarly, we can draw the less than ogive using following example.

In this example, we will plot less than Ogive graph which will show the less than values of class intervals. Dataset: [44,27,5,2,43,56,77,53,89,54,11,23, 51,5,79,25,39]. For this data the table can be created as follows:

Table 2.19: Data for drawing less than Ogive

Class-Interval (x)	Frequency (f)	Cumulative Frequency (More than)
0-10	3	17
10-20	1	14
20-30	3	13
30-40	1	10
40-50	2	9
50-60	4	7
60-70	0	3
70-80	2	3
80-90	1	1

Approach is same as above only the cumulative sum that we will calculate will be reversed using **flipud()** function present in the numpy library.

```
# Program 2.5: Less than Ogive
# importing modules
import numpy as np
import matplotlib.pyplot as plt
# creating dataset
data = [44, 27, 5, 2, 43, 56, 77, 53, 89, 54, 11, 23, 51, 5, 79, 25, 39]
# creating class interval
classInterval = [0, 10, 20, 30, 40, 50, 60, 70, 80, 90]
# calculating frequency and intervals
values, base = np.histogram(data, bins=classInterval)
# calculating cumulative frequency
cumsum = np.cumsum(values)
# reversing cumulative frequency
res = np.flipud(cumsum)
# plotting ogive
plt.plot(base[1:], res, color='brown', marker='o', linestyle='-')
plt.title('Ogive Graph')
plt.xlabel('Marks in End-Term')
plt.ylabel('Cumulative  Frequency')
```

Output:

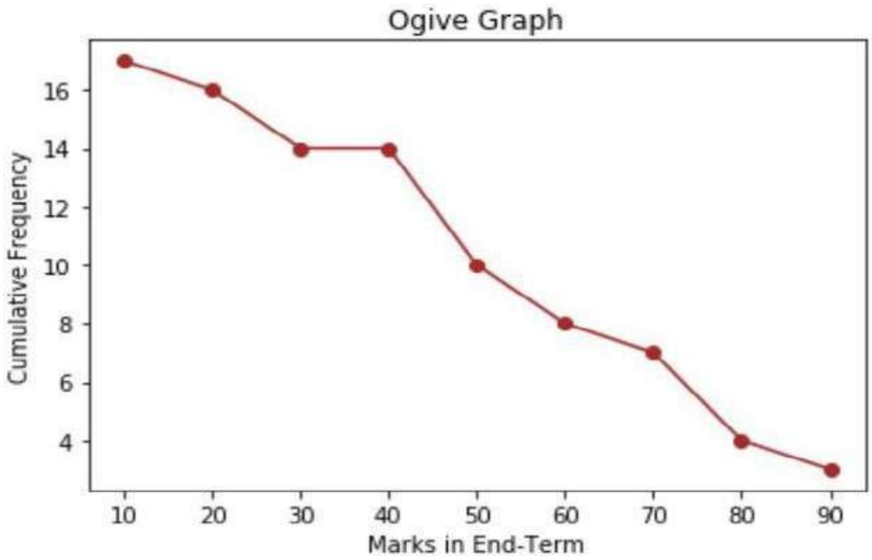


Figure 2.9: Less than Ogive

2.2.5 Pie Charts for Data Representation

A pie chart is a type of graph that represents the data in the circular graph. The slices of pie show the relative size of the data. It is a type of pictorial representation of data. A pie chart requires a list of categorical variables and the numerical variables. Here, the term "pie" represents the whole, and the "slices" represent the parts of the whole. Each slice denotes a proportionate part of the whole. The pie chart is an important type of data representation. It contains different segments and sectors in which each segment and sectors of a pie chart forms a certain portion of the total(percentage). The total of all the data is equal to 360°. **The total value of the pie is always 100%.**

The steps to design a pie chart are:

- Categorize the data
- Calculate the total
- Divide the categories
- Convert into percentages
- Finally, calculate the degrees
- Therefore, the pie chart formula is given as

Formula for pie chart = (Given Data/Total value of Data) × 360°.

Consider an example to draw a pie chart in a step by step manner:

Imagine a teacher surveys her class on the basis of their favourite Sports:

Football	**Hockey**	**Cricket**	**Basketball**	**Badminton**
10	**5**	**5**	**10**	**10**

This can be represented in the following table using above-said steps:

Table 2.20: Data based on favourite sports

Sports	No. of Students	Percentage	Degree(pie)
Football	10	(10/40)*100=25%	(10/40)*360=90^0
Hockey	5	(5/40)*100=12.5%	(5/40)*360=45^0
Cricket	5	(5/40)*100=12.5%	(5/40)*360=45^0
Basketball	10	(10/40)*100=25%	(10/40)*360=90^0
Badminton	10	(10/40)*100=25%	(10/40)*360=90^0
Total	40		

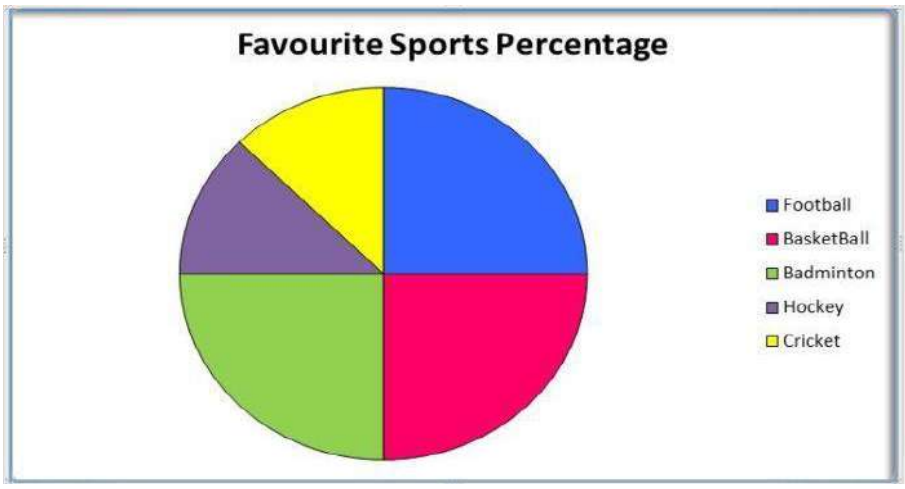


Figure 2.10: Pie Chart for table 2.20

Pie-chart using Python

Matplotlib API has pie() function in its pyplot module which create a pie chart representing the data in an array.

Syntax: matplotlib.pyplot.pie(data, explode=None, labels=None, colors=None, autopct=None, shadow=False)

Where,

***data** represents the array of data values to be plotted, the fractional area of each slice is represented by data/sum(data). If sum(data)<1, then the data values returns the fractional area directly, thus resulting pie will have empty wedge of size 1-sum(data).* ***labels** is a list of sequence of strings which sets the label of each wedge.**color** attribute is used to provide color to the wedges.**autopct** is a string used to label the wedge with their numerical value.* ***shadow** is used to create shadow of wedge.*

Let‟s create a simple pie chart using the pie() function:

```
# Program 2.6: Pie Chart

# Import libraries

from matplotlib import pyplot as plt
```

```
import numpy as np
# Creating dataset
cars = ['AUDI', 'BMW', 'FORD', 'TESLA', 'JAGUAR', 'MERCEDES']
data = [23, 17, 35, 29, 12, 41]
# Creating plot
fig = plt.figure(figsize =(10, 7))
plt.pie(data, labels = cars)
# show plot
plt.show()
```

Output:

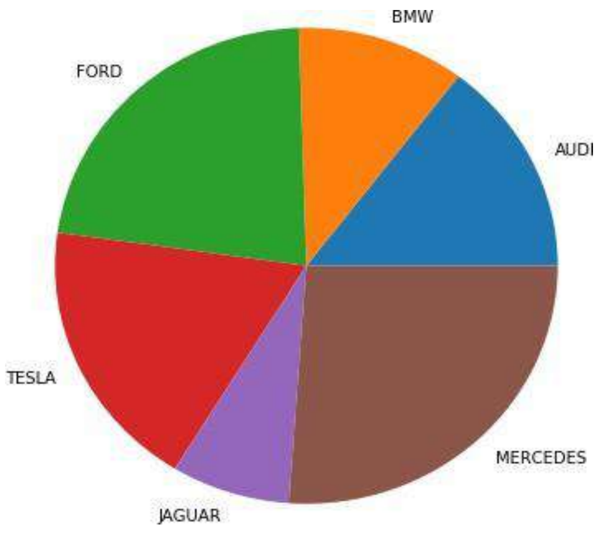


Figure 2.11: Pie Chart

2.2.6 Bar Diagram for Data Representation

Bar graphs are the pictorial representation of data in the form of vertical or horizontal rectangular bars, where the length of bars are proportional to the measure of data. They are also known as bar charts. Bar graphs are one of the means of data handling in statistics.

The bars drawn are of uniform width, and the variable quantity is represented on one of the axes. Also, the measure of the variable is depicted on the other axes. The heights or the lengths of the bars denote the value of the variable, and these graphs are also used to comparecertain quantities. The frequency distribution tables can be easily represented using bar chartswhich simplify the calculations and understanding of data.

The three major attributes of bar graphs are:

- The bar graph helps to compare the different sets of data among different groups easily.
- It shows the relationship using two axes, in which the categories on one axis and the discrete values on the other axis.
- The graph shows the major changes in data over time.

Types of Bar Charts

The bar graphs can be vertical or horizontal. The primary feature of any bar graph is its length or height. If the length of the bar graph is more, then the values are greater than any given data.

Bar graphs normally show categorical and numeric variables arranged in class intervals. They consist of an axis and a series of labelled horizontal or vertical bars. The bars represent frequencies of distinctive values of a variable or commonly the distinct values themselves. The number of values on the x-axis of a bar graph or the y-axis of a column graph is calledthe scale.

The types of bar charts are as follows:

- Vertical bar chart
- Horizontal bar chart

Even though the graph can be plotted using horizontally or vertically, the most usual type of bar graph used is the vertical bar graph. The orientation of the x-axis and y-axis are changed depending on the type of vertical and horizontal bar chart. Apart from the vertical and horizontal bar graph, the two different types of bar charts are:

- Grouped Bar Graph
- Stacked Bar Graph

Now, let us discuss the four different types of bar graphs.

Vertical Bar Graphs: When the grouped data are represented vertically in a graph or chart with the help of bars, where the bars denote the measure of data, such graphs are called vertical bar graphs. The data is represented along the y-axis of the graph, and the height ofthe bars shows the values.

Horizontal Bar Graphs: When the grouped data are represented horizontally in a chart with the help of bars, then such graphs are called horizontal bar graphs, where the bars show the measure of data. The data is depicted here along the x-axis of the graph, and the length of the bars denote the values.

Grouped Bar Graph: The grouped bar graph is also called the clustered bar graph, which is used to represent the discrete value for more than one object that shares the same category. In this type of bar chart, the total number of instances are combined into a single bar. In other words, a grouped bar graph is a type of bar graph in which different sets of data items are compared. Here, a single colour is used to represent the specific series across the set. The grouped bar graph can be represented using both vertical and horizontal bar charts.

Stacked Bar Graph: The stacked bar graph is also called the composite bar chart, which divides the aggregate into different parts. In this type of bar graph, each part can be represented using different colours, which helps to easily identify the different categories.The stacked bar chart requires specific labelling to show the different parts of the bar. In a stacked bar graph, each bar represents the whole and each segment represents the different parts of the whole.

Drawing a Bar Graph: In order to visually represent the data using the bar graph, we need to follow the steps given below.

- First, decide the title of the bar graph.
- Draw the horizontal axis and vertical axis.
- Now, label the horizontal axis.
- Write the names on the horizontal axis.
- Now, label the vertical axis.
- Finalise the scale range for the given data.
- Finally, draw the bar graph.

Bar Graph Examples: To understand the above types of bar graphs, consider the following examples:

In a firm of 400 employees, the percentage of monthly salary saved by each employee is given in the following table. Represent it through a bar graph.

Table 2.21: Data of Savings

Savings (in percentage)	Number of Employees (Frequency)
20	105
30	199
40	29
50	73
Total	400

The given data can be represented as a vertical bar graph:

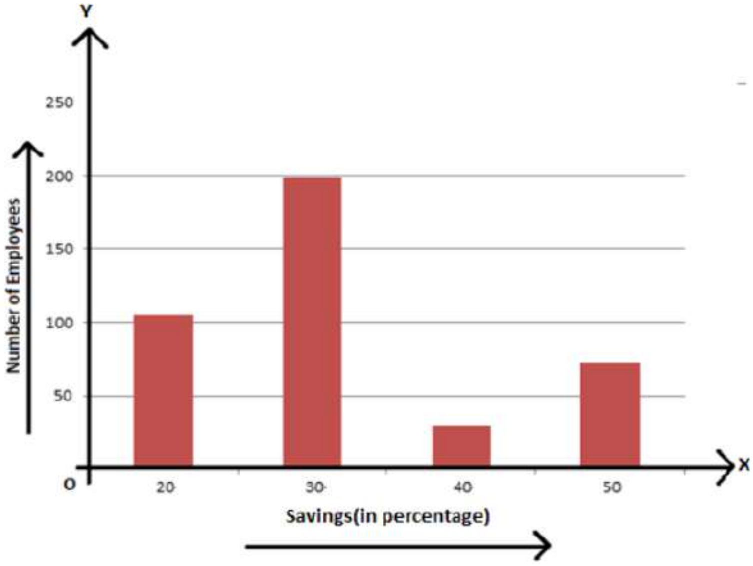


Figure 2.12: Vertical Bar Diagram for table 2.21

This can also be represented using a horizontal bar graph as follows:

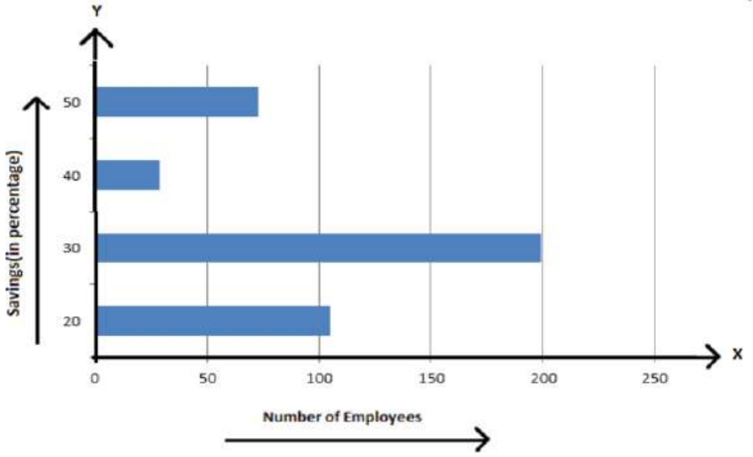


Figure 2.13: Horizontal Bar Diagram for table 2.21

Let as consider another example of grouped bar diagram: **A cosmetic companymanufactures 4 different shades of lipstick. The sale for 6 months is shown in the table. Represent it using bar charts.**

Table 2.22: Data of Lipsticks

Month	Sales (in units)			
	Shade 1	Shade 2	Shade 3	Shade 4
January	4500	1600	4400	3245
February	2870	5645	5675	6754
March	3985	8900	9768	7786
April	6855	8976	9008	8965
May	3200	5678	5643	7865
June	3456	4555	2233	6547

The graph given below depicts the following data:

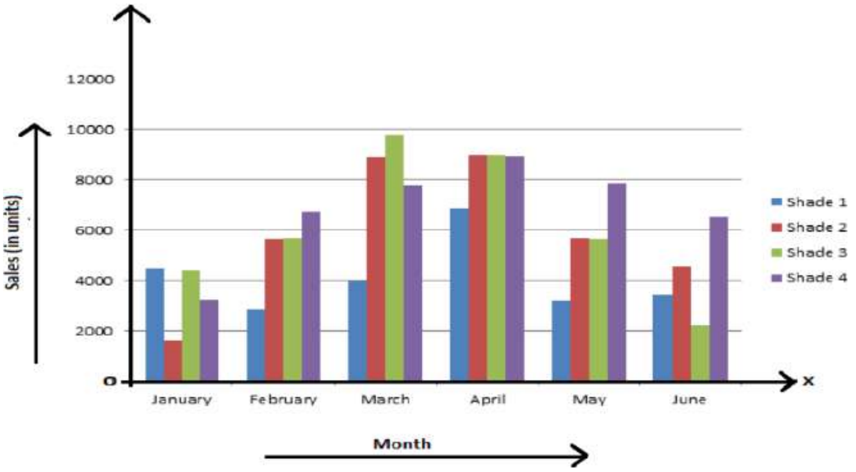


Figure 2.13: Grouped Bar Diagram for table 2.22

Bar diagram in Python

The **matplotlib** API in Python provides the bar() function. The syntax of the bar() function to be used with the axes is as follows:-

plt.bar(x, height, width, bottom, align)

The following program creates a bar plot bounded with a rectangle depending on the given parameters. Following is a simple example of the bar plot, which represents the number of students enrolled in different courses of an institute.

```
# Program 2.7: Bar Diagram
import numpy as np
import matplotlib.pyplot as plt
# creating the dataset
data = {'C':20, 'C++':15, 'Java':30,     'Python':35}
courses = list(data.keys())
values = list(data.values())
fig = plt.figure(figsize = (10, 5))
# creating the bar plot
plt.bar(courses, values, color ='maroon',       width = 0.4)
plt.xlabel("Courses offered")
plt.ylabel("No. of students enrolled")
plt.title("Students enrolled in different courses")
plt.show()
```

Output:

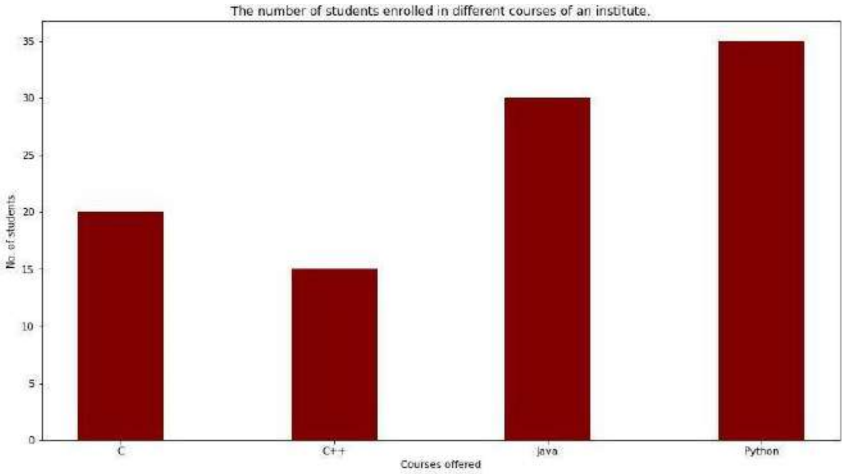


Figure 2.14: Bar Diagram

Horizontal charts can also be designed using Matplotlib. To create a horizontal bar chart:

```
#Program 2.8: Horizontal Bar Diagram

import matplotlib.pyplot as plt; plt.rcdefaults()
import numpy as np
import matplotlib.pyplot as plt
objects = ('Python', 'C++', 'Java', 'Perl', 'Scala', 'Lisp')
y_pos = np.arange(len(objects))
performance = [10,8,6,4,2,1]
plt.barh(y_pos, performance, align='center', alpha=0.5)
plt.yticks(y_pos, objects)
plt.xlabel('Usage')
plt.title('Programming language usage')
plt.show()
```

Output:

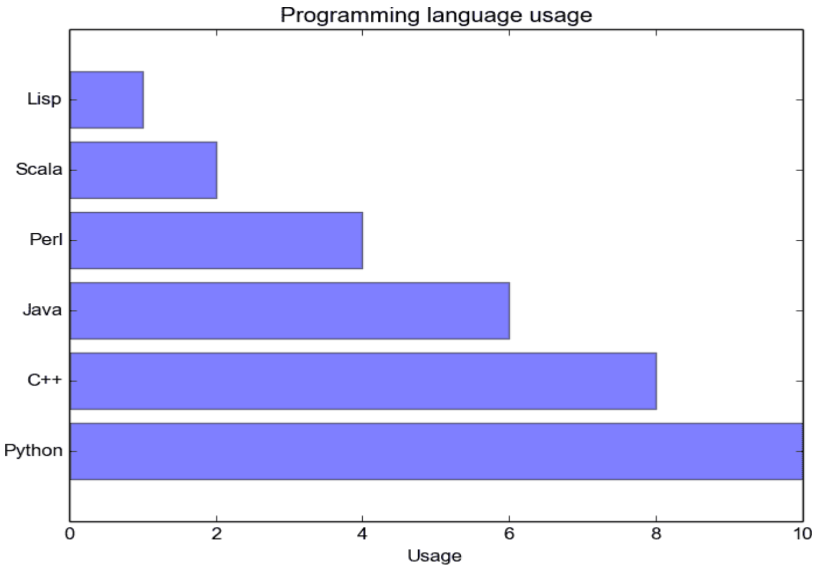


Figure 2.15: Horizontal Bar Diagram

Multiple bar plots: Multiple bar plots are used when comparison among the data set is to be done when one variable is changing. It can be drawn using python as shown in the following program.

```
#Program 2.9: Multiple Bar Plot
import numpy as np
import matplotlib.pyplot as plt
# set width of bar
barWidth = 0.25
fig = plt.subplots(figsize =(12, 8))
# set height of bar
IT = [12, 30, 1, 8, 22]
ECE = [28, 6, 16, 5, 10]
CSE = [29, 3, 24, 25, 17]
# Set position of bar on X axis
br1 = np.arange(len(IT))
br2 = [x + barWidth for x in br1]
br3 = [x + barWidth for x in br2]
# Make the plot
plt.bar(br1, IT, color ='r', width = barWidth, edgecolor ='grey', label ='IT')
plt.bar(br2, ECE, color ='g', width = barWidth, edgecolor ='grey', label ='ECE')
plt.bar(br3, CSE, color ='b', width = barWidth, edgecolor ='grey', label ='CSE')
# Adding Xticks
plt.xlabel('Branch', fontweight ='bold', fontsize = 15)
plt.ylabel('Students passed', fontweight ='bold', fontsize = 15)
plt.xticks([r + barWidth for r in range(len(IT))], ['2015', '2016', '2017', '2018', '2019'])
plt.legend()
plt.show()
```

Output:

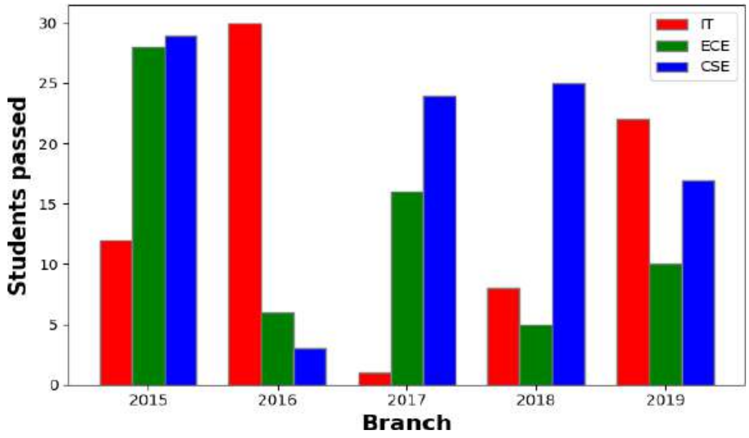


Figure 2.16: Multiple Bar Diagram

Stacked bar plot: Stacked bar plots represent different groups on top of one another. The height of the bar depends on the resulting height of the combination of the results of the groups. It goes from the bottom to the value instead of going from zero to value. The following bar plot represents the contribution of boys and girls in the team.

```
#Program 2.10: Stacked Bar Plot
import numpy as np
import matplotlib.pyplot as plt
N = 5
boys = (20, 35, 30, 35, 27)
girls = (25, 32, 34, 20, 25)
boyStd = (2, 3, 4, 1, 2)
girlStd = (3, 5, 2, 3, 3)
ind = np.arange(N)
width = 0.35
```

```
fig = plt.subplots(figsize =(10, 7))
p1 = plt.bar(ind, boys, width, yerr = boyStd)
p2 = plt.bar(ind, girls, width, bottom = boys, yerr = girlStd)
plt.ylabel('Contribution')
plt.title('Contribution by the teams')
plt.xticks(ind, ('T1', 'T2', 'T3', 'T4', 'T5'))
plt.yticks(np.arange(0, 81, 10))
plt.legend((p1[0], p2[0]), ('boys', 'girls'))
plt.show()
```

Output:

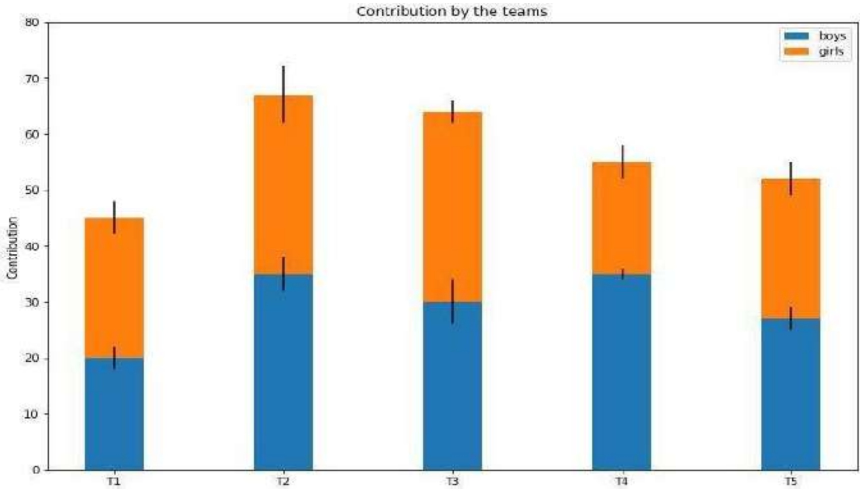


Figure 2.17: Stacked Bar Diagram

2.3 SUMMARY

In this module, representation of statistical data in the tabular and graphical forms has been discussed. The data representation in tabular form has been discussed terms of both grouped and ungrouped data. The presentation of continuous and discrete data has also been explained in this module. Further, the classification of data and its graphical representation has alsobeen discussed. The various types of graphs such as histogram, frequency polygon, bar diagram for presentation of data has been discussed at length. The concept of cumulative frequency distribution along with its tabular and graphical representation has also beenexplained in detail. The implementation of various types of graphs has been done in python. Overall, this module provides an insight of how to represent the statistical data.

2.4 PRACTICE QUESTION

Q.1 What are grouped and ungrouped frequency distributions? What are their uses? What are the considerations that one has to bear in mind while forming the frequency distribution?

Q.2 Explain the method of constructing Histogram and Frequency Polygon. Which out of these two is better representative of frequencies of (i) a particular group and (ii) whole group.

Q.3 What are the principles governing the choice of (i) Number of class intervals, (ii) The length of the class interval, (iii) The mid-point of the class interval.

Q.4 Write short notes on: (i) Frequency distribution, (ii) Histogram, frequency. polygon and frequency curve, (iii) Ogive.

Q.5 Write a program in python to draw various types of bar diagram considering your own data.

Q.6 Explain various types of bar diagram using suitable examples.

Q.7 What is cumulative frequency distribution? How can you represent c.f. graphically?

Q.8 The following numbers give the weights of 55 students of a class. Prepare a suitable frequency table.

42 74 40 60 82 115 41 61 75 83 63 53 110 76 84 50 67 65 78 77 56 95 68 69 104 80 79 79 54 73 59 81 100 66 49 77 90 84 76 42 64 69 70 80 72 50 79 52 103 96 51 86 78 94 71

(i) Draw the histogram and frequency polygon of the above data. (ii) For the above weights, make a cumulative frequency table and draw the less than ogive.

Q.9 A sample consists of 34 observations recorded correct to the nearest integer, ranging in value from 20l to 331. If it is decided to use seven classes of width 20 integers and to begin the first class at 199·5, find the class limits and class marks of the seven classes.

REFERENCES

- A. Abebe, J. Daniels, J.W.Mckean, "Statistics and Data Analysis".
- A. Martelli, A. Ravenscroft, S. Holden, "Python in a Nutshell", OREILLY.
- Clarke, G.M. & Cooke, D., "A Basic course in Statistics", Arnold.
- David M. Lane, "Introduction to Statistics".
- Eric Matthes, "Python Crash Course: A Hands-On, Project-Based Introduction to Programming".
- S.C.Gupta and V.K.Kapoor, "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, New Delhi.
- Weiss, N.A., "Introductory Statistics", Addison Wesley

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

UNIT III: MEASUREMENT OF CENTRAL TENDENCY

STRUCTURE

3.0 OBJECTIVES

In this module, we will try to understand about the various measures of central tendency. The various measures of central tendency i.e., mean, median and mode will be discussed in detail. The various ways to find these measures will be explored in this module. Various measures of dispersion will also be discussed in this module. The various measures of dispersion that would be discussed are range, standard deviation, variance, coefficient of variation and mean deviation. In the end the concept of skewness and kurtosis will also be deliberated.

3.1 INTRODUCTION

This module is designed to know about the different measures of central tendency. In statistics, the measure of central tendency plays very important role as most of the analysis and interpretation surrounds these measures. In most of the cases either mean or median is used for interpreting any data. Mode is used to know about the locality of references in the data. Further, the scatteredness of data also plays an important role in data analysis. The scatteredness can be easily identified by measures of dispersion. There are number of measures of dispersion viz., range, mean deviation, variance, standard deviation, etc. Even though the combination of measures of central tendency and dispersion are good enough for very good interpretation about data, yet there are two more measures named as skewness and kurtosis with the help of which it can be known about the shape of the curve and hence better interpretation of data. This module is a helping hand to the basics of interpreting and analysing the data.

3.2 MAIN CONTENT

3.2.1 Measures of Central Tendency

According to Professor Bowley, averages are "statistical constants which enable us to comprehend in a single effort the significance of the whole." They give us an idea about the concentration of the values in the central part of the distribution. Plainly speaking, an average of a statistical series is the value of the variable which is representative of the entire distribution. The following are the five measures of central tendency that are in common use:
(i) Arithmetic Mean or Simply Mean, (ii) Median, (iii) Mode, (iv) Geometric Mean, and (v) Harmonic Mean.

However, in this course, we will be focussing only on first three measures.

Requirements for an Ideal Measure of Central Tendency

According to Professor Yule, the following are the characteristics to be satisfied by-an ideal measure of central tendency;

(i) It should be rigidly defined.
(ii) It should be readily comprehensible and easy to calculate.
(iii) It should be based on all the observations.
(iv) It should be suitable for further mathematical treatment.
(v) It should be affected as little as possible by fluctuations of sampling.
(vi) It should not be affected much by extreme values.

Arithmetic Mean

Arithmetic mean of a set of observations is their sum divided by the number of observations. e.g the arithmetic mean x of n observations $\bar{x}$ of n observations x_1, x_2, …, x_n, is given by

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In case of frequency distribution, $x_i|f_i$, i = 1, 2, ..., n. where f_i is the frequency of the variable x_i;

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \dots + x_n f_n}{f_1 + f_2 + \dots + f_n} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1}{N}\sum_{i=1}^{n} f_i x_i, \text{ where } \sum_{i=1}^{n} f_i = N.$$

In case of grouped or continuous frequency distribution. X is taken as the mid. value of the corresponding class.

Let us understand it via some examples.

- First consider the following data: 1600, 1590, 1560, 1610, 1640, 1630. Find the arithmetic mean

$$\bar{x} = \frac{1600+1590+1560+1610+1640+1630}{6} = 1605.$$

- Now consider the following frequency distribution and find its arithmetic mean.

X	1	2	3	4	5	6	7
F	5	9	12	17	14	10	6

X	F	fx
1	5	5
2	9	18
3	12	36
4	17	68
5	14	70
6	10	60
7	6	42
Total	73	299

Arithmetic Mean = $\bar{x} = \frac{299}{73} = 4.09$

- Let us consider another example:

Marks (x)	0-10	10-20	20-30	30-40	40-50	50-60
No. of students (f)	12	18	27	20	17	6

Marks	No. of students (f)	Mid-Point (x)	fx
0-10	12	5	60
10-20	18	15	270
20-30	27	25	675
30-40	20	35	700
40-50	17	45	765
50-60	6	55	330
Total	100		2800

Arithmetic Mean = $\bar{x} = \frac{2800}{100} = 28$

It may be noted that if the values of x or/and f are large, the calculation of mean is quite time consuming and tedious. The arithmetic is reduced to a great extent, by taking the deviations of the given values from any arbitrary point 'A', as explained below.

Let $d_i = x_i - A$, then $f_i d_i = f_i(x_i - A) = f_i x_i - f_i A$.

Applying summation both sides, we get

$$\sum_{i=1}^{n} f_i d_i = \sum_{i=1}^{n} f_i x_i - A\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} f_i x_i - AN$$

$$\frac{1}{N}\sum_{i=1}^{n} f_i d_i = \frac{1}{N}\sum_{i=1}^{n} f_i x_i - A = \bar{x} - A$$

$$\bar{x} = A + \frac{1}{N}\sum_{i=1}^{n} f_i d_i$$

This formula is easy to handle with as compared to the earlier formula.

Any number can serve the purpose of arbitrary point 'A' but. usually, the value of x corresponding to the middle part of the distribution will be much more convenient.

In case of grouped or continuous frequency distribution, the arithmetic is reduced to a still greater extent by taking

$d_i = \frac{x_i - A}{h}$, where A is an arbitrary point and h is the common magnitude of class interval. In this case, we have $hd_i = x_i - A$, and proceeding exactly similarly as above, we get

$$\bar{x} = A + \frac{h}{N}\sum_{i=1}^{n} f_i d_i$$

Let us understand by an example:

- Consider the following distribution

Class Interval	0-8	8-16	16-24	24-32	32-40	40-48
Frequency	8	7	16	24	15	7

Let h = 8, A = 28

Class Interval	Mid-point (x)	Frequency (f)	D = (x – A) / h	fd
0-8	4	8	-3	-24
8-16	12	7	-2	-14
16-24	20	16	-1	-16

24-32	28	24	0	0
32-40	36	15	1	15
40-48	44	7	2	14
Total		77		-25

$\bar{x} = A + \frac{h}{N}\sum_{i=1}^{n} f_i d_i = 28 + \frac{8}{28}(-25) = 25.404$

Properties of Arithmetic Mean:

Property 1. Algebraic sum of the deviations of a set of values from their arithmetic mean is zero. If $x_i|f_i$, i = 1, 2, ... ,n is the frequency distribution, then

$\sum_{i=1}^{n} f_i (x_i - \bar{x}) = 0$, $\bar{x}$ being the mean of distribution.

Property 2. The sum of the squares of the deviations of a set of values is minimum when taken about mean.

Property 3. (Mean of the composite series). If $\bar{x}_i$, (i = 1, 2, ... , k) are the means of k-component series of sizes n_i, (i = 1, 2, ... , k) respectively, then the mean $\bar{x}$ of this composite series obtained on combining the component series given by the formula:

$$\bar{x} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2 + \cdots + n_k\bar{x}_k}{n_1 + n_2 + \cdots + n_k} = \frac{\sum_{i=1}^{k} n_i\bar{x}_i}{\sum_{i=1}^{k} n_i}$$

Merits and Demerits of Arithmetic Mean

Merits. (i) It is rigidly defined. (ii) It is easy to understand and easy to calculate. (iii) It is based upon all the observations. (iv) Of all the averages, arithmetic mean is affected least by fluctuations of sampling. This property is sometimes described by saying that arithmetic mean is, a stable average.

Demerits. (i) It cannot be determined by inspection nor it can be located graphically. (ii) Arithmetic mean cannot be used if we are dealing with qualitative characteristics which cannot be measured quantitate; such as, intelligence, honesty, beauty, etc. In such cases median is the only average to be used. (iii) Arithmetic mean cannot be obtained if a single observation is missing or lost or is illegible unless we drop it out and compute the arithmetic mean of the remaining values. (iv) Arithmetic mean is affected very much by extreme values.In case of extreme items, arithmetic mean gives a distorted picture of the distribution and no longer remains representative of the distribution. (v) Arithmetic mean may lead to wrong conclusions if the details of the data from which it is computed are not given. (vi) Arithmetic mean cannot be calculated if the extreme class is open. Moreover, even if a single observation is missing mean cannot be calculated. (vii) In extremely asymmetrical (skewed) distribution, usually arithmetic mean is not a suitable measure of location.

Median

Median of a distribution is the value of the variable which divides it into two equal parts. It is the va1ue which exceeds and is exceeded by the same number of observations, i.e., it is the value such that the number of observations above it is equal to the number of observations below it. The median is thus a positional average. In case of ungrouped data, if the number of observations is odd then median is the middle value after the values have been arranged in

ascending or descending order of magnitude. In case of even number of observations, there are two middle terms and median is obtained by taking the arithmetic mean of the middle terms. For example, the median of the values 25, 20, 15, 35, 18, i.e., 15, 18, 20, 25, 35 is 20
and the median of 8, 20, 50, 25, 15, 30, i.e., of 8, 15, 20, 25, 30, 50 is (20 + 25)/2 = 22·5.

In case of discrete frequency distribution median is obtained by considering the cumulative frequencies. The steps for calculating median are given below:

(i) Find N/2, where N = $\sum_{i=1}^{n} f_i$.
(ii) See the (less than) cumulative frequency (cf.) just greater than N/2.
(iii) The corresponding value of x is median.

Let us understand it via an example: Obtain the median for the following frequency distribution:

X	1	2	3	4	5	6	7	8	9
F	8	10	11	16	20	25	15	9	6

X	F	c.f.
1	8	8
2	10	18
3	11	29
4	16	45
5	20	65
6	25	90
7	15	105
8	9	114
9	6	120

Here N = 120, Therefore N/2 = 60 and cumulative frequency just greater than 60 is 65. The value corresponding to 65 is 5. Hence median is 5.

In the case of continuous frequency distribution, the class corresponding to the cf. just greater than N/2 is called the median class and the value of median is obtained by the following formula:

Median = l + $\frac{h}{f}(\frac{N}{2} - c)$

where l is the lower limit of the median class,

f is the frequency of the median class,

h is the magnitude of the median class,

'c' is the c.f. of the class preceding the median class,

And N = $\sum_{i=1}^{n} f_i$.

Let us consider another example using above formula.

- Find the median wage of the following distribution:

Wages	20-30	30-40	40-50	50-60	60-70
No. of Labourer	3	5	20	10	5

Wages	No. of Labourer (f)	C.F.
20-30	3	3
30-40	5	8
40-50	20	28
50-60	10	38
60-70	5	43

N/2 = 21.5 and the value just greater than this is 28 in the c.f. column. Therefore, according the above-said formula:

Median = l + $\frac{h}{f}(\frac{N}{2} - c)$ = 40 + $\frac{10}{20}(\frac{43}{2} - 8)$ = 40 + 6.75 = 46.75.

Merits and Demerits of Median

Merits. (i) It is rigidly defined. (ii) It is easily understood and is easy to calculate. In some cases, it can be located merely by inspection. (iii) It is not at all affected by extreme values. (iv) It can be calculated for distributions with open-end classes.

Demerits. (i) In case of even number of observations median cannot be determined exactly. We merely estimate it by taking the mean of two middle terms. (ii) It is not based on all the observations. (iii) It is not amenable to algebraic treatment. (iv) As compared with mean, it is affected much by fluctuations of sampling.

Mode

In statistics, the **mode** is the value which is repeatedly occurring in a given set. We can also say that the value or number in a data set, which has a high frequency or appears more frequently is called mode or modal value.

For example, mode of the set {3, 7, 8, 8, 9}, is 8. Therefore, for a finite number of observations, we can easily find the mode. A set of values may have one mode or more than one mode or no mode at all.

Bimodal, Trimodal & Multimodal (More than one mode)

- When there are two modes in a data set, then the set is called **bimodal**

For example, the mode of Set A = {2,2,2,3,4,4,5,5,5} is 2 and 5, because both 2 and 5 is repeated three times in the given set.

- When there are three modes in a data set, then the set is called **trimodal**

For example, the mode of set A = {2,2,2,3,4,4,5,5,5,7,8,8,8} is 2, 5 and 8

- When there are four or more modes in a data set, then the set is called **multimodal**

In the case of discrete frequency distribution mode is the value of x corresponding tomaximum frequency. For example, in the following frequency distribution:

X	1	2	3	4	5	6	7	8
f	4	9	16	25	22	15	7	3

The value of X corresponding to the maximum frequency, viz., 25 is 4. Hence mode is 4.

In the case of grouped frequency distribution, calculation of mode just by looking into the frequency is not possible. To determine the mode of data in such cases we calculate the modal class. Mode lies inside the modal class. The mode of data is given by the formula:

$$Mode = l + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where,

l = lower limit of the modal class

h = size of the class interval

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class

f_2 = frequency of the class succeeding the modal class

Let us take an example to understand this clearly.

- Find the mode for the following distribution:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

Here maximum frequency is 28. Thus the class 40-50 is the modal class. The value of mode is given by

Mode = 40 + $\frac{10(28-12)}{(2*28-12-20)}$ = 40 + 6.66 = 46.66

Sometimes mode is estimated from the mean and the median. For a symmetrical distribution, mean, median and mode coincide. If the distribution is moderately asymmetrical, the mean, median and mode obey the following empirical relationship (due to Karl Pearson):

Mean - Median = $\frac{1}{3}$ (Mean - Mode)

Mode = 3 Median - 2 Mean

Merits and Demerits or Mode

Merits. (i) Mode is readily comprehensible and easy to calculate. Like median, mode can be located in some cases merely by inspection. (ii) Mode is not at all affected by extreme values. (iii) Mode can be conveniently located even if the frequency distribution has class-intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. Open end classes also do not pose any problem in the location of mode.

Demerits. (i) Mode is ill-defined. It is not always possible to find a clearly defined mode. (ii) It is not based upon all the observations. (iii) It is not capable of further mathematical treatment. (iv) As compared with mean, mode is affected to a greater extent by fluctuations ofsampling.

Measures of Central Tendencies using Python:

We can find mean of given data using following function:

mean([data-set])

Parameters :[data-set] : List or tuple of a set of numbers.

Returns: Simple arithmetic mean of the provided data-set.

```
# Program 3.1: Python program to demonstrate mean() function from the statistics module
# Importing the statistics module
import statistics
# list of positive integer numbers
data1 = [1, 3, 4, 5, 7, 9, 2]
x = statistics.mean(data1)
# Printing the mean
print("Mean is :", x)


Output:
Mean is : 4.428571428571429
```

We can find median of given data using following function:

median([data-set])

Parameters: [data-set]: List or tuple or an iterable with a set of numeric values

Returns: Return the median (middle value) of the iterable containing the data

```
# Program 3.2: Python code to demonstrate the working of median() function.
# importing statistics module
import statistics
# unsorted list of random integers
data1 = [2, -2, 3, 6, 9, 4, 5, -1]
```

```
# Printing median of the
# random data-set
print("Median of data-set is : % s " % (statistics.median(data1)))


Output:
Median of data-set is : 3.5
```

We can find mode of given data using following function:

mode([data-set])

Parameters : [data-set] *which is a tuple, list or a iterator of real valued numbers as well as Strings.*

Return type : *Returns the most-common data point from discrete or nominal data.*

```
# Program 3.3: Python code to demonstrate the use of mode() function
import statistics


# declaring a simple data-set consisting of real valued positive integers.
set1 =[1, 2, 3, 3, 4, 4, 4, 5, 5, 6]
# In the given data-set we can infer that 4 has the highest population distribution
# So mode of set1 is 4
# Printing out mode of given data-set
print("Mode of given data set is % s" % (statistics.mode(set1)))


Output:
Mode of given data set is 4
```

3.2.2 Measures of Dispersion

Literal meaning of dispersion is scatteredness. We study dispersion to have an idea about the homogeneity or heterogeneity of the distribution. The measures of central tendency serve to locate the center of the distribution, but they do not reveal how the items are spread out on either side of the center. This characteristic of a frequency distribution is commonly referred to as dispersion. In a series all the items are not equal. There is difference or variation among the values. The degree of variation is evaluated by various measures of dispersion. Small dispersion indicates the high uniformity of the items, while large dispersion indicates less uniformity.

Characteristics of a good measure of Dispersion

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

- It should be rigidly defined
- It should be based on all the items.
- It should not be unduly affected by extreme items.
- It should lend itself for algebraic manipulation.
- It should be simple to understand and easy to calculate.

The following are the measures of dispersion:

(i) Range, (ii) Mean deviation, (iii) Standard deviation, (iv) Variance, and (v) Coefficient of Variation.

Range

The range is the difference between two extreme observations, or the distribution. If A and B are the greatest and smallest observations respectively in a distribution, then its range is A-B. Range is the simplest but a crude measure of dispersion. Since it is based on two extreme observations which themselves are subject to chance fluctuations, it is not at all a reliable measure of dispersion.

In individual observations and discrete series, A and B are easily identified. In continuous series, the following method is followed. A = Upper boundary of the highest class B = Lower boundary of the lowest class.

Find the value of range for the following data: 8,10, 5,9,12,11

A = 12, B = 5

Range = A – B = 7

Calculate the range from the following distribution:

Class Interval	0-10	10-20	20-30	30-40	40-50	50-60	60-70	70-80
Frequency	5	8	7	12	28	20	10	10

Here the range = 80 – 0 = 80

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

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

Mean Deviation

If $x_i|f_i$, i = 1, 2, …, n is the frequency distribution, then mean deviation from the average A, (either mean, median or mode), is given by

Mean deviation = $\frac{1}{N}\sum_i f_i|x_i - A|, \sum_i f_i = N$

where |Xi – A| represents the modulus or the absolute value of the deviation (Xi - A).

Since mean deviation is based on all the observations. it is a better measure of dispersion than range. But the step of ignoring the signs of the deviations |Xi – A| creates artificiality and, renders it useless for further mathematical treatment. It may be pointed out here that mean deviation is least when taken from median.

Let us consider few examples:

- Calculate Mean Deviation about Mean for the numbers given below: 1,2,3,4,5.

Here, Mean $=\bar{x}= \frac{\sum_i x_i}{n} = \frac{15}{5} = 3$

x	$\lvert x - \bar{x}\rvert$
1	2
2	1
3	0
4	1
5	2
$\sum_i x_i$=15	$\sum_i \lvert x - \bar{x}\rvert = 6$

M.D. about Mean $= \frac{\sum_i \lvert x - \bar{x}\rvert}{N} = \frac{6}{5} = 1.2$

Consider another examples based of discrete frequency distribution.

- Calculate the M.D. from Mean for the following data:

X	F	Fx	$\lvert x - \bar{x}\rvert = \lvert x - 6\rvert$	$f\lvert x - \bar{x}\rvert$
2	1	2	4	4
4	4	16	2	8
6	6	36	0	0
8	4	32	2	8
10	1	10	4	4
Total	16	96	14	24

Mean $=\bar{x}= \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{96}{16} = 6$

M.D. about Mean $= \frac{\sum_i f_i \lvert x - \bar{x}\rvert}{\sum_i f_i} = \frac{24}{16} = 1.5$

- Consider another examples based of continuous frequency distribution.

Marks	No. of Students(f)	Middle-Point(x)	fx	$\lvert x - \bar{x}\rvert = \lvert x - 33.4\rvert$	$f\lvert x - \bar{x}\rvert$
0-10	6	5	30	28.4	170.4
10-20	5	15	75	18.4	92
20-30	8	25	200	8.4	67.2
30-40	15	35	525	1.6	24
40-50	7	45	315	11.6	81.2
50-60	6	55	330	21.6	129.6
60-70	3	65	195	31.6	94.8
Total	50		1670		659.2

Mean = $\bar{x} = \frac{\sum_i f_i x_i}{\sum_i f_i} = \frac{1670}{50} = 33.4$

M. D. about Mean = $\frac{\sum_i f_i |x - \bar{x}|}{\sum_i f_i} = \frac{659.2}{50} = 13.18$

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

Demerits: 1. It is not a very accurate measure of dispersion. 2. It is not suitable for further mathematical calculation. 3. It is rarely used. It is not as popular as standard deviation. 4. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

Mean Deviation using Python:

Absolute mean deviation can be calculated in python using following code:

Using Numpy

```
# Program 3.4: Mean Deviation using Numpy
# Importing mean, absolute from numpy
data = [75, 69, 56, 46, 47, 79, 92, 97, 89, 88, 36, 96, 105, 32, 116, 101, 79, 93, 91, 112]
# Absolute mean deviation
mean(absolute(data - mean(data)))
Output:
20.055
```

Using Pandas

```
#Program 3.5: Mean Deviation using Pandas
# Import the pandas library as pd
import pandas as pd
data = [75, 69, 56, 46, 47, 79, 92, 97, 89, 88, 36, 96, 105, 32, 116, 101, 79, 93, 91, 112]
# Creating data frame of the given data
df = pd.DataFrame(data)
# Absolute mean deviation
df.mad() # mad() is mean absolute deviation function
Output:
20.055
```

Standard Deviation and Variance

Standard deviation, usually denoted by the sigma (σ), is the positive square root of the arithmetic mean of the squares of the deviations of the given values from their arithmetic mean. For the frequency distribution $x_i|f_i$, i = 1,2, ... , n.

$\sigma = \sqrt{\frac{1}{N}\sum_i f_i(x_i - \bar{x})^2}$, where $\bar{x}$ is the arithmetic mean of the distribution and $\sum_i f_i = N$.

The step of squaring the deviations $(x_i - \bar{x})$ overcomes the drawback of ignoring the signs in mean deviation. Standard deviation is also suitable for further mathematical treatment. Moreover, of all the measures, standard deviation is affected least by fluctuations of sampling.

The square of standard deviation is called the variance and is given by

$$\sigma^2 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^2$$

For individual series, standard deviation and variance can be calculated as follows:

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

Different formulae for calculating variance:

Now we know that variance for any frequency distribution may be written as:

$$\sigma^2 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^2$$

However, it can better be written as:

$$\sigma_x^2 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^2$$

If $\bar{x}$ is not a whole number, the calculation of σ_x^2 is very cumbersome and time consuming. Therefore, formula can be changed to

$$\sigma_x^2 = \frac{1}{N}\sum_i f_i x_i^2 - \left(\frac{1}{N}\sum_i f_i x_i\right)^2$$

If the values of x and f are large, the calculation of fx, fx^2 is quite tedious. In that case we take the deviations from any arbitrary point 'A '. Generally, the point in the middle of the distribution is much convenient and therefore we have

$$\sigma_x^2 = \frac{1}{N}\sum_i f_i d_i^2 - \left(\frac{1}{N}\sum_i f_i d_i\right)^2, \text{ where } d_i = x_i - A$$

We can make the calculations easier by using

$$\sigma_x^2 = h^2\left[\frac{1}{N}\sum_i f_i d_i^2 - \left(\frac{1}{N}\sum_i f_i d_i\right)^2\right], \text{ where } d_i = \frac{x_i - A}{h}$$

Now let us take some examples to understand the standard deviation and variance.

- Calculate S.D. for the data given below.

Sr. No.	Marks (X)	X^2
1	5	25
2	10	100
3	20	400
4	25	625
5	40	1600
6	42	1764
7	45	2025
8	48	2304
9	70	4900
10	80	6400
Total	385	20143

Mean = $\bar{x} = \frac{385}{10} = 38.5$, $\sigma = \sqrt{\frac{\Sigma x^2}{n} - (\frac{\Sigma x}{n})^2} = \sqrt{\frac{20143}{10} - 38.5^2} = 23.07$

- Calculate S.D. for the following data.

X	F	Fx	X^2	Fx^2
6	7	42	36	252
9	12	108	81	972
12	13	156	144	1872
15	10	150	225	2250
18	8	144	324	2592
Total	50	600		7938

$$\sigma = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - [\frac{\Sigma fx}{\Sigma f}]^2} = \sqrt{\frac{7938}{50} - 12^2} = 3.84$$

- Calculate the mean and standard deviation for tile following table giving tile age distribution of 542 members.

Age	20-30	30-40	40-50	50-60	60-70	70-80	80-90
No. of Members	3	61	132	153	140	51	2

In this data let us take d = $\frac{x-A}{h} = \frac{x-55}{10}$

Age	Mid-Value(x)	Members(f)	D=$\frac{x-55}{10}$	Fd	Fd^2
20-30	25	3	-3	-9	27
30-40	35	61	-2	-132	244
40-50	45	132	-1	-132	132
50-60	55	153	0	0	0
60-70	65	140	1	140	140
70-80	75	51	2	102	204
80-90	85	2	3	6	18
Total		542		-15	765

$$\bar{x} = A + h\frac{\sum fd}{N} = 55 + 10(\frac{-15}{542}) = 54.72$$

$$\sigma_x^2 = h^2[\frac{1}{N}\sum_i f_i d_i^2 - (\frac{1}{N}\sum_i f_i d_i)^2] = 100[\frac{765}{542} - (\frac{-15}{542})^2] = 133.3$$

σ = 11.55 years.

Standard Deviation and Variance in Python:

It can be calculated in python using many ways. However, we will be considering only two ways i.e. using Numpy and Statistics packages.

Using Numpy:

One can calculate the standard deviation by using **numpy.std()** function in python.

numpy.std(a, axis=None, dtype=None, out=None, ddof=0, keepdims=<no value>)

Parameters:

a: Array containing data to be averaged

axis: Axis or axes along which to average a

dtype: Type to use in computing the variance.

out: Alternate output array in which to place the result.

ddof: Delta Degrees of Freedom

keepdims: If this is set to True, the axes which are reduced are left in the result as dimensions with size one

```
# Program 3.6: Python program to get standard deviation of a list

# Importing the NumPy module

import numpy as np

# Taking a list of elements

list = [2, 4, 4, 4, 5, 5, 7, 9]

# Calculating standard deviation using std()

print(np.std(list))

Output:

2.0
```

One can calculate the variance by using **numpy.var()** function in python.

numpy.var(a, axis=None, dtype=None, out=None, ddof=0, keepdims=<no value>)

Parameters:

a: Array containing data to be averaged *axis: Axis or axes along which to average a* *dtype: Type to use in computing the variance.*

out: Alternate output array in which to place the result.

ddof: Delta Degrees of Freedom

keepdims: If this is set to True, the axes which are reduced are left in the result as dimensions with size one

```
# Program 3.7: Python program to get variance of a list

# Importing the NumPy module

import numpy as np

# Taking a list of elements

list = [2, 4, 4, 4, 5, 5, 7, 9]

# Calculating variance using var()

print(np.var(list))

Output:

4.0
```

Using Statistics:

variance([data], xbar)

Parameters :
[data] : An iterable with real valued numbers.
xbar (Optional) : Takes actual mean of data-set as value.

Returnype : Returns the actual variance of the values passed as parameter.

```
# Program 3.8: Python code to demonstrate the working of variance() function of Statistics

# Importing Statistics module

import statistics

# Creating a sample of data

sample = [2.74, 1.23, 2.63, 2.22, 3, 1.98]

# Prints variance of the sample set

# Function will automatically calculate it's mean and set it as xbar

print("Variance of sample set is % s" %(statistics.variance(sample)))


Output:

Variance of sample set is 0.40924
```

Another example of coding for finding the variance is:

```
# Program 3.9: Python code to demonstrate the use of xbar parameter for variance
# Importing statistics module
import statistics
# creating a sample list
sample = (1, 1.3, 1.2, 1.9, 2.5, 2.2)
# calculating the mean of sample set
m = statistics.mean(sample)
# calculating the variance of sample set
print("Variance of Sample set is % s" %(statistics.variance(sample, xbar = m)))


Output:
Variance of Sample set is 0.3656666666666667
```

Standard Deviation in Python using Statistics package can be coded as:

stdev([data-set], xbar)

Parameters :

[data] : An iterable with real valued numbers.

xbar (Optional): Takes actual mean of data-set as value.

Returnype : Returns the actual standard deviation of the values passed as parameter.

```
# Program 3.10: Python code to demonstrate stdev() function
# importing Statistics module
import statistics
# creating a simple data - set
sample = [1, 2, 3, 4, 5]
# Prints standard deviation
# xbar is set to default value of 1
print("Standard Deviation of sample is % s " % (statistics.stdev(sample)))


Output:
Standard Deviation of the sample is 1.5811388300841898
```

Another example of coding for finding the standard deviation is:

```
# Program 3.11: Python code to demonstrate use of xbar parameter while using stdev() function

# Importing statistics module

import statistics

# creating a sample list

sample = (1, 1.3, 1.2, 1.9, 2.5, 2.2)

# calculating the mean of sample set

m = statistics.mean(sample)

# xbar is nothing but stores the mean of the sample set

# calculating the standard deviation of sample set

print("Standard Deviation of Sample set is % s" %(statistics.stdev(sample, xbar = m)))


Output:

Standard Deviation of Sample set is 0.6047037842337906
```

Coefficient of Variation

Coefficient of variation can be found using the following formula:

Coefficient of Variation (C.V.) $= \frac{Standard\ Deviation}{Arithmetic\ Mean} * 100 = \frac{\sigma}{\bar{x}} * 100$.

Let us consider an example for this:

- The means and standard deviation values for the number of runs of two players A and B are 55; 65 and 4.2; 7.8 respectively. Who is the more consistent player?

Coefficient of variation of Player A $= \frac{\sigma}{x} * 100 = \frac{4.2}{55} * 100 = 7.64$

Coefficient of variation of Player B $= \frac{\sigma}{x} * 100 = \frac{7.8}{65} * 100 = 12$

Coefficient of variation of player A is less. Therefore, Player A is the more consistent player.

Coefficient of variation can be calculated using python as follows:

scipy.stats.variation(arr, axis = None) *function computes the coefficient of variation. It is defined as the ratio of standard deviation to mean.*

Parameters:

arr: *[array_like] input array.*

axis: *[int or tuples of int] axis along which we want to calculate the coefficient of variation.*

-> *axis = 0 coefficient of variation along the column.*

-> *axis = 1 coefficient of variation working along the row.*

Results: *Coefficient of variation of the array with values along specified axis.*

```
# Program 3.12: Coefficient of Variation
from scipy.stats import variation
import numpy as np
arr = np.random.randn(5, 5)
print ("array : \n", arr)
# rows: axis = 0, cols: axis = 1
print ("\nVariation at axis = 0: \n", variation(arr, axis = 0))
print ("\nVariation at axis = 1: \n", variation(arr, axis = 1))



Output:
array :
 [[-1.16536706 -1.29744691 -0.39964651  2.14909277 -1.00669835]
 [ 0.79979681  0.91566149 -0.823054    0.9189682  -0.01061181]
 [ 0.9532622   0.38630077 -0.79026789 -0.70154086  0.79087801]
 [ 0.53553389  1.46409899  1.89903817 -0.35360202 -0.14597738]
 [-1.53582875 -0.50077039 -0.23073327  0.32457064 -0.43269088]]

Variation at axis = 0:
 [-12.73042404   5.10272979 -14.6476392    2.15882202  -3.64031032]

Variation at axis = 1:
 [-3.73200773  1.90419038  5.77300406  1.29451485 -1.27228112]
```

3.2.3 Skewness

Literally skewness means 'lack of symmetry‟. We study skewness to have an idea about the shape of the curve which we can draw with the help of the given data. A distribution is said to be skewed if

(i) Mean, median and mode fall at different points. i.e., Mean G Median G Mode

(ii) The curve drawn with the help of the given data is not symmetrical but stretched more to one side than to the other.

Measures of Skewness: Various measures of skewness are

- $S_k = M - M_d$
- $S_k = M - M_o$,

where M is the mean, M_d the median and M_o is the mode of the distribution.

These are the absolute measures of skewness. As in dispersion, for comparing two series we do not calculate these absolute measures but we calculate the relative measures called the co-efficient of skewness which are pure numbers independent of units of measurement. The following is the coefficients of Skewness.

Prof. Karl Pearson's Coefficient of Skewness (S_k) = $\frac{(M-M0)}{\sigma}$, where σ is the standard deviation of the distribution.

If mode is ill-defined, then using the relation, $M_o = 3M_d - 2M$, for a moderately asymmetrical distribution, we get

$$S_k = \frac{3(M-M_d)}{\sigma}$$

The limits for Karl Pearson's coefficient of skewness are ±3. In practice, these limits are rarely attained. Skewness is positive if $M > M_o$ or $M > M_d$ and negative if $M < M_o$ or $M < M_d$.

Skewness can be represented graphically as shown in figure 3.1:

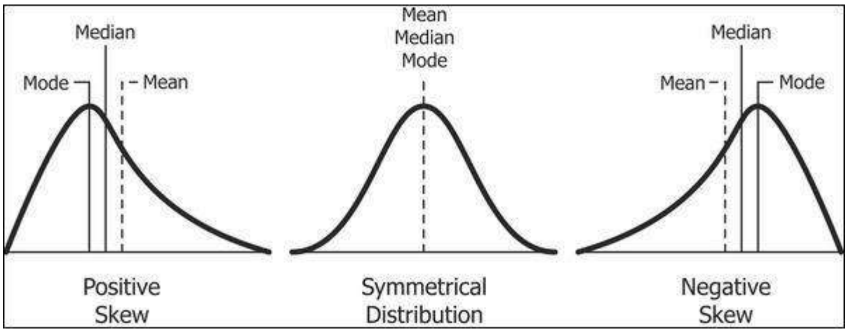


Figure 3.1: Skewness

Skewness can be calculated as per following examples:

- From the marks secured by 120 students in Section A and B of a class, the Following measures are obtained:

Section A: $\bar{X}$= 46.83; S.D = 14.8; Mode = 51.67

Section B: $\bar{X}$= 47.83; S.D = 14.8; Mode = 47.07

Determine which distribution of marks is more skewed.

Karl Pearson's Coefficient of Skewness for Section A = $S_k = \frac{(M-M0)}{\sigma} = \frac{(46.83-51.67)}{14.8} = -0.3270$

Karl Pearson's Coefficient of Skewness for Section B = $S_k = \frac{(M-M0)}{\sigma} = \frac{(47.83-47.07)}{14.8} = 0.05135$

Marks of Section A is more Skewed. But marks of Section A are negatively Skewed. Marks of Section B are Positively skewed.

- From a moderately skewed distribution of retail prices for men‟s shoes, it is found that the mean price is Rs. 20 and the median price is Rs. 17. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness of the distribution.

Coefficient of Variation (C.V.) $= \frac{Standard\ Deviation}{Arithmetic\ Mean} * 100 = \frac{\sigma}{\bar{x}} * 100$

$20 = \frac{\sigma}{20} * 100 \implies \sigma = 4$

$S_k = \frac{3(M-M_d)}{\sigma} = \frac{3(20-17)}{4} = 2.25.$

Skewness may be calculated using python as follows:

scipy.stats.skew(array, axis=0, bias=True) *function calculates the skewness of the data set.*

Parameters :
array : *Input array or object having the elements.*
axis : *Axis along which the skewness value is to be measured. By default axis = 0.* ***bias :*** *Bool; calculations are corrected for statistical bias, if set to False.*

Returns: *Skewness value of the data set, along the axis.*

```
#Program 3.13: finding Skewness
from scipy.stats import skew
import numpy as np
# random values based on a normal distribution
x = np.random.normal(0, 2, 10000)
print ("X : \n", x)
print('\nSkewness for data : ', skew(x))


Output:
X :
 [ 0.03255323 -6.18574775 -0.58430139 ... 3.22112446  1.16543279 0.84083317]
Skewness for data: 0.03248837584866293
```

3.2.4 Kurtosis

If we know the measures of central tendency, dispersion and skewness, we still cannot form a complete idea about the distribution as will be clear from the figure 3.2 in which all the three curves A, B and C are symmetrical about the mean and have the same range.

In addition to these measures we should know one more measure which Prof. Karl Pearson calls as the Convexity of curve or Kurtosis. Kurtosis enables us to have an idea about the flatness -or peakedness of the curve, It is measured by the co-efficient β_2 or its derivation γ_2 given by

$$\beta_2 = \frac{\mu_4}{\mu_2^2}, \gamma_2 = \beta_2 - 3$$

Where, $\mu_2, and\ \mu_4$ are moments about mean and can be calculated using following formulae:

The rth moment of a variable about the mean $\bar{X}$ usually denoted by μ_r is given by:

$\mu_r = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^r$, In particular $\mu_0 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^0 = \frac{\sum_i f_i}{N} = 1$.

$\mu_1 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^1 = 0$, being the algebraic sum of deviations from the mean.

Also, $\mu_2 = \frac{1}{N}\sum_i f_i(x_i - \bar{x})^2 = \sigma^2$

The rth moment of a variable x about any arbitrary point x = A, usually denoted by μ'_r is given by:

$$\mu'_r = \frac{1}{N}\sum_i f_i(x_i - A)^r = \frac{1}{N}\sum_i f_i d_i$$

Where $d_i = x_i - A$.

The relations between moments about an arbitrary point and about mean is represented as:

$\mu_2 = \mu'_2 - \mu'^2_1$,

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1$,

$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1$.

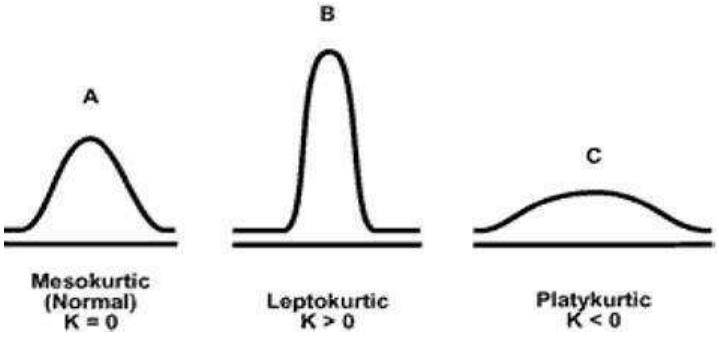


Figure 3.2: Kurtosis

Curve of the type 'A' which is neither flat nor peaked is called the normal curve or mesokurtic curve and for such a curve β_2 = 3, i.e., γ_2 = O. Curve of the type 'C' which is flatter than the normal curve is known as platykurtic curve and for such a curve β_2 < 3, i.e., γ_2 < O. Curve of the type 'B' which is more peaked than the normal curve is called leptokurtic and for such a curve, β_2 > 3 i.e.; γ_2 > 0.

Let us understand the concept by using an example.

- The data on daily wages of 45 workers of a factory are given. Compute skewness and kurtosis using moment about the mean. Comment on the results.

Wages	100-120	120-140	140-160	160-180	180-200	200-220	220-240
No. of Workers	1	2	6	20	11	3	2

Solution:

Wages	No. of Workers(f)	Mid-Point(x)	D=$\frac{x-170}{20}$	Fd	Fd^2	Fd^3	Fd^4
100-120	1	110	-3	-3	9	-27	81
120-140	2	130	-2	-4	8	-16	32
140-160	6	150	-1	-6	6	-6	6
160-180	20	170	0	0	0	0	0
180-200	11	190	1	11	11	11	11
200-220	3	210	2	6	12	24	48
220-240	2	230	3	6	18	54	162
Total	45			10	64	40	330

$\mu'_1 = \frac{\Sigma fd}{N} * \text{h} = \frac{10}{45} * 20 = 4.44$

$\mu'_2 = \frac{\Sigma fd^2}{N} * \text{h}^2 = \frac{64}{45} * 20^2 = 568.88$

$\mu'_3 = \frac{\Sigma fd^3}{N} * \text{h}^3 = \frac{40}{45} * 20^3 = 7111.11$

$\mu'_4 = \frac{\Sigma fd^4}{N} * \text{h}^4 = \frac{330}{45} * 20^4 = 1173333.33$

Moments about mean are:

$\mu_2 = \mu'_2 - \mu'^2_1 = 549.16$

$\mu_3 = \mu'_3 - 3\mu'_2\mu'_1 + 2\mu'^3_1 = -291.32$

$\mu_4 = \mu'_4 - 4\mu'_3\mu'_1 + 6\mu'_2\mu'^2_1 - 3\mu'^4_1 = 1113162.18$

Therefore,

Skewness = $\beta_1 = \mu_3^2$ = 0.00051

Kurtosis = $\beta_2 = \frac{\mu_4}{\mu_2^2}$ = 3.69

From the above calculations, it can be concluded that skewness is almost zero, thereby indicating that the distribution is almost symmetrical. Kurtosis, has a value greater than 3, thus implying that the distribution is leptokurtic.

Kurtosis in python can be calculated as follows:

***scipy.stats.kurtosis(array, axis=0, fisher=True, bias=True)** function calculates the kurtosis (Fisher or Pearson) of a data set.*

Parameters:

***array:** Input array or object having the elements.*

***axis:** Axis along which the kurtosis value is to be measured. By default, axis = 0.*

***fisher:** Bool; Fisher's definition is used (normal 0.0) if True; else Pearson's definition is used (normal 3.0) if set to False.*

***bias:** Bool; calculations are corrected for statistical bias, if set to False.*

***Returns:** Kurtosis value of the normal distribution for the data set.*

3.3 SUMMARY

In this module, various measures of central tendency viz., mean, median and mode have been discussed. Various methods for calculating mean, median and mode have been discussed for grouped, ungrouped, discrete and continuous data. The measure of dispersion such as range, mean deviation, standard deviation, variance and coefficient of variation have been discussed in detail. Different techniques for evaluating these measures have also been elaborated. Following this, skewness and kurtosis along with their measures have also been inspected. All these types of measures have been implemented in python using numpy, pandas, matplotlib and statistics packages.

3.4 PRACTICE QUESTION

Q.1 Compare mean, median and mode as measures of location of a distribution.

Q.2 Describe the different measures of central tendency of a frequency distribution, mentioning their merits and demerits.

Q.3 Given below is the distribution of 140 candidates obtaining marks X or higher in it certain examination (all marks are given in whole numbers):

X	10	20	30	40	50	60	70	80	90	100
c.f.	140	133	118	100	75	45	25	9	2	0

Calculate the mean, median and mode of the distribution.

Q.4 The mean of marks obtained in an examination by a group of 100 students was found to be 49·96. The mean of the marks obtained in the same examination by another group of 200 students was 52·32. Find the mean of the marks obtained by both the groups of students taken together.

Q.5 Which measure of location will be suitable to compare: (i) heights of students in two classes; (ii) size of agricultural holdings; (iii) average sales for various years; (iv) intelligence of students; (v) per capita income in several countries.

Q.6 Explain with suitable examples the term dispersion. State the relative and absolute measures of dispersion and describe the merits and demerits of these.

Q.7 Explain the main difference between mean deviation and standard deviation.

Q.8 Age distribution of hundred life insurance policy holders is as follows:

Age	17-19.5	20-25.5	26-35.5	36-40.5	41-50.5	51-55.5	56-60.5	61-70.5

Number	9	16	12	26	14	12	6	5

Calculate mean deviation from median age.

Q.9 What is standard deviation? Explain its superiority over other measures of dispersion.

Q.10 Calculate the mean and standard deviation of the following distribution:

X	2.5-7.5	7.5-12.5	12.5-17.5	17.5-22.5	22.5-27.5	27.5-32.5	32.5-37.5
f	65	121	175	198	176	120	66

Q. 11 The mean of 5 observations is 4·4 and variance is 8·24. It three of the five observation are 1, 2 and 6; find the other two.

Q.12 Lives of two models of refrigerators turned in for new models in a recent survey are:

Life	Model A	Model B
0-2	5	2
2-4	16	7
4-6	13	12
6-8	7	19
8-10	5	9
10-12	4	1

What is the average life of each model of these refrigerators? Which model shows more uniformity?

Q. 13 What do you understand by skewness? How is it measured? Distinguish clearly, by giving figures, between positive and negative skewness.

Q.14 Explain the methods of measuring skewness and kurtosis of a frequency distribution.

Q.15 Obtain Karl Pearson"s measure of skewness and kurtosis for the following data:

Values	5-10	10-15	15-20	20-25	25-30	30-35	35-40
Frequency	6	8	17	21	15	11	2

REFERENCES

- A. Abebe, J. Daniels, J.W.Mckean, "Statistics and Data Analysis".
- A. Martelli, A. Ravenscroft, S. Holden, "Python in a Nutshell", OREILLY.
- Clarke, G.M. & Cooke, D., "A Basic course in Statistics", Arnold.
- David M. Lane, "Introduction to Statistics".
- Eric Matthes, "Python Crash Course: A Hands-On, Project-Based Introduction to Programming".
- S.C.Gupta and V.K.Kapoor, "Fundamentals of Mathematical Statistics", Sultan Chand & Sons, New Delhi.

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

UNIT IV: DESCRIPTIVE STATISTICS

STRUCTURE

4.0 OBJECTIVES

In this module, we will try to understand about the descriptive statistics along with its importance and limitations followed by its implementation on python. Then a detailed discussion about exploratory data analysis will be undertaken. Various tools used in exploratory data analysis will also be explored. The implementation of these tools in python will also be deliberated upon. The concept of data visualization will also be taken up in the last part. The various types of data visualization methods will be considered along with their implementation in python.

4.1 INTRODUCTION

This module is designed to know about the overall presentation of data both in a textual manner as well as graphical manner. Descriptive statistics is used to describe the data as a whole, which mean the various measures of data are evaluated using descriptive statistics. It provides an overall description about the data and that‟s why known descriptive statistics. The exploratory data analysis presents the various tools used for interpreting and analysingthe data. Various types of graphical methods are used in this type of analysis. Further, data visualization is a concept, which is backbone of data analysis. If we are able to analyse data in a good manner, but not able to present it in an effective manner, it will be no or little use. That‟s where comes the need of data visualization. Data visualization provides an effective way of presenting our data.

4.2 MAIN CONTENT

4.2.1 Descriptive Statistics

Descriptive statistics summarize and organize characteristics of a data set. A data set is a collection of responses or observations from a sample or entire population. In quantitative research, after collecting data, the first step of statistical analysis is to describe characteristics of the responses, such as the average of one variable (e.g., age), or the relation between two variables (e.g., age and creativity).

Purpose of Descriptive Statistics

The essential features of the data of a study are defined using descriptive statistics. The sample and measurements are summarized. They form the basis of practically any quantitative analysis of the data in combination with simple graphic analysis. Also, a data set can be summarised and described using descriptive statistics through a variety of tabulated and graphical explanations and discussion of the observed results. Complex quantitative data are summed up in descriptive statistics. Descriptive statistics can be helpful to

- providing essential data on variables in a dataset
- highlighting possible relationships between variables.

Types of descriptive statistics

There are 3 main types of descriptive statistics:

- The distribution concerns the frequency of each value.
- The central tendency concerns the averages of the values.

- The variability or dispersion concerns how spread out the values are.

We have already discussed all three types of descriptive statistics in the previous units.

We can apply these to assess only one variable at a time, in univariate analysis, or to compare two or more, in bivariate and multivariate analysis.

Univariate descriptive statistics: Univariate descriptive statistics focus on only one variable at a time. It"s important to examine data from each variable separately using multiplemeasures of distribution, central tendency and spread.

Bivariate and Multivariate descriptive statistics: If we"ve collected data on more than one variable, we can use bivariate or multivariate descriptive statistics to explore whether there are relationships between them. In bivariate analysis, you simultaneously study the frequency and variability of two variables to see if they vary together. You can also compare the central tendency of the two variables before performing further statistical tests. Similarly, multivariate analysis can be applied for more than two variables.

Importance of Descriptive Statistics

Descriptive statistics allow for the ease of data visualization. It allows for data to be presented in a meaningful and understandable way, which, in turn, allows for a simplified interpretation of the data set in question. Raw data would be difficult to analyse, and trend and pattern determination may be challenging to perform. In addition, raw data makes it challenging to visualize what the data is showing.

Limitations of Descriptive Statistics

Descriptive statistics are so small that only the individuals or items you have calculated are summed up. The data you have obtained cannot be used to generalize to others or objects. When testing a drug to beat cancer in your patients, for example, you cannot say it would operate only based on descriptive statistics in other cancer patients.

Examples for Descriptive Statistics

- Take a simple number to sum up how well a batter does in baseball, the average batting. This figure is just the number of hits divided by the number of times at bat.
- You want to study the popularity of different leisure activities by gender. You distribute a survey and ask participants how many times they did each of the following in the past year: Go to a library, Watch a movie at a theatre, Visit a nationalpark. Your data set is the collection of responses to the survey. Now you can use descriptive statistics to find out the overall frequency of each activity (distribution), the averages for each activity (central tendency), and the spread of responses for each activity (variability).
- There are 100 students enrolled for a particular module. To find the overall performance of the students taking the respective module and the distribution of the marks, descriptive statistics must be used. Getting the marks as raw data would prove the determination of the overall performance and the distribution of the marks to be challenging.

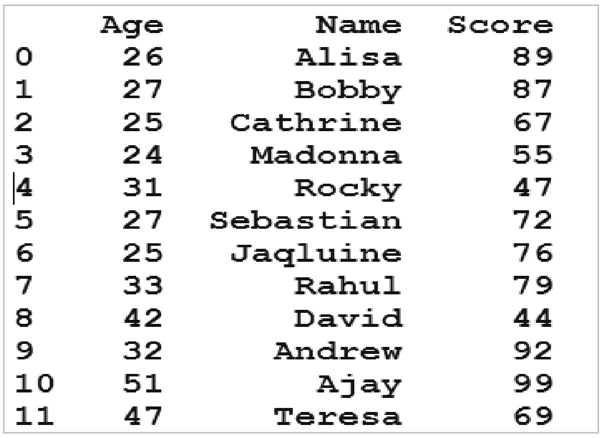
Descriptive Statistics in Python

It can be done in python using Pandas Describe() function. Describe Function gives the mean, std and IQR values. Generally describe() function excludes the character columns and gives summary statistics of numeric columns. We need to add a variable named include=‟all‟ to get the summary statistics or descriptive statistics of both numeric and character column.

Let‟s see with an example

```
# Program 4.1: Creation of Data Frame

Import pandas as pd

Import numpy as np

# Create a Dictionary of Series

d = {'Name':pd.Series(['Alisa', 'Bobby', 'Cathrine', 'Madonna', 'Rocky', 'Sebastian', 'Jaqluine', „Rahul', 'David', 'Andrew', 'Ajay', 'Teresa']), 'Age':pd.Series([26, 27, 25, 24, 31, 27, 25, 33, 42, 32, 51, 47]), 'Score':pd.Series([89,87,67,55,47,72,76,79,44,92,99,69])}

# Create a Data Frame

df = pd.DataFrame(d)

print df
```

Output:



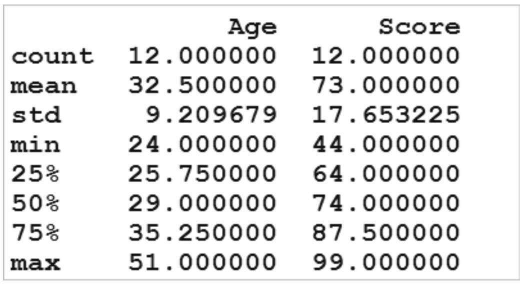
```
    Age       Name  Score
0    26      Alisa     89
1    27      Bobby     87
2    25   Cathrine     67
3    24    Madonna     55
4    31      Rocky     47
5    27  Sebastian     72
6    25   Jaqluine     76
7    33      Rahul     79
8    42      David     44
9    32     Andrew     92
10   51       Ajay     99
11   47     Teresa     69
```

Describe() Function gives the mean, std and IQR values. It excludes character column and calculate summary statistics only for numeric columns.

If we write the following statement:

Print df.describe()

The output will be:



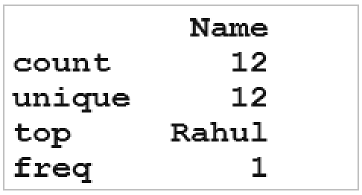
```
             Age      Score
count  12.000000  12.000000
mean   32.500000  73.000000
std     9.209679  17.653225
min    24.000000  44.000000
25%    25.750000  64.000000
50%    29.000000  74.000000
75%    35.250000  87.500000
max    51.000000  99.000000
```

describe() Function with an argument named include along with value object i.e include=‟object‟ gives the summary statistics of the character columns.

If we write the following statement:

Print df.decribe(include=[„object‟])

Then the output will be:



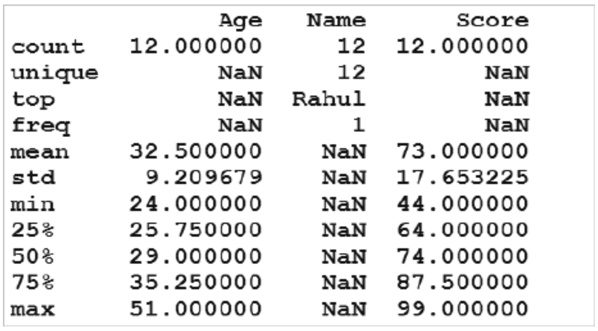
```
         Name
count      12
unique     12
top     Rahul
freq        1
```

describe() Function with include=‟all‟ gives the summary statistics of all the columns.

If we write the following statement:

Print df.decribe(include=‟all‟)

Then the output will be:



```
              Age   Name      Score
count   12.000000     12  12.000000
unique        NaN     12        NaN
top           NaN  Rahul        NaN
freq          NaN      1        NaN
mean    32.500000    NaN  73.000000
std      9.209679    NaN  17.653225
min     24.000000    NaN  44.000000
25%     25.750000    NaN  64.000000
50%     29.000000    NaN  74.000000
75%     35.250000    NaN  87.500000
max     51.000000    NaN  99.000000
```

However, if we don‟t require the output for all the functions then we can use a specific function for the output purpose.

Sr.No.	Function	Description
1	count()	Number of non-null observations
2	sum()	Sum of values
3	mean()	Mean of Values
4	median()	Median of Values

5	mode()	Mode of values
6	std()	Standard Deviation of the Values
7	min()	Minimum Value
8	max()	Maximum Value
9	abs()	Absolute Value
10	prod()	Product of Values
11	cumsum()	Cumulative Sum
12	cumprod()	Cumulative Product

For example, if we wish to find standard deviation, then we can write

```
# Program 4.2: To find standard deviation
import pandas as pd
import numpy as np
#Create a Dictionary of series
d = {'Name':pd.Series(['Tom', 'James', 'Ricky', 'Vin', 'Steve', 'Smith', 'Jack',   'Lee', 'David', 'Gasper',   'Betina',   'Andres']),   'Age':pd.Series([25,26,25,23,30,29,23,34,40,30,51,46]), 'Rating':pd.Series([4.23,3.24,3.98,2.56,3.20,4.6,3.8,3.78,2.98,4.80,4.10,3.65]) }
#Create  a  DataFrame
df = pd.DataFrame(d)
print df.std()


Output :
Age       9.232682
Rating    0.661628
dtype: float64
```

4.2.2 Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyse and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

Importance of Exploratory Data Analysis

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.

Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals. EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modelling, including machine learning.

Exploratory data analysis tools

Specific statistical functions and techniques you can perform with EDA tools include:

- Clustering and dimension reduction techniques, which help create graphical displays of high-dimensional data containing many variables.
- Univariate visualization of each field in the raw dataset, with summary statistics.
- Bivariate visualizations and summary statistics that allow you to assess the relationship between each variable in the dataset and the target variable you‟re looking at.
- Multivariate visualizations, for mapping and understanding interactions between different fields in the data.
- K-means Clustering is a clustering method in unsupervised learning where data points are assigned into K groups, i.e. the number of clusters, based on the distance from each group‟s centroid. The data points closest to a particular centroid will be clusteredunder the same category. K-means Clustering is commonly used in market segmentation, pattern recognition, and image compression.
- Predictive models, such as linear regression, use statistics and data to predict outcomes.

Types of Exploratory Data Analysis

There are four primary types of EDA:

- Univariate non-graphical. This is simplest form of data analysis, where the data being analysed consists of just one variable. Since it‟s a single variable, it doesn‟t deal with causes or relationships. The main purpose of univariate analysis is to describe the data and find patterns that exist within it.

- Univariate graphical. Non-graphical methods don"t provide a full picture of the data. Graphical methods are therefore required. Common types of univariate graphics include: (i) Stem-and-leaf plots, which show all data values and the shape of the distribution. (ii) Histograms, a bar plot in which each bar represents the frequency (count) or proportion (count/total count) of cases for a range of values. (iii) Box plots, which graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum.
- Multivariate non-graphical: Multivariate data arises from more than one variable. Multivariate non-graphical EDA techniques generally show the relationship between two or more variables of the data through cross-tabulation or statistics.
- Multivariate graphical: Multivariate data uses graphics to display relationships between two or more sets of data. The most used graphic is a grouped bar plot or bar chart with each group representing one level of one of the variables and each bar within a group representing the levels of the other variable. Other common types of multivariate graphics include: (i) Scatter plot, which is used to plot data points on a horizontal and a vertical axis to show how much one variable is affected by another.

  (ii) Multivariate chart, which is a graphical representation of the relationships between factors and a response. (iii) Run chart, which is a line graph of data plotted over time. (iv) Bubble chart, which is a data visualization that displays multiple circles (bubbles) in a two-dimensional plot. (v) Heat map, which is a graphical representation of data where values are depicted by color.

Exploratory Data Analysis Tools

Some of the most common data science tools used to create an EDA include:

Python: An interpreted, object-oriented programming language with dynamic semantics. Its high-level, built-in data structures, combined with dynamic typing and dynamic binding, make it very attractive for rapid application development, as well as for use as a scripting or glue language to connect existing components together. Python and EDA can be used together to identify missing values in a data set, which is important so you can decide how to handle missing values for machine learning.

R: An open-source programming language and free software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians in data science in developing statistical observations and data analysis.

Steps in Exploratory Data Analysis in Python

There are many steps for conducting Exploratory data analysis.

- Description of data
- Handling missing data
- Handling outliers
- Understanding relationships and new insights through plots

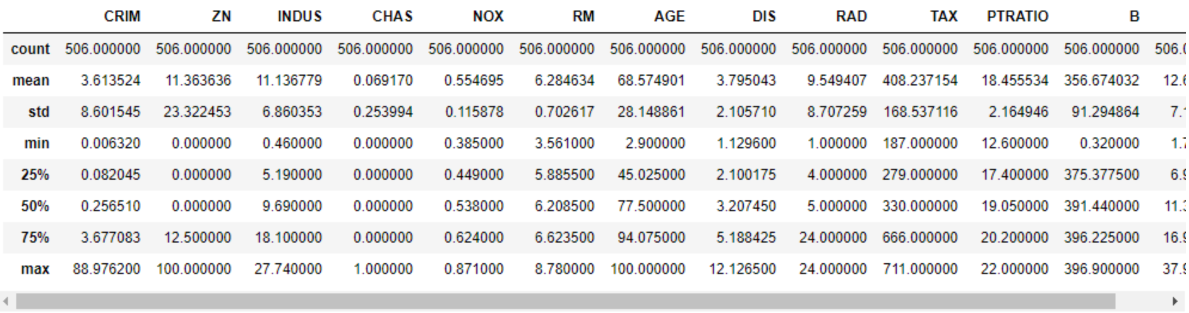
Description of data: We will be using the Boston Data Set for our examples, which can be imported from sklearn.datasets import load_boston.

We need to know the different kinds of data and other statistics of our data before we can move on to the other steps. A good one is to start with the describe() function in python.In Pandas, we can apply describe() on a DataFrame which helps in generating descriptive statistics that summarize the central tendency, dispersion, and shape of a dataset‟s distribution, excluding NaN values.

The result‟s index will include count, mean, std, min, max as well as lower, 50 and upper percentiles. By default, the lower percentile is 25 and the upper percentile is 75. The 50 percentile is the same as the median.

```
# Program 4.3: Loading the Dataset

import pandas as pd

from sklearn.datasets import load_boston

boston = load_boston()

x = boston.data

y = boston.target

columns = boston.feature_names

# creating dataframes

boston_df = pd.DataFrame(boston.data)

boston_df.columns = columns

boston_df.describe()
```

Output:



```
            CRIM          ZN       INDUS        CHAS         NOX          RM         AGE         DIS         RAD         TAX     PTRATIO           B
count 506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.000000  506.0
mean    3.613524   11.363636   11.136779    0.069170    0.554695    6.284634   68.574901    3.795043    9.549407  408.237154   18.455534  356.674032   12.6
std     8.601545   23.322453    6.860353    0.253994    0.115878    0.702617   28.148861    2.105710    8.707259  168.537116    2.164946   91.294864    7.1
min     0.006320    0.000000    0.460000    0.000000    0.385000    3.561000    2.900000    1.129600    1.000000  187.000000   12.600000    0.320000    1.7
25%     0.082045    0.000000    5.190000    0.000000    0.449000    5.885500   45.025000    2.100175    4.000000  279.000000   17.400000  375.377500    6.9
50%     0.256510    0.000000    9.690000    0.000000    0.538000    6.208500   77.500000    3.207450    5.000000  330.000000   19.050000  391.440000   11.3
75%     3.677083   12.500000   18.100000    0.000000    0.624000    6.623500   94.075000    5.188425   24.000000  666.000000   20.200000  396.225000   16.9
max    88.976200  100.000000   27.740000    1.000000    0.871000    8.780000  100.000000   12.126500   24.000000  711.000000   22.000000  396.900000   37.9
```

Handling missing data: Data in the real-world are rarely clean and homogeneous. Data can either be missing during data extraction or collection due to several reasons. Missing values need to be handled carefully because they reduce the quality of any of our performance matrix. It can also lead to wrong prediction or classification and can also cause a high bias forany given model being used. There are several options for handling missing values. However,the choice of what should be done is largely dependent on the nature of our data and the missing values. Some of the techniques for this are (i) Drop NULL or missing values, (ii) Fill Missing Values, and (iii) Predict Missing values with an ML Algorithm.

The dropping of NULL values is the fastest and easiest step to handle missing values. However, it is not generally advised. This method reduces the quality of our model as it reduces sample size because it works by deleting all other observations where any of the variables is missing. It can be achieved by using dropna() function as shown below:

```
Boston_df.shpae

Output:

(506,13)
```

```
Boston_df = boston_df.dropna()

Boston_df.shape

Output:

(506,13)
```

After implementing the above code, it will indicate that there are no null values in our data set.

Another way of handling missing values is by filling missing values. This is a process whereby missing values are replaced with a test statistic like mean, median or mode of the particular feature the missing value belongs to. Let‟s suppose we have a missing value of agein the boston data set. Then the below code will fill the missing value with the 30.

```
Boston_df[„AGE‟] = boston_df[„AGE‟].fillna(30)

Boston_df.shpae

Output:

(506,13)
```

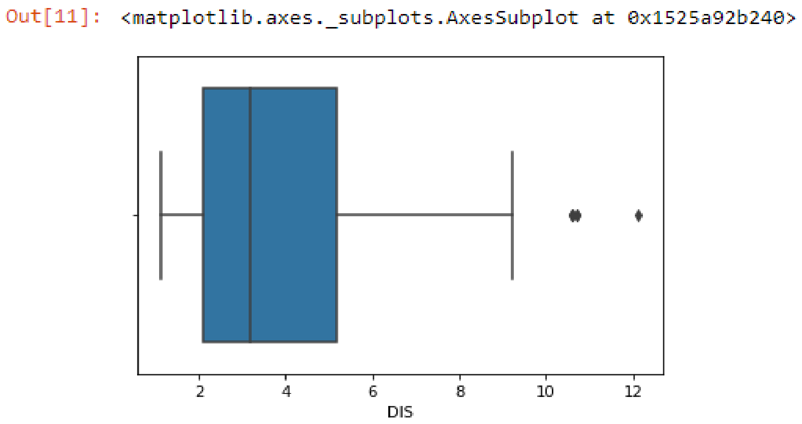
The third way is via predicting missing values with an ML Algorithm. This is by far one of the best and most efficient methods for handling missing data. Depending on the class of datathat is missing, one can either use a regression or classification model to predict missing data.

Handling outliers: An outlier is something which is separate or different from the crowd. Outliers can be a result of a mistake during data collection or it can be just an indication of variance in your data. Some of the methods for detecting and handling outliers are (i) BoxPlot (ii) Scatterplot (iii) Z-score and (iv) IQR(Inter-Quartile Range).

BoxPlot: A box plot is a method for graphically depicting groups of numerical data through their quartiles. The box extends from the Q1 to Q3 quartile values of the data, with a line at the median (Q2). The whiskers extend from the edges of the box to show the range of the data. Outlier points are those past the end of the whiskers. Boxplots show robust measures of location and spread as well as providing information about symmetry and outliers.

```
# Program 4.4: BoxPlot
import seaborn as sns
sns.boxplot(x=boston_df['DIS'])
```

Output:

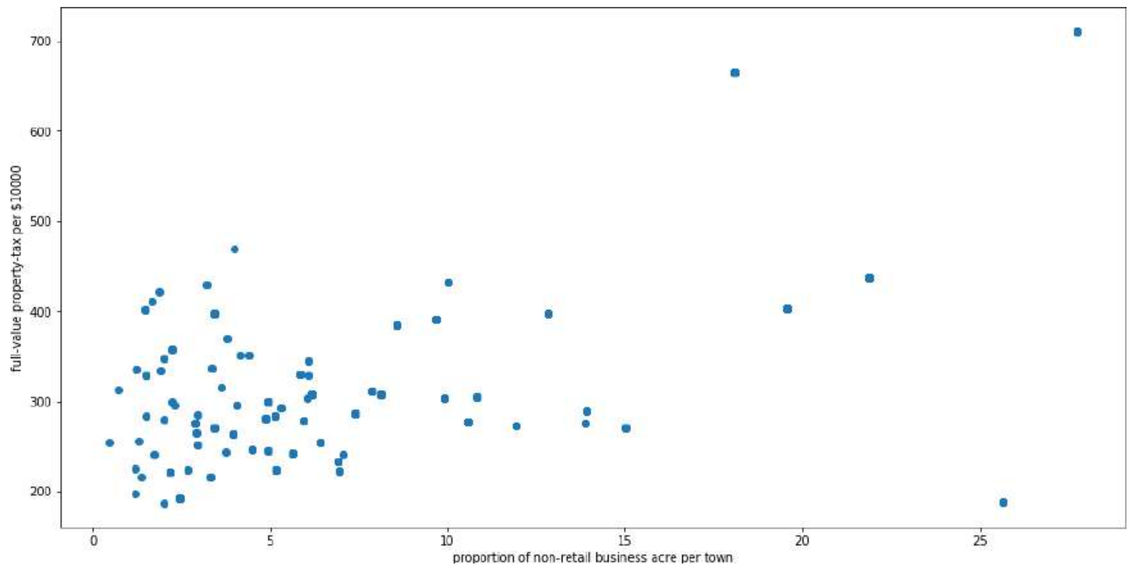


```
Out[11]: <matplotlib.axes._subplots.AxesSubplot at 0x1525a92b240>
```

Scatterplot: A scatter plot is a mathematical diagram using Cartesian coordinates to display values for two variables for a set of data. The data are displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis. The points that arefar from the population can be termed as an outlier.

```
# Program 4.5: Scatter Plot
import matplotlib.pyplot as plt
fig, ax = plt.subplots(figsize=(16,8))
ax.scatter(boston_df['INDUS'] , boston_df['TAX'])
ax.set_xlabel('proportion of non-retail business acre per town')
ax.set_ylabel('full-value property-tax per $10000')
plt.show()
```

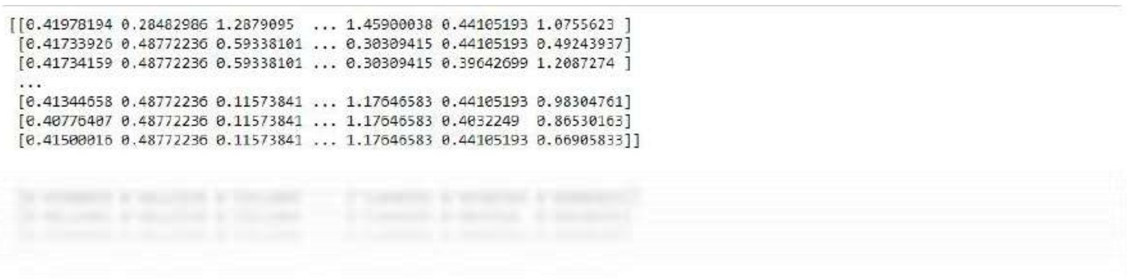
Output:



Z-score: The Z-score is the signed number of standard deviations by which the value of an observation or data point is above the mean value of what is being observed or measured. While calculating the Z-score we re-scale and center the data and look for data points that are too far from zero. These data points which are way too far from zero will be treated as the outliers. In most of the cases a threshold of 3 or -3 is used i.e. if the Z-score value is greater than or less than 3 or -3 respectively, that data point will be identified as outliers.

```
# Program 4.6: Z-Score

From scipy import stats

Import numpy as np

Z = np.abs(stats.zscore(boston_df))

Print(Z)
```

Output:



```
[[0.41978194 0.28482986 1.2879095  ... 1.45900038 0.44105193 1.0755623 ]
 [0.41733926 0.48772236 0.59338101 ... 0.30309415 0.44105193 0.49243937]
 [0.41734159 0.48772236 0.59338101 ... 0.30309415 0.39642699 1.2087274 ]
 ...
 [0.41344658 0.48772236 0.11573841 ... 1.17646583 0.44105193 0.98304761]
 [0.40776407 0.48772236 0.11573841 ... 1.17646583 0.4032249  0.86530163]
 [0.41500016 0.48772236 0.11573841 ... 1.17646583 0.44105193 0.66905833]]
```

For removing outliers, following statements may be used:

```
Boston_df_outlier_Zscore = boston_df[(z<3).all(axis=1)]

Boston_df_outlier_Zscore.shape
```

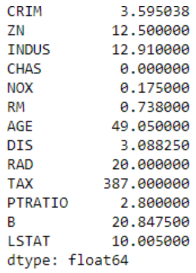
```
Output:

(415,13)
```

We can see from the above code that the shape changes, which indicates that our dataset has some outliers.

IQR: The interquartile range (IQR) is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

IQR = Q3 − Q1 and can be found using following statements:

```
Q1 = boston_df.quantile(0.25)

Q3 = boston_df_quantile(0.75)

IQR = Q3 – Q1

Print(IQR)


Output:
```



```
CRIM         3.595038
ZN          12.500000
INDUS       12.910000
CHAS         0.000000
NOX          0.175000
RM           0.738000
AGE         49.050000
DIS          3.088250
RAD         20.000000
TAX        387.000000
PTRATIO      2.800000
B           20.847500
LSTAT       10.005000
dtype: float64
```

Once we have IQR scores below code will remove all the outliers in our dataset.

```
Boston_df_outlier-IQR= boston_df[~((boston_df < (Q1 – 1.5 * IQR)) | (boston_df > (Q3 +
1.5 * IQR))).any(axis=1)]

Boston_df_outlier_IQR.shape

Output:

(274, 13)
```

4.2.3 Data Visualization

Data visualization is a graphical representation of quantitative information and data by using visual elements like graphs, charts, and maps. Data visualization convert large and small data sets into visuals, which is easy to understand and process for humans. Data visualization tools provide accessible ways to understand outliers, patterns, and trends in the data.

In the world of Big Data, the data visualization tools and technologies are required to analyse vast amounts of information. Data visualizations are common in our everyday life, but they always appear in the form of graphs and charts.

Data visualizations are used to discover unknown facts and trends. We can see visualizations in the form of line charts to display change over time. Bar and column charts are useful for observing relationships and making comparisons. A pie chart is a great way to show parts-of-a-whole and maps are the best way to share geographical data visually.

Today's data visualization tools go beyond the charts and graphs used in the Microsoft Excel spreadsheet, which displays the data in more sophisticated ways such as dials and gauges, geographic maps, heat maps, pie chart, and fever chart.

Effective Data Visualization

American Statistician and Yale Professor Edward Tufte saya useful data visualizations consist of complex ideas communicated with clarity, precision, and efficiency.

To craft an effective data visualization, we need to start with clean data that is well-sourced and complete. Once the data is ready to visualize, we need to pick the right chart to visualize. After that, we need to design and customize our visualization according to requirements. Simplicity is essential as we don't want to add any elements that distract from the data.

Importance of Data Visualization

- Data visualization is important because of the processing of information in human brains. Using graphs and charts to visualize a large amount of the complex data sets is more comfortable in comparison to studying the spreadsheet and reports.
- Data visualization is an easy and quick way to convey concepts universally. We can experiment with a different outline by making a slight adjustment.
- Data visualization can identify areas that need improvement or modifications.
- Data visualization can clarify which factor influence customer behavior.
- Data visualization helps you to understand which products to place where.
- Data visualization can predict sales volumes.

Uses of Data Visualization

- To make easier in understand and remember.
- To discover unknown facts, outliers, and trends.
- To visualize relationships and patterns quickly.
- To ask a better question and make better decisions.
- To competitive analyse.
- To improve insights.

Types of Data Visualizations

The earliest form of data visualization can be traced back the Egyptians in the pre-17th century, largely used to assist in navigation. As time progressed, people leveraged data visualizations for broader applications, such as in economic, social, health disciplines. Perhaps most notably, Edward Tufte published “The Visual Display of Quantitative Information” , which illustrated that individuals could utilize data visualization to present data in a more effective manner. His book continues to stand the test of time, especially as companies turn to dashboards to report their performance metrics in real-time. Dashboards are effective data visualization tools for tracking and visualizing data from multiple data sources, providing visibility into the effects of specific behaviors by a team or an adjacent one on performance. Dashboards include common visualization techniques, such as:

Tables: This consists of rows and columns used to compare variables. Tables can show a great deal of information in a structured way, but they can also overwhelm users that are simply looking for high-level trends.

Pie charts and Stacked Bar Charts: These graphs are divided into sections that represent parts of a whole. They provide a simple way to organize data and compare the size of each component to one other.

Line graphs and Area charts: These visuals show change in one or more quantities by plotting a series of data points over time. Line graphs utilize lines to demonstrate these changes while area charts connect data points with line segments, stacking variables on top ofone another and using color to distinguish between variables.

Histograms: This graph plots a distribution of numbers using a bar chart (with no spaces between the bars), representing the quantity of data that falls within a particular range. This visual makes it easy for an end user to identify outliers within a given dataset.

Scatter plots: These visuals are beneficial in revealing the relationship between two variables, and they are commonly used within regression data analysis. However, these can sometimes be confused with bubble charts, which are used to visualize three variables via the x-axis, the y-axis, and the size of the bubble.

Heat maps: These graphical displays are helpful in visualizing behavioral data by location. This can be a location on a map, or even a webpage.

Tree maps: These display hierarchical data as a set of nested shapes, typically rectangles. Treemaps are great for comparing the proportions between categories via their area size.

Data Visualization in Python

Data visualization in python is perhaps one of the most utilized features for data science with python in today‟s day and age. The libraries in python come with lots of different features that enable users to make highly customized, elegant, and interactive plots.

Useful packages for visualizations in python are Matplotlib, Seaborn, Statistics, Pandas and Numpy.

Matplotlib: Matplotlib is a visualization library in Python for 2D plots of arrays. Matplotlib is written in Python and makes use of the NumPy library. It can be used in Python and IPython shells, Jupyter notebook, and web application servers. Matplotlib comes with a wide variety of plots like line, bar, scatter, histogram, etc., which can help us, into understanding trends, patterns, correlations.

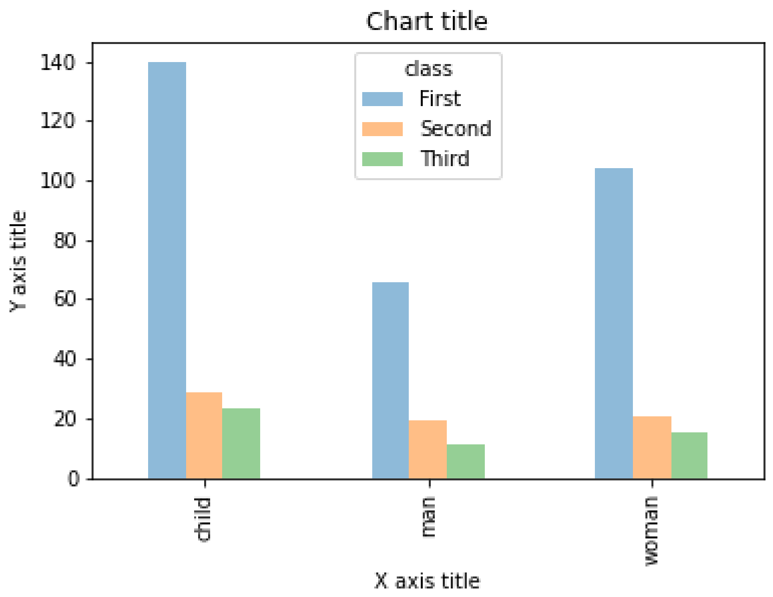
Seaborn: Seaborn is a dataset-oriented library for making statistical representations in Python. It is developed atop matplotlib and to create different visualizations. It is integrated with pandas data structures. The library internally performs the required mapping and aggregation to create informative visuals.

Grouped bar chart: A grouped bar chart is used when we want to compare the values in certain groups and sub-groups

Grouped bar chart using Matplotlib: In the following code a data set named as Titanic has been used which can easily be downloaded from Kaggle.

```
# Program 4.7: Grouped Bar Chart using Matplotlib
#Creating the dataset
import seaborn as sns
import matplotlib.pyplot as plt
df = sns.load_dataset('titanic')
df_pivot = pd.pivot_table(df, values="fare", index="who", columns="class",
aggfunc=np.mean)
#Creating a grouped bar chart
ax = df_pivot.plot(kind="bar", alpha=0.5)
#Adding the aesthetics
plt.title('Chart title')
plt.xlabel('X axis title')
plt.ylabel('Y axis title')
# Show the plot
plt.show()
```

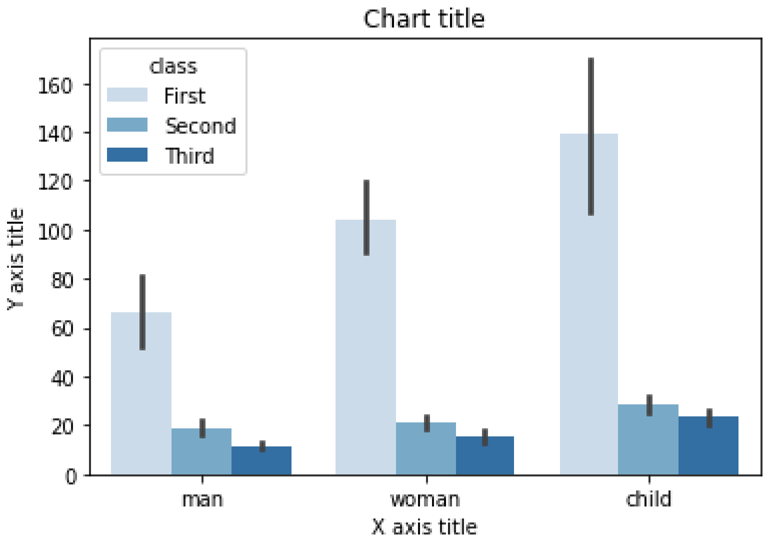
Output:



Grouped bar chart using Seaborn:

```
# Program 4.8: Grouped bar chart using seaborn
#Reading the dataset
import seaborn as sns
import matplotlib.pyplot as plt
titanic_dataset = sns.load_dataset('titanic')
#Creating the bar plot grouped across classes
sns.barplot(x = 'who', y = 'fare', hue = 'class', data = titanic_dataset, palette = "Blues")
#Adding the aesthetics
plt.title('Chart title')
plt.xlabel('X axis title')
plt.ylabel('Y axis title')
# Show the plot
plt.show()
```

Output:



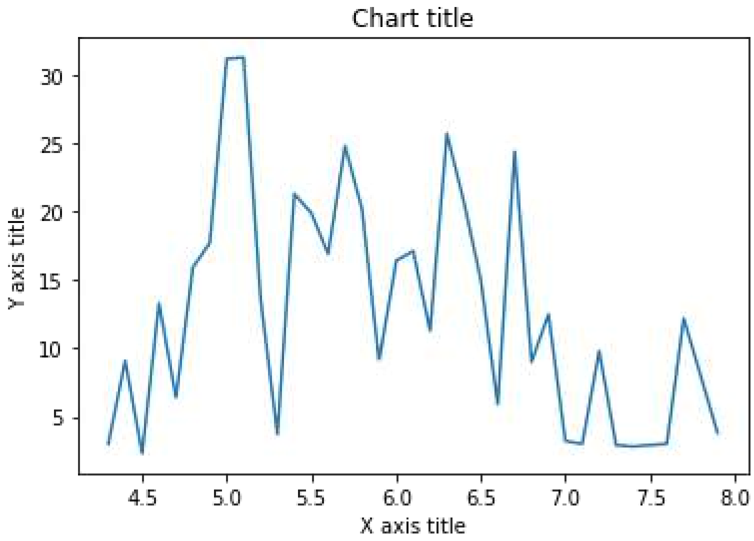
Line chart: A line chart is used for the representation of continuous data points. This visual can be effectively utilized when we want to understand the trend across time.

In the following example "iris" dataset has been used which can be downloaded from Kaggle.

```
#Program 4.9: Line Chart
#Creating the dataset
import seaborn as sns
import matplotlib.pyplot as plt
```

```
df = sns.load_dataset("iris")

df=df.groupby('sepal_length')['sepal_width'].sum().to_frame().reset_index()

#Creating the line chart

plt.plot(df['sepal_length'], df['sepal_width'])

#Adding the aesthetics

plt.title('Chart title')

plt.xlabel('X axis title')

plt.ylabel('Y axis title')

#Show the plot

plt.show()
```

Output:

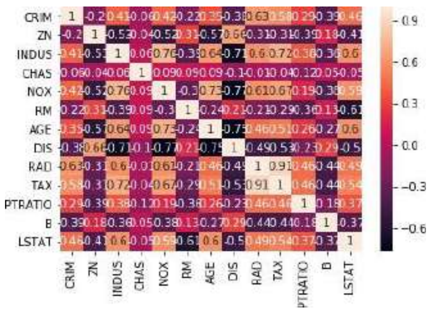


HeatMaps: The Heat Map procedure shows the distribution of a quantitative variable over all combinations of 2 categorical factors. If one of the 2 factors represents time, then the evolution of the variable can be easily viewed using the map. A gradient color scale is used to represent the values of the quantitative variable. The correlation between two random variables is a number that runs from -1 through 0 to +1 and indicates a strong inverse relationship, no relationship, and a strong direct relationship, respectively.

It can be created using following statement using boston data set as used in the earlier section

```
Corr_mat = boston_df.corr().round(2)

Sns.heatmap(data=corr_mat, annot=True)
```

Output:



Other types of graphs have already been discussed in the previous modules.

4.3 SUMMARY

In this module, three important topics have been discussed. The topics discussed in this module are descriptive statistics, exploratory data analysis and data visualization. All thethree topic are related to each other. Starting with descriptive statistics involves the steps to analyse the data using the measures discussed in previous two modules i.e. module-II and module-III. The descriptive statistics explain the measures of central tendency, measure of dispersion along with the frequency distribution used for the given data. Following this exploratory data analysis involves the various steps of data analysis such as identifying data sources, cleaning, presenting and interpreting. When we talk about data presenting, data visualization comes into picture. Data visualization involves the presentation of data using suitable means i.e. graphs so that it becomes easy to understand and interpret. The implementation of these topics have been explained using python.

4.4 PRACTICE QUESTIONS

Q.1 What is meant by descriptive statistics? Describe the purpose of descriptive statistics.

Q.2 Explain various types of descriptive statistics in detail.

Q.3 Describe various functions in python for performing descriptive statistics.

Q.4 What do you mean by exploratory data analysis. Explain its significance.

Q.5 Write and explain the various steps for performing exploratory data analysis.

Q.6 What are the various packages in python that can be used for performing exploratory data analysis? Explain in detail.

Q.7 How can you draw (i) histogram (ii) bar chart (iii) heat map (iv) grouped bar chart (v) pie chart, in python? Explain by writing suitable programs.

Q.8 What is meant by data visualization? Explain the various steps in data visualization.

Q.9 Discuss various tools used for data visualization.

REFERENCES

1. A. Abebe, J. Daniels, J.W.Mckean, “Statistics and Data Analysis”.
2. A. Martelli, A. Ravenscroft, S. Holden, “Python in a Nutshell”, OREILLY.

3. Clarke, G.M. & Cooke, D., “A Basic course in Statistics”, Arnold.
4. David M. Lane, “Introduction to Statistics”.
5. Eric Matthes, “Python Crash Course: A Hands-On, Project-Based Introduction to Programming”.\
6. S.C.Gupta and V.K.Kapoor, “Fundamentals of Mathematical Statistics”, Sultan Chand & Sons, New Delhi.
7. Weiss, N.A., “Introductory Statistics”, Addison Wesley

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

UNIT V: CORRELATION AND REGRESSION

STRUCTURE

5.0 OBJECTIVES

The main goal of this module is to help students learn, understand and practice the basics of statistics which will helpful to do the research in the social sciences. In this module you will learn the basics of statistics which covers two fundamentals concepts correlation and regression. The examples of this module were calculated manual as well as using programming language. Python language is used in this module.

5.1 INTRODUCTION

Data science become a buzzword that everyone talks about the data science. Data science is an interdisciplinary field that combines different domain expertise, computer programming skills, mathematics and statistical knowledge to find or extract the meaningful or unknown patterns from unstructured and structure dataset.

Data science is useful for extraction, preparation, analysis and visualization of data. Various statistical methods can be applied to get insight in the data.

Data science is all about using data to solve problems. Data has become the fuel of industries. It is most demandable field of 21st century. Every industry require data to functioning, searching, marketing, growing, expanding their business.

The application of areas of data science are health care, fraud detection, disease predicting, real time shipping routes, speech recognition, targeting advertising, gaming and many more.

5.2 CORRELATION

Correlation is a statistical measure that expresses the extent to which two or more variables are changes together at a constant rate. It is a relationship between two or more variables. Thedata can be represented by the pairs (x, y) where x is an independent variable and y is a dependent variable.

Correlations are useful for describing simple relationships among the data. It quantifies the degree and direction to which two variables are related. It is a measure of the extent to which two variables are related.

5.3 TYPES of CORRELATION

The different types of correlations on the basis of: the degree of correlation, numbers of variables used and linearity.

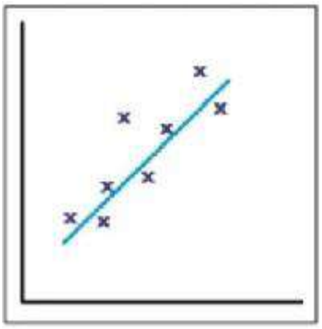
5.3.1 Positive, Negative and Zero Correlation

The positive, negative and zero correlations are basis on the degree of correlation.

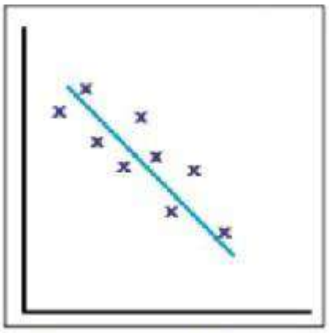
- **Positive Correlation** – A relationship between two variables in which both the variables move in same direction, it means that when one variable is increases as the other variable increases, or one variable is decreases while the other decreases. The examples of positive correlation are: height and weight of person, price and supply of a commodity, etc.
- **Negative Correlation** – A relationship between two variables in which both the variable moves in opposite direction. That means when one variable is increase as the other variable decreases, or one variable decrease while the other increases.

The examples of negative correlation are: no. of absent and grade of student, speed of train and time to reach destination, etc.

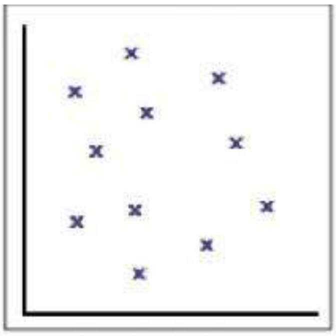
- **No / Zero Correlation –** There is no linear dependence or no relationship between two variables. The example of no / zero correlation is the coffee drunk and level of IQ.



Positive Correlation



Negative Correlation



No Correlation

5.3.2 Simple, Multiple and Partial Correlation

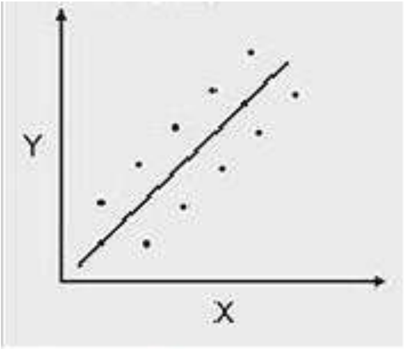
The simple, multiple and partial correlations are depending on the number of variables studied in the analysis.

- **Simple Correlation –** There are only two variables are studied and check the correlation between them is called simple correlation. Examples of simple correlation are age and height of students, price and demand of commodities.
- **Multiple Correlation –** There are three or more variables are studied for correlation simultaneously is called multiple correlation. Example of multiple correlation is to study the relationship between the yield of any crop, amount of fertilizers used and amount of rainfall.
- **Partial Correlation –** There are three or more variables are studied. When one or more variables are kept constant and the relationship is studied between others is called partial correlation. Example of partial correlation is the price of ice-cream, temperature and demand. If we kept the price of ice-cream is constant and studied the correlation between temperature and demand of ice-cream.

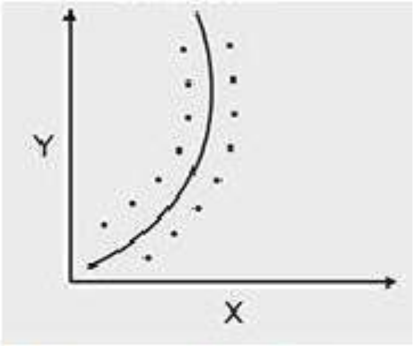
5.3.3 Linear and Non-Linear Correlation

The linear and non-linear correlation are on the basis of linearity of data.

- **Linear Correlation –** The correlation is said to be a linear if the ratio of change of the two variables is constant. If we plot all the points on the scatter diagramtend to lie near a line which like a straight line.
- **Non-Linear Correlation –** The correlation is said to be a non-linear or curvilinear if the ratio of change of the two variables is not constant. If we plot all the points on the scatter diagram tend to lie near a smooth curve which like a curve.



Linear Correlation



Non-Linear Correlation

The **corr()** function is used to find the correlation between two variables in Python. Here we take an example of boys age and weight to calculate the correlation.

Example: Find the correlation of boys age and weight.

Age	10	20	30	40	50	60	70	80
Weight	32	48	59	64	70	67	78	82

The below code will find the correlation of boys age and weight.

```
# Importing library
import pandas as pd

# Data of Age and Weight
Age = pd.Series([10, 20, 30, 40, 50, 60, 70, 80])
Weight = pd.Series([32, 48, 59, 64, 70, 67, 78, 82])

#Calculating Correlation
correlation = Age.corr(Weight)
correlation
```

The above code will calculate the correlation between age and weight of boys as follow:

```
0.9501687384314103
```

Example: Find the correlation of boys age and weight.

Age	10	20	30	40	50	60	70	80
Weight	32	48	59	64	70	67	78	82

The below code will find the correlation of boys age and weight.

```
# Importing library
import pandas as pd

# Data of Age and Weight
data = {
    "Age" : [10, 20, 30, 40, 50, 60, 70, 80],
    "Weight" : [32, 48, 59, 64, 70, 67, 78, 82]
            }

# Creation of Dataframe
df = pd.DataFrame(data)

# Creation of Correlation Matrix
df.corr()
```

The above code will calculate the correlation between age and weight of boys as follow:

	Age	**Weight**
Age	1.000000	0.950169
Weight	0.950169	1.000000

Now we take an example of three subjects mark of science student to calculate the correlation.

Example: Find the correlation between marks of three subjects maths, chemistry and physics of science students.

Maths	100	86	90	80	96	95	92	99
Chemistry	88	90	80	90	90	86	94	88
Physics	92	88	89	94	90	87	93	91

The below code will find the correlation between marks of three subjects such as maths, chemistry and physics of science students.

```
# Importing library
import pandas as pd

data = {
    "Maths" : [100, 86, 90, 80, 96, 95, 92, 99],
    "Chemistry" : [88, 90, 80, 90, 90, 86, 94, 88],
    "Physics" : [92, 88, 89, 94, 90, 87, 93, 91]
             }

df = pd.DataFrame(data)
```

```
df.corr()
```

The above code will calculate the correlation between maths, chemistry and physicssubject marks of science students as follow:

	Maths	Chemistry	Physics
Maths	1.000000	-0.096004	-0.180719
Chemistry	-0.096004	1.000000	0.502519
Physics	-0.180719	0.502519	1.000000

From the above table we can see that, the correlation between maths and chemistry is negative (-0.096004), between maths and physics is also negative (-0.180719) and between chemistry and physics is positive (0.502519).

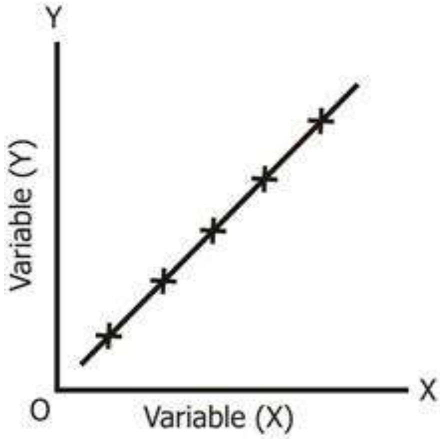
5.4 TECHNIQUES FOR MEASURING CORRELATION

The different techniques for measuring correlation are: graphical method and algebraic method. Scatter diagram is a graphical method. Karl Pearson’s correlation coefficient and Spearman’s rank correlation coefficient are algebraic methods.

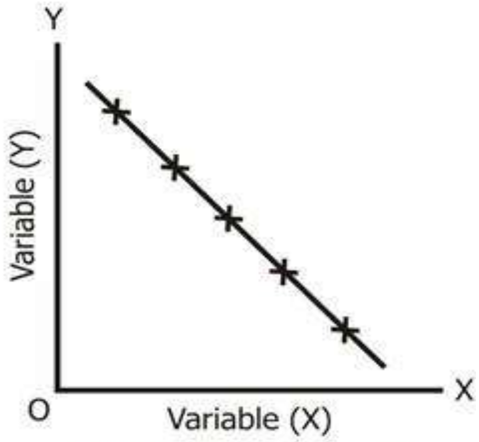
5.4.1 Scatter Plot

Scatter diagram is a graph in which the values of two variables are plotted along with two axes. It is a most basic type of plot that helps you visualize the relationship between two variables. Each value in this plot is represent by a dot. It is a set of dottedpoints to represent the data on both horizontal and vertical axis.

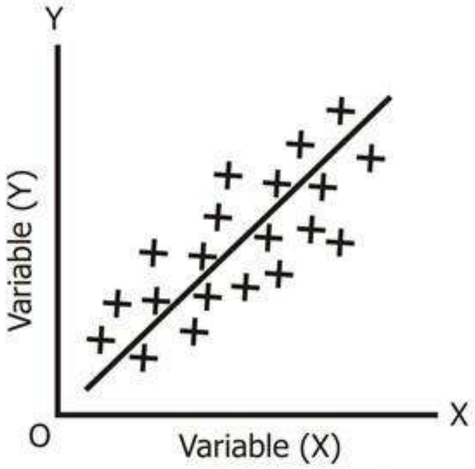
- **Perfect positive correlation :** All the plotted points are on a straight line which is rising from lower left corner to the upper right corner in a scatter diagram.
- **Perfect negative correlation :** All the plotted points are on a straight line which is rising from upper left corner to the lower right corner in a scatter diagram.
- **Strong positive correlation :** All the plotted points are closer to a straight line which is rising from lower left corner to the upper right corner in a scatterdiagram.
- **Strong negative correlation :** All the plotted points are closer to a straight line which is rising from upper left corner to the lower right corner in a scatterdiagram.
- **Weak positive correlation :** All the plotted points are not closer (lie away) to a straight line which is rising from lower left corner to the upper right corner in a scatter diagram.
- **Weak negative correlation :** All the plotted points are not closer (lie away) to a straight line which is rising from upper left corner to the lower right corner in a scatter diagram.
- **No correlation :** All the plotted points are scattered randomly across the graph.



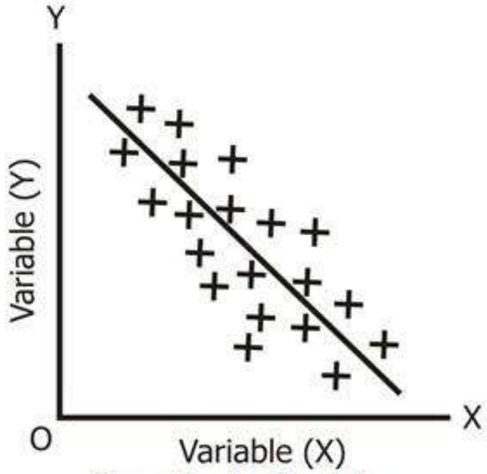
Perfect Positive Correlation



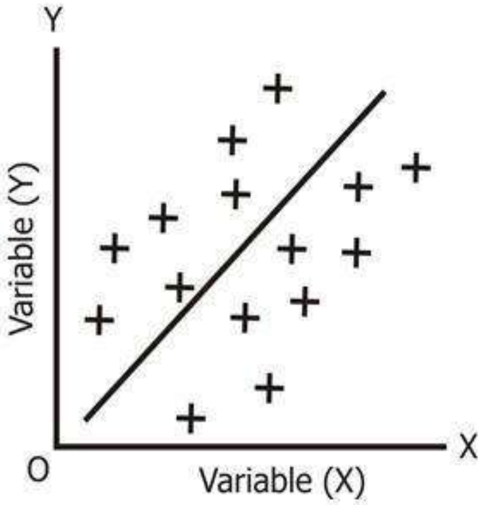
Perfect Negative Correlation



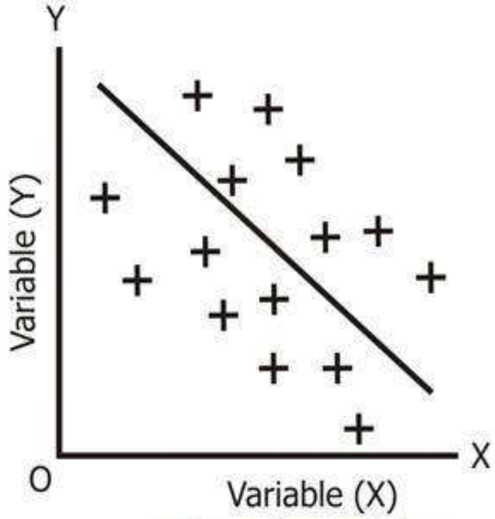
Strong Positive Correlation



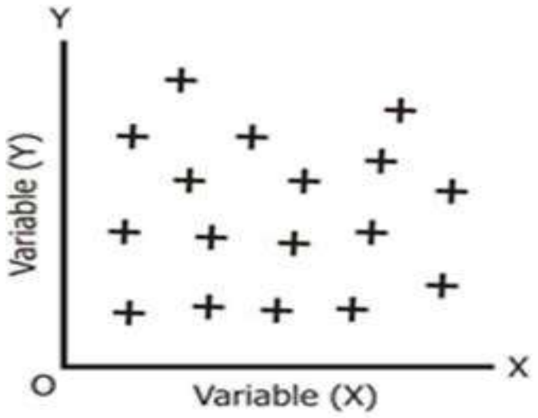
Strong Negative Correlation



Weak Positive Correlation



Weak Negative Correlation



No Correlation

The **scatter()** function is used to draw the scatter plot in Python. It plots one dot for each observation. It required two different set of observation with same length for both the axis.

Here we take an example of boys age and weight. The x-axis represents age and y-axis represents weight of boys.

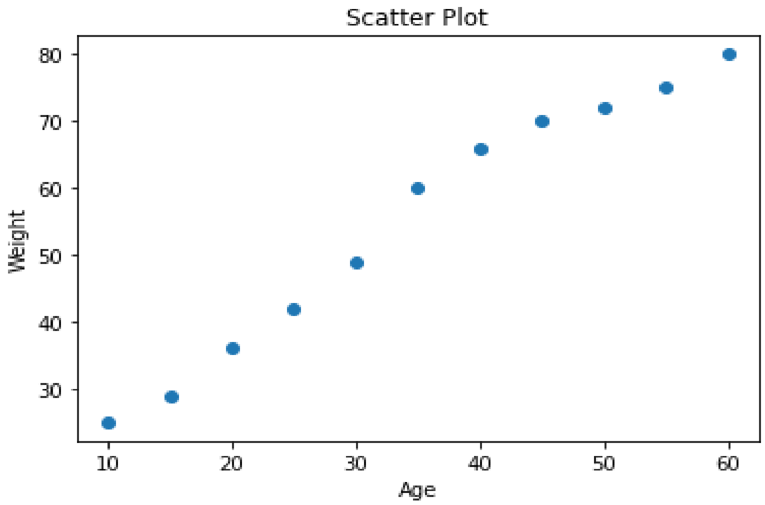
Example: Plot the scatter diagram of age and weight of boys.

Age	10	15	20	25	30	35	40	45	50	55	60
Weight	25	29	36	42	49	60	66	70	72	75	80

The below code will plot the scatter diagram of age and weight of boys.

```
# Importing library
import matplotlib.pyplot as plt

# Data of Age and Weight
Age = [10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60]
Weight = [25, 29, 36, 42, 49, 60, 66, 70, 72, 75, 80]

# Plotting scatter plot with title and label
plt.scatter(Age,  Weight)
plt.title("Scatter Plot")
plt.xlabel("Age")
plt.ylabel("Weight")

# Show plot
plt.show()
```

The above code will create scatter plot as follow:



In the above scatter plot we can see the relationship between two variables age and weight. There is a positive correlation between boys age and weight.

Another method is to draw scatter plot using ***"kind"*** parameter. Here we take an example of two subject marks. The x-axis represents chemistry and y-axis represents physics subject marks.

Example: Plot the scatter diagram of chemistry and physics subject marks.

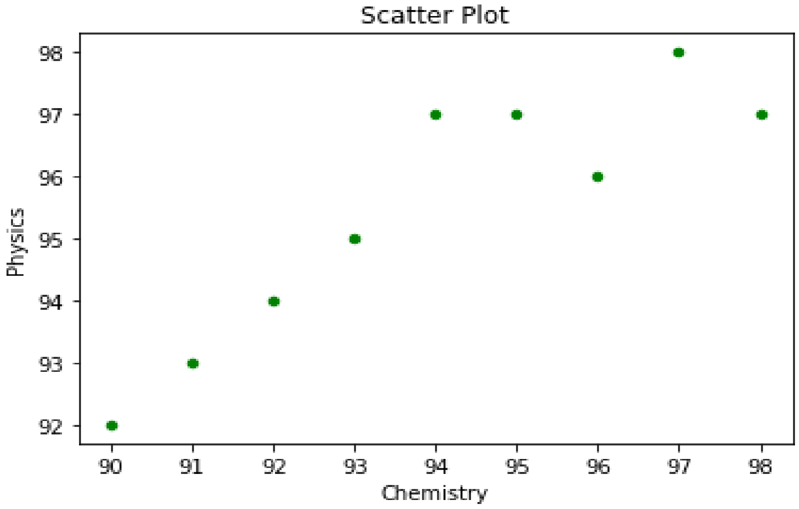
Chemistry	90	92	97	98	96	94	91	93	95
Physics	92	94	98	97	96	97	93	95	97

The below code will plot the scatter diagram of chemistry and physics subject marks.

```
# Importing library
import pandas as pd
import matplotlib.pyplot as plt

# Data of Chemistry and Physics subject marks
data = {
  "Chemistry_Marks" : [90, 92, 97, 98, 96, 94, 91, 93, 95],
  "Physics_Marks" : [92, 94, 98, 97, 96, 97, 93, 95, 97]
             }



# Dataframe Creation
df = pd.DataFrame(data)

# Plotting scatter plot with title and label
df.plot(x='Chemistry_Marks', y='Physics_Marks', kind = 'scatter', color = 'Green')
plt.title("Scatter Plot")
plt.xlabel("Chemistry")
plt.ylabel("Physics")
```

```
# show the plot
plt.show()
```

The above code will create scatter plot as follow:



In the above scatter plot we can see the relationship between chemistry and physics subject marks.

5.4.2 Karl Pearson Coefficient of Correlation

This is the most common measure of correlation. It is also known as Pearson Product Moment Correlation (PPMC). It is used to measure the strength and direction of the linear relationship between two variables in correlation analysis. It is used to measure the degree of association between variables. The correlation coefficient is symbolized by r which lies between -1 and 1.

- The value of r is 1 indicates perfect positive correlation.
- The value of r is -1 indicates perfect negative correlation.
- The value of r is 0 indicates no relationship.

The following formula is used to calculate the Pearson correlation.

$$r = \frac{n\,(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n \sum x^2 - (\sum x)^2][n \sum y^2 - (\sum y)^2]}}$$

Here,

n = No. of values
ΣX = Total of the First Variable Value
ΣY = Total of the Second Variable Value
ΣXY = Sum of the Product of first and Second Value
ΣX^2 = Sum of the Squares of the First Value
ΣY^2 = Sum of the Squares of the Second Value

Example : Calculate the Karl Pearson’s Coefficient of correlation from the following data.

X	10	20	30	40	50	60	70	80
Y	32	48	59	64	70	67	78	82

Solution:

X	**Y**	**XY**	**X^2**	**Y^2**
10	32	320	100	1024
20	48	960	400	2304
30	59	1770	900	3481
40	64	2560	1600	4096
50	70	3500	2500	4900
60	67	4020	3600	4489
70	78	5460	4900	6084
80	82	6560	6400	6724
ΣX = 360	**ΣY = 500**	**ΣXY = 25150**	**ΣX^2 = 20400**	**ΣY^2 = 33102**

Now, we calculate

$$r = \frac{n\,(\sum xy) - (\sum x)\,(\sum y)}{\sqrt{[\,n \sum x^2 - (\sum x)^2\,][\,n \sum y^2 - (\sum y)^2\,]}}$$

$$r = \frac{8\,(25150) - (360)(500)}{\sqrt{[\,8\,(20400) - (360)(360)][\,8\,(33102) - (500)(500)\,]}}$$

$$r = \frac{201200 - 180000}{\sqrt{[\,163200 - 129600\,]\,[\,264816 - 250000\,]}}$$

$$r = \frac{21200}{\sqrt{(\,33600\,)\,(14816)}}$$

$$r = \frac{21200}{\sqrt{497817600}}$$

$$r = \frac{21200}{22311.826}$$

$$r = 0.9501687$$

Now we will calculate the Karl Pearson’s Coefficient of correlation using Python.

The below code will calculate the Karl Pearson’s Coefficient.

```
# Importing library
import pandas as pd

# Data of Age and Weight
data = {"Age" : [10, 20, 30, 40, 50, 60, 70, 80],
       "Weight" : [32, 48, 59, 64, 70, 67, 78, 82]}

df = pd.DataFrame(data)

#Calculating Correlation
df.corr(method='pearson')
```

The above code will give the following result :

	Age	**Weight**
Age	1.000000	0.950169
Weight	0.950169	1.000000

Here we can see that the value of Karl Pearson's Coefficient of correlation (r) is 0.950169 which is same as the manual calculation as above.

5.4.3 Spearman's Rank Correlation Coefficient

In year 1904, Charles Edward Spearman introduced a new method of measuring the correlation between two variables. It is applicable to individual observation. In this method, the rank (or order) of the observation is taken instead of the value of variable. This correlation coefficient is called rank correlation coefficient. This is useful to measure the qualitative characteristics such as beauty, honesty, height etc.

The following formula is used to calculate the Spearman's Rank Coefficient of correlation.

$$r = 1 - \frac{6\ \sum D^2}{n\ (n^2 - 1)}$$

Here,

r = Spearman rank correlation coefficient
n = No. of pairs of observation
D = Differences in ranks between pair observation

Example : Calculate the Spearman's Rank Coefficient of correlation from the following data.

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

Solution:

Rank (X)	**Rank (Y)**	**D = R (X) – (Y)**	**D^2**
1	6	-5	25
6	4	2	4
5	9	-4	16
10	8	2	4
3	1	2	4
2	2	0	0
4	3	1	1
9	10	-1	1
7	5	2	4
8	7	1	1
N = 10			**$\sum$D^2 = 60**

Now, we calculate

$$r = 1 - \frac{6 \sum D^2}{n\ (n^2 - 1)}$$

$$r = 1 - \frac{6\ (60)}{10\ (100 - 1)}$$

$$r = 1 - \frac{360}{990}$$

$$r = 1 - 0.36363$$

$$r = 0.63636$$

Now we will calculate the Spearman’s Rank Coefficient of correlation using Python.

The below code will calculate the Spearman’s Rank Coefficient.

```
# Importing library
import pandas as pd

# Data
data = {"X" : [1,6,5,10,3,2,4,9,7,8],
       "Y" : [6,4,9,8,1,2,3,10,5,7]}

df = pd.DataFrame(data)

#Calculating Correlation
df.corr(method='spearman')
```

The above code will give the following result :

	X	Y
X	1.000000	0.636364
Y	0.636364	1.000000

Here we can see that the value of Spearman's Rank Coefficient of correlation (r) is 0.636364 which is same as the manual calculation as above.

5.5 REGRESSION ANALYSIS

Regression analysis is a set of various statistical methods for estimating the relationship between a dependent variable (also called outcome variable) and one or more independent variables (also called predictor variable). This is widely used to predict the outputs, forecasting, time series analysis and finding the causal effect dependencies.

There are many different types of regression techniques based on the number of dependent variables, the dimensionality of regression line and the types of dependent variable such as linear regression, polynomial regression, decision tree regression, support vector regression, random forest regression, logistic regression, etc. The most frequently used regression analysis is linear regression.

The following terms are related to regression analysis:

- **Dependent variable or target variable:** Variable to predict.
- **Independent variable or predictor variable:** Variables to estimate the dependent variable.
- **Outlier:** The observation which differs significantly from the other observation.
- **Normality**: The data follows a normal distribution.
- **Multicollinearity:** It is a situation in which two or more independent variables are highly linearly related.
- **Homoscedasticity:** It is a situation in which the error term is the same across all values of the independent variables. It is also called homogeneity of variance.

5.6 APPLICATIONS OF REGRESSION ANALYSIS

Regression analysis refers to a group of techniques for studying the relationships among two or more variables based on a sample. Linear regression is one of the most commonly used techniques in statistics which is used to quantify the relationship between one or more predictor variables and a response variable.

Regression analysis is used for prediction and forecasting. This statistical method is used in various sectors. The most common applications of regression analysis are:

- **Financial Sector** – Regression analysis is used to calculate the Beta (volatility of returns relative to the overall market) which is used as a measure of risk. A company with a higher beta has a greater risk and greater expected returns also. It is also used in stock market to understand the trend of stocks and forecast the prices of different stocks.

- **Marketing Sector** – Regression analysis is used to understand the effectiveness of various marketing campaigns, forecasting the sales and pricing of the products. It is also used to measure the strength of a relationship between different variables such as customer satisfaction with product quality and price of product.

- **Manufacturing Sector** – Regression analysis is used to evaluate the relationship of variables that determine to define a better engine to provide better performance. An important application of regression analysis in manufacturing sector is to estimate the cost of product. In manufacturing industries, an accurate cost prediction during a new product development process is most important factor for manufacturing firms to survive in this competition era.

- **Sales and Promotion Sector** – Regression analysis is used to analyses promotions on sales of product. It is used to find the relationship between the amount spent on advertising on a product and determines the amount of its sales. It finds the return on investment such as a company want to find the amount that have invested in marketing of particular products or brands and sales of that product.

- **Medical Sector** – Regression analysis is an important statistical method for theanalysis of medical data. It enables the identification and characterization of relationships among multiple factors in online public health data. It is used for prediction and clarification can both be appropriate for public health data analysis for better understanding of public health outcomes. It is also used to forecast the different combination of medicines to prepare generic medicines for diseases.

5.7 TYPES OF REGRESSION ANALYSIS

There are three different types of regression analysis as follows:

- Simple Linear Regression
- Multiple Linear Regression
- Non-Linear Regression

5.7.1 Simple Linear Regression

Simple linear regression analysis is a statistical tool for finding the best relationship between one independent (predictor or explanatory) variable and one dependent (response, outcome) variable which is continuous in nature. Linear Regression is a predictive model used for finding the linear relationship between a dependent variable and one or more independent variables. This relationship represents how an input variable is related to the output variable and how it is represented by a straight line.

The simple linear regression model is expressed by using the following equation:

Y = a + bX + □

Here,

Y = Dependent variable
a = Intercept
b = Slop
X = Independent variable
□ = Error (residual)

Example : Obtain the equation of the lines of regression from the following data. Also estimate the value of Y for X = 28.

X	10	20	30	40	50	60	70	80
Y	32	48	59	64	70	67	78	82

Solution:

X	**Y**	**XY**	**X^2**	**Y^2**
10	32	320	100	1024
20	48	960	400	2304
30	59	1770	900	3481
40	64	2560	1600	4096
50	70	3500	2500	4900
60	67	4020	3600	4489
70	78	5460	4900	6084
80	82	6560	6400	6724
ΣX = 360	**ΣY = 500**	**ΣXY = 25150**	**ΣX^2 = 20400**	**ΣY^2 = 33102**

Now, we calculate

$$\bar{X} = \frac{\sum x}{n} = \frac{360}{8} = 45$$

$$\bar{Y} = \frac{\sum y}{n} = \frac{500}{8} = 62.5$$

$$SSxy = \sum xy - \frac{(\sum x)(\sum y)}{n} = 25150 - \frac{(360)(500)}{8} = 25150 - 22500 = 2650$$

$$SSxx = \sum x^2 - \frac{(\sum x^2)}{n} = 20400 - \frac{(360)^2}{8} = 20400 - 16200 = 4200$$

$$SSyy = \sum y^2 - \frac{(\sum y^2)}{n} = 33102 - \frac{(500)^2}{8} = 33102 - 31250 = 1852$$

$$b = \frac{SSxy}{SSxx} = \frac{2650}{4200} = 0.63095$$

$$a = \bar{Y} - b\bar{X} = 62.5 - (0.63095)(45) = 62.5 - 28.39 = 34.11$$

So, we get the following equation for regression line.

Y = 34.22 + 0.63095 X

Now, we can use this equation for prediction. So, we can predict the value of Y, when X = 28.

So, Y = 34.22 + 0.63095 X = 34.22 + 0.63095 (28) = 34.22 + 7.66 = 51.77

From the above example, we get the predicted value of Y is 51.77 when the value of X is 28.

Now we will calculate the intercept and slope using Python.

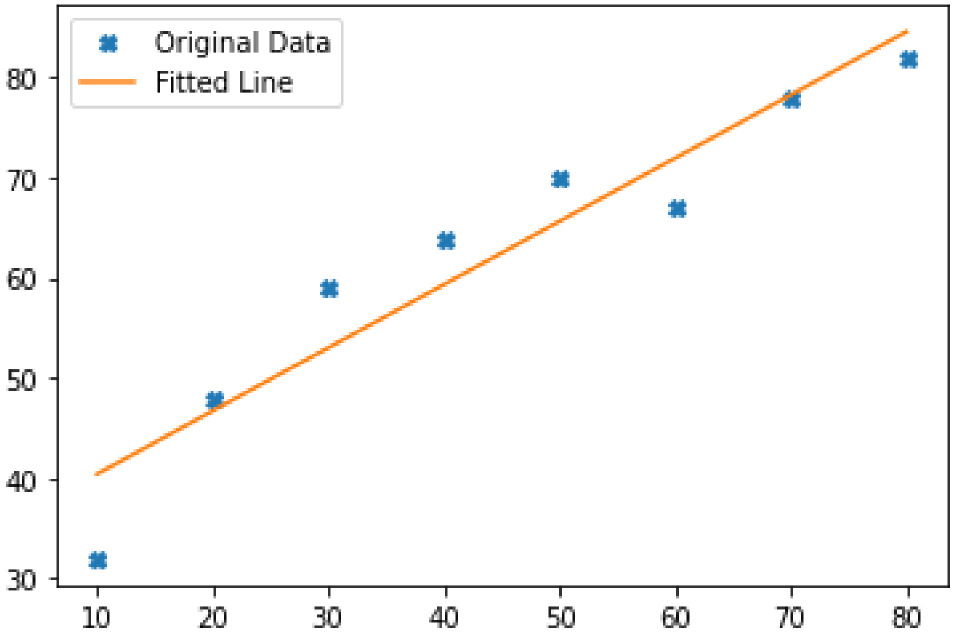
The below code will calculate the intercept and slope and to get the equation of the line of regression. It will also estimate the value of Y for X = 28.

```
# Importing library
import matplotlib.pyplot as plt
from scipy import stats

# Data
x = [10, 20, 30, 40, 50, 60, 70, 80]
y = [32, 48, 59, 64, 70, 67, 78, 82]

# Calculation slop and intercept
slope, intercept, r, p, std_err = stats.linregress(x, y)
print("Slop =", slope)
print("Intercept =", intercept)
print("R value =", r)
print("P value =", p)
print("Standard Error =", std_err)

# Function
def reglinefun(x):
  return intercept + slope * x

# Line
line = list(map(reglinefun, x))

# Plotting
```

```
plt.plot(x, y, 'X', label="Original Data")
plt.plot(x, line, label="Fitted Line")
plt.legend()
plt.show()

# Prediction
predict = reglinefun(28)
print("Predicted value =", predict)
```

The above code will give the following result :

```
Slop = 0.6309523809523809
Intercept = 34.10714285714286
R value = 0.9501687384314103
P value = 0.0002979006863309245
Standard Error = 0.08450983564345486
```



From the above code we can get the estimated value of Y is 51.77 when X = 28.

```
Predicted value = 51.773809523809526
```

5.8 SUMMARY

The students will learn many things related to basic statistics in this module and they will be able to calculate the measures of statistics such as correlation and regression. They will also able to perform the various statistical analysis using Python.

- Ability to understand the correlation in the statistics.

- Ability to understand the various types of correlation such as positive, negative, zero, simple, partial, multiple, linear and non-linear correlation.
- Ability to do understand the methods of studying correlation using scatter diagram.
- Ability to calculate the Karl Pearson correlation coefficient and Spearman rank correlation coefficient.
- Ability to understand the regression analysis and applications of regression analysis.
- Ability to obtain the line of regression and predication using simple linear regression.

5.9 PRACTICE QUESTIONS

Short Answer:

1. What is correlation?
2. Define positive and negative correlation.
3. Define simple correlation and multiple correlation.
4. What is partial correlation?
5. Define linear and non-linear correlation
6. What is regression?
7. List types of regression.

Long Answer:

1. What is correlation? Explain types of correlation.
2. Explain scatter diagram for correlation.
3. Explain Karl Pearson correlation coefficient with example.
4. Explain Spearman Rank correlation coefficient with example.
5. Explain applications of regression analysis.
6. Explain simple linear regression with example.

PRACTICALS

1. Find the correlation between height of father and son (in cm) of the following.

Height of Father	65	66	67	67	68	69	71	73
Height of Son	64	68	65	69	71	70	69	71

2. Find the correlation between marks obtained by eight students in physics, chemistry and biology.

Physics	75	56	70	82	64	78	49	59
Chemistry	80	67	67	75	70	60	55	68
Biology	67	64	68	77	72	73	58	62

3. Draw the scatter plot of age of father and daughter (in year) of the following.

Age of Father	40	45	48	50	51	54	58	60
Age of Daughter	12	13	18	20	22	26	28	32

4. Calculate the Karl Pearson correlation coefficient of the following:

Price	10	20	30	40	50	60	70
Supply	8	7	13	15	11	18	20

5. Calculate the Karl Pearson correlation coefficient of the following:

Age of Husband	25	27	28	30	30	31	34	38
Age of Wife	23	26	27	29	26	28	34	36

6. Calculate the Spearman Rank correlation coefficient of the following:

X	65	66	67	67	68	69	71	73
Y	64	68	65	69	71	70	69	71

7. Calculate the Spearman Rank correlation coefficient of the following:

X	4	3	8	7	1	5	2	6
Y	6	2	7	5	3	4	1	8

8. Obtain the equation of line of regression from the following data.

X	65	66	67	67	68	69	71	73
Y	64	68	65	69	71	70	69	71

9. Obtain the line of regression equation age and blood pressure of the following and predict the value for blood pressure when age is 50 year.

Age	42	55	61	38	68	74	64	46	58	70
Blood Pressure	126	140	148	121	145	160	155	130	142	156

REFERENCES

Books

1. Gupta, S.C. and Kapoor, V.K.: Fundamentals of Mathematical Statistics, Sultan Chand & Sons, New Delhi, 11th Ed
2. Hastie, Trevor, et al. The Elements of Statistical Learning, Springer
3. Ross, S.M.: Introduction to Probability and Statistics for Engineers and Scientists, Academic Press
4. Papoulis, A. and Pillai, S.U.: Probability, Random Variables and Stochastic Processes, McGraw Hill

Web References

1. https://www.geeksforgeeks.org

2. https://www.tutorialspoint.com
3. https://www.w3schools.com
4. https://pandas.pydata.org
5. https://pbpython.com
6. https://www.statisticshowto.com
7. https://realpython.com

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

Unit 6: Mathematical and Statistical Probability

STRUCTURE

6.0 OBJECTIVES

The main goal of this module is to teach you the basics of random experiment and related terms like sample space, sample points, events, type of events and algebra of events which are the building blocks for learning basic concepts of probability and probability theory. By studying this module you should be able to:

- Understand what is random experiment,
- define the sample space of a random experiment
- identify events, types of event,
- Learn about the algebra of events and some algebraic properties of events.

Some examples are given at the end of each definition to understand the basic concept. Few examples are done using python programming language.

6.1 INTRODUCTION

In our day to day life, we came across situations, where we cannot predict the result of our action or outcome of experiment we are conducting. But we may know that the outcome has to one of the several possibilities. For example, (a) gender of newly born baby cannot be predicted before birth but we know that the gender of newly born baby has to be either “Male” or “Female”, (b) result of class 12 student cannot be predicted before the declaration of the result, but we know that the result has to be either “Pass” or “Fail”, (c) when a coin is tossed, we cannot predict the outcome of experiment before the completion of the experiment, but we know that the outcome has to be either “Head” (H) or “Tail” (T). Such experiments, whose outcomes cannot be predicted in advance before the completion of experiment, are called random experiment.

The concepts of such experiments are fundamental to study the theory of probability. Before learning the probability theory, we need to understand some basic terminology like random experiment, sample space, event, types of events and some algebraic operations on events. In subsequent sections, you will learn about all these basic terms with some illustrations.

6.2 RANDOM EXPERIMENT

An experiment is a process that generates well defined outcomes (result). In an experiment, if all possible outcomes of an experiment are known in advance but which particular outcome will occur can be determined with certainty only after the completion of an experiment, then such experiment is called a *random experiment*.

In short, an experiment whose outcome is not predictable with certainty in advance is called a random experiment.

Examples:

a) Tossing of a fair coin,
b) Rolling a fair die,
c) Experiment of tossing two fair coins,
d) Rolling two fair dice,
e) Sex of newly born baby,
f) Recording a person's Blood group,

6.3 Sample Point and Sample Space

Sample Point

Outcome of a random experiment is known as *sample point*. In case of tossing a fair coin, the sample point may be either H or T.

Sample Space

The set of all possible outcomes of a random experiment is called a *sample space* of that random experiment. Usually sample space is denoted by S or U.

Examples:

a) If a coin is tossed once, then the sample space is
S = {H, T}, H : head, T : tail.

b) If a die is thrown once, then the sample space is
S = {1, 2, 3, 4, 5, 6}.

c) If two coins are tossed, then the sample space is
S = {HH, HT, TH, TT}.

d) If two dice are thrown, then the sample space is
S = {(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)}.

e) If we record sex of newly born baby, then the sample space is
S = {Male, Female}

f) If we record blood type of a person, then the sample space is
S = {O, A, B, AB}.

6.4 Event

Any subset of the sample space that together has some property we are interested in is called event. Generally, events are denoted by A, B, C,… or A_1, A_2, A_3,….

Examples:

(a) In tossing two coins, the sample space is S = {HH, HT, TH, TT}. Following are some events:

i. A = one head occurs, i.e., A = {HT, TH},
ii. B = two head occurs, i.e., B = {HH},
iii. C = no head occurs, i.e., C = {TT}.

(b) In throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}. Following are some events:

i. A = even number appear on the top, i.e., A = {2, 4, 6},
ii. B = a number divisible by 3 appear on the top, i.e., B = {3, 6},
iii. C = a number less than 4 appear on the top, i.e., C = {1, 2, 3, 4}.

6.5 TYPES OF EVENTS

6.5.1 Simple and Compound Event

Simple Event

An event consisting of single element of the sample space is called a simple event.

For example, in tossing two coins the sample space is S = {HH, HT, TH, TT}. Let A be the event that two head occurs and B be the event that only one head occurs and it is on the first coin. Then A = {HH} and B = {HT} both are simple events.

Compound Event

An event consisting of more than one element of the sample space is called a compound event.

For example, in tossing two coins the sample space is S = {HH, HT, TH, TT}. Let C be the event that at least one head occurs and B be the event that one head occurs. Then C = {HT, TH, HH} and D = {HT, TH} both are compound events.

6.5.2 Elementary Events

All singleton subsets of sample space of random experiment are called elementary events. That is, events that cannot be broken down into other events are called elementary events. Generally they are denoted by E_1, E_2, E_3, …. All elementary events are simple events.

Examples:

(a) In a random experiment of tossing a coin, the sample space is S = {H, T}. The events E_1 = {H} and E_2 = {T} are elementary events.

(b) In a random experiment of throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}. The events E_1 = {1}, E_2 = {2}, E_3 = {3}, E_4 = {4}, E_5 = {5} and E_6 ={6} are elementary events.

6.5.3 Equally Likely (equiprobable) Events

Two events are said to be equally likely if one of them cannot be expected to occur in preference to other. Clearly they have equal chance or likelihood of individual occurrence.

Examples:

(a) In tossing a coin, the sample space is S ={H, T}. Here likelihood (probability) of head is same as that of likelihood of tail. So the event A = {H} and B = {T} are equally likely events or equiprobable events.

(b) In throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}. The likelihood of simple events 1, 2, …, 6 are equally likely.

6.5.4 Impossible and Sure Event

Impossible Event

In a random experiment, the event that is logically impossible is called an impossible event or null event. Event corresponding to null set (empty) is an impossible event.

Examples:

(a) In a random experiment of throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}. Let A be the event corresponding to the number 8 appears on the top. Clearly the event A = { } =ϕ is an impossible event.

(b) In a random experiment of tossing two coins, the sample space is S = {HH, HT, TH, TT}. Let A be the event that at least three heads occurs. Clearly the event A = { } = ϕ is an impossible event.

Sure Event

In a random experiment, the event that is certain or sure to occur is called certain or sure event. Event corresponding to the sample space is called certain event.

Examples:

(a) In a random experiment of tossing a coin, the sample space is S = {H, T}. If A = event that either head or tail occurs, then clearly event A = {H, T} = S is sure event.

(b) In a random experiment of throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}. Let A be the event that a natural number less than 7 appears on the top. Clearly the event A = {1, 2, 3, 4, 5, 6} = S is sure event.

6.6 Algebra of Events

6.6.1 Complementary Event

Consider event A associated with a sample space S. The event corresponding to set of points in the sample space S, not belonging to A is called complementary event of A and is denoted by A' or A^c. In words, A^c means "not A". In set notation, it can be written as $A^c = \{x \mid x \in S; x \notin A\}$.

Examples:

(a) In tossing of two coins, the sample space is S = {HH, HT, TH, TT}. Let A be the event that two head occurs, i.e., A = {HH}. Then complementary event of A is A^c = {HT, TH, TT}.

(b) In throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}. Let B be the event that a number divisible by 3 appear on the top, i.e., B = {3, 6}. Then the complementary event of B is B^c = {1, 2, 4, 5}

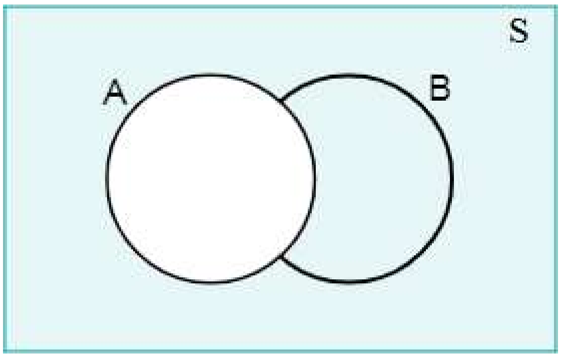


Figure 6.1: Complementary event

In the above Venn-diagram, the shaded region is A^c (Complementary event to A).

Python Code for Complementary event

```
# Create Universal Set S as follows
S = {1,2,3,4,5,6}
# Create Set A as follow
A = {2,5,6}
print("\nComplementary of A using Difference '-' operator")
S - A
```

The above code will give the following result:

```
Complementary of A using Difference '-' operator
{1, 3, 4}
```

6.6.2 Equality of Events

Let A and B be the events of sample space S. Then A and B are equal event if $A \subset B$ and $B \subset A$, i.e., two events are said to be equal if all the sample points of A belongs to B and vice versa.

Example:

In a random experiment of throwing a die, the sample space is S = {1, 2, 3, 4, 5, 6}. Let A be the event that even number appear on the top of the die and B be the event that a number divisible by 2 appear on the top of the die. Then A = {2, 4, 6} and B = {2, 4, 6}. As $A \subset B$ and $B \subset A$, event A and B are equal.

Python Code to check equality of events

```
A = {10, 12, 16, 18}
B = {12, 18, 10, 16}
A == B
print("Set A and B are equal")
C = {12,17,10,18}
A == C
print("Set A and C are not equal")
```

The above code will give the following results:

```
True
Set A and B are equal
False
Set A and C are not equal
```

6.6.3 Union of Events

Consider event A and B as two events associated with a sample space S. The union of two events A and B is denoted by $A \cup B$ and is defined as a set consisting of all those members which belongs to either A or B or both. In words, $A \cup B$ means “A or B”. In set notation, it can be written as $A \cup B = \{x \mid x \in A \text{ OR } x \in B\}$.

Examples:

(a) If event A is A = {1, 2, 3, 5, 6} and event B is B = {1, 2, 4, 6}. Then $A \cup B$ = {1, 2, 3, 4, 5, 6}.

(b) Let the sample space be S = {1, 2, 3, 4, 5, 6, 7, 8}. Let A = {1, 2, 3, 4, 5, 6} and B = {2, 3, 7, 8}. Then $A \cup B$ = {1, 2, 3, 4, 5, 6, 7, 8}.

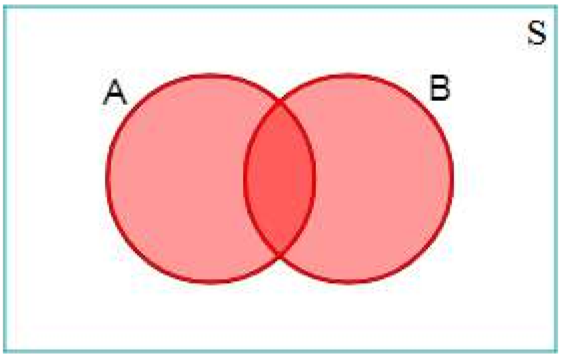


Figure 6.2: Shaded Region is the A ∪ B

In the above Venn-diagram, the shaded region is the union of A and B.

Python Code for Union of two events

```
# Create Set A as follows
A = {1,2,3,4,5,6}
# Create Set B as follow
B = {2,3,7,8}
print("\nUnion using '|' Operator")
A | B
print("\nUnion using 'union()' Function")
A.union(B)
```

The above code will give the following results:

```
Union using '|' Operator
{1, 2, 3, 4, 5, 6, 7, 8}
Union using 'union()' Function
{1, 2, 3, 4, 5, 6, 7, 8}
```

6.6.4 Intersection of Events

Consider event A and B as two events associated with a sample space S. The intersection of two events A and B is denoted by A ∩ B and is defined as a set consisting of all those members which belongs to both A and B. In words, A ∩ B means “A and B”. In set notation, it can be written as $A \cap B = \{x \mid x \in A \text{ and } x \in B\}$.

Examples:

(a) If event A is A = {1, 3, 5, 6} and event B is B = {1, 2, 4, 6}. Then A ∩ B = {1, 6}.

(b) Let the sample space be S = {1, 2, 3, 4, 5, 6, 7, 8}. Let A = {1, 2, 3, 4, 5, 6} and B = {2, 3, 7, 8}. Then A ∩ B = {2, 3}.

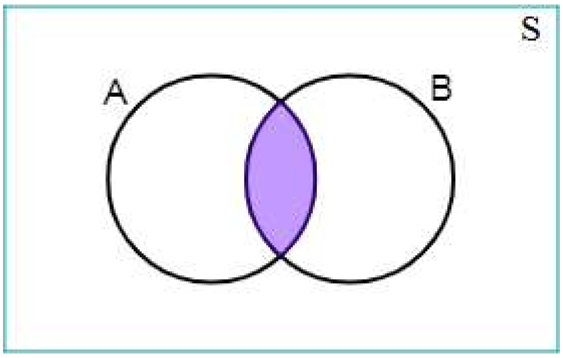


Figure 6.3: Intersection of Two events

In the above Venn-diagram, the shaded region is the intersection of A and B.

Python Code for Intersection of two events

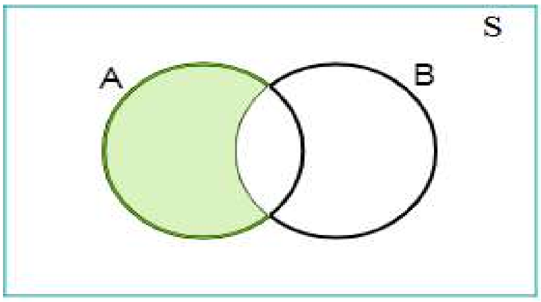
```
# Create Set A as follows
A = {1,2,3,4,5,6}
# Create Set B as follow
B = {2,3,7,8}
print("\nIntersection using '&' Operator")
A & B
print("\nIntersection using 'intersection()' Function")
A.intersection(B)
```

The above code will give the following results:

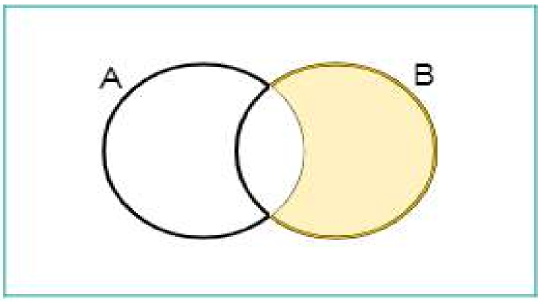
```
Intersection using '&' Operator
{2, 3}
Intersection using 'intersection()' Function
{2, 3}
```

6.6.5 Difference of Events

Let A and B be the events of the sample space S. Then A – B is known as difference of event A from event B. In difference A – B, the event A occur but B does not occur. i.e., $A – B = \{x \mid x \in A; x \notin B\}$. It can also be written as $A – B = A \cap B^c$. Similarly, B – A is known as difference event of B from A, i.e., $B – A = \{x \mid x \in B, x \notin A\}$. It can also be written as $B – A = B \cap A^c$.



$A – B = A \cap B^c$



$B – A = B \cap A^c$

Figure 6.4: Difference of Two Events

In the above Venn-diagram, the first figure shows A – B and second shows B – A.

Example:

(a) In a random experiment of tossing two coins, the sample space is S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}. Let A be the event that at least two head occurs, i.e., A = {HHH, HHT, HTH, THH} and B be the event that at least one tail occurs, i.e., B = {HHT, HTH, THH, HTT, THT, TTH, TTT}.
Then A – B = {HHH} = outcomes belongs to A and do not belong to B.

(b) Let the sample space be S = {1, 2, 3, 4, 5, 6, 7, 8}. Let A = {1, 2, 3, 4, 5, 6} and B = {2, 3, 7, 8}. Then A – B = {1, 4, 5, 6}.

Python Code for difference of two events

```
# Create Set A as follows
A = {1,2,3,4,5,6}
# Create Set B as follow
B = {2,3,7,8}
print("Difference using '-' operator\n")
A - B
print("Difference using 'difference()' function\n")
A.difference(B)
```

The above code will give the following result:

```
Difference using '-' operator
{1, 4, 5, 6}
Difference using 'difference()' function
{1, 4, 5, 6}
```

Mutually Exclusive Events

Events are said to be mutually exclusive if occurrence of one of them prevents the occurrence of any of the remaining ones. That is, there is no possibility of A and B both being true, then they are said to be mutually exclusive or disjoint. Symbolically, A and B are mutually exclusive if their intersection is null or empty set, i.e., $A \cap B = \phi$.

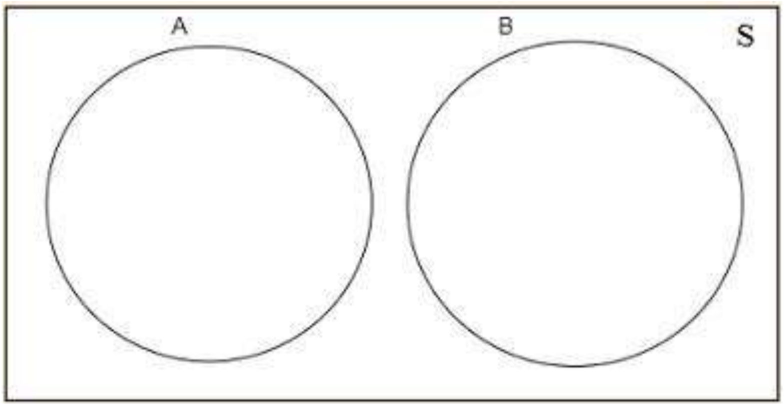


Figure 6.5: Mutually Exclusive Events

In the above Venn-diagram, the two events A and B are mutually exclusive.

Examples:

(a) In tossing of a coin experiment, the sample space is S = {H, T}. Let A = {H}, head occur and B = {T}, tail occur. Then A and B are mutually exclusive events ($A \cap B = \phi$).

(b) In a random experiment of throwing an unbiased die, the sample space is S = {1, 2, 3, 4, 5, 6}. Let A be the event that even number appear on the top, B be the event that odd number appear on the top and C be the event that a multiple of 3 appear on the top. Then A = {2, 4, 6}, B = {1, 3, 5} and C = {3, 6}. The events A and B are mutually exclusive because $A \cap B = \phi$. The events A and C are not mutually exclusive because $A \cap C = \{6\} \neq \phi$. The events B and C are not mutually exclusive because $B \cap C = \{3\} \neq \phi$.

Python Code to check mutually exclusive events

```
# Create Set A as follows
A = {2,4,6}
# Create Set B as follow
B = {1,3,5}
# Create Set C as follow
C = {3,6}
print("Intersection of A and B is ")
A.intersection(B)
print("A and B are mutually exclusive.")
print("Mutually exclusive using 'isdisjoint()' function")
A.isdisjoint(B)
print("Intersection of A and C is ")
A.intersection(C)
print("A and C are not mutually exclusive")
print("Mutually exclusive using 'isdisjoint()' function")
A.isdisjoint(C)
print("Intersection of B and C is ")
B.intersection(C)
print("B and C are not mutually exclusive")
print("Mutually exclusive using 'isdisjoint()' function")
B.isdisjoint(C)
```

The above code will give the following results:

```
Intersection of A and B is
set()
A and B are mutually exclusive.
Mutually exclusive using 'isdisjoint()' function
True
Intersection of A and C is
{6}
A and C are not mutually exclusive
Mutually exclusive using 'isdisjoint()' function
False
Intersection of B and C is
```

```
{3}
B and C are not mutually exclusive
Mutually exclusive using 'isdisjoint()' function
False
```

6.7 Exhaustive Events

If the union of two or more events is the sample space S then they are calledexhaustive events. Simple (elementary) events of random experiments always constitute an exhaustive event.

Examples:

(a) Consider an experiment of throwing die. The sample space is S = {1, 2, 3, 4, 5, 6}. Let A_1, A_2, …, A_6 be the events that 1, 2, … 6 appears on the top respectively. Clearly, at any throw at least one of these events will occur and $A_1 \cup A_2 \cup A_3 \cup A_4 \cup A_5 \cup A_6 = S$. Events A_1, A_2, …, A_6 constitute an exhaustive event.

(b) Consider an experiment of tossing a coin. The sample space is S = {H, T}. Let B_1 = {H} and B_2 = {T}. Clearly, at any toss at least one of these event will occur and $B_1 \cup B_2 = S$. Events A and B are exhaustive events.

6.8 Mutually Exclusive and Exhaustive Events

If $A \cap B = \phi$ and $A \cup B = S$, then A and B are called mutually exclusive and exhaustive events. All the elementary events are always mutually exclusive and exhaustive, so they form partition of the sample space.

Example:

(a) In a random experiment of tossing three coins, the sample space is S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}. Let A be the event that at least two head occurs, B be the event that one or less head occurs. Then A = {HHH, HHT, HTH, THH} and B = {HTT, THT, TTH, TTT}. As $A \cap B = \phi$ and $A \cup B = S$, the events A and B are mutually exclusive and exhaustive.

(b) In a random experiment of throwing an unbiased die, the sample space is S = {1, 2, 3, 4, 5, 6}. Let A be the event that even number appear on the top and B be the event that odd number appear on the top. Then A = {2, 4, 6} and B = {1, 3, 5}. The events A and B are mutually exclusive and exhaustive because $A \cap B = \phi$ and $A \cup B = S$.

6.9 Algebraic Properties of Events

Algebraic properties of events are the rules similar to rules of algebra on the events to perform unions, intersections and complementary of events.

Let A, B, C be the events of sample space S, then following laws holds:

- **Identity Law**
  a) $A \cup \phi = A$
  b) $A \cap \phi = \phi$
  c) $A \cup S = S$
  d) $A \cap S = A$
- **Idempotent Law**
  a) $A \cup A = A$

b) $A \cap A = A$

- **Complement Law**
  a) $A \cup A^c = S$
  b) $A \cap A^c = \phi$
  c) $(A^c)^c = A$
- **Commutative Law**
  a) $A \cup B = B \cup A$
  b) $A \cap B = B \cap A$
- **Associative Law**
  a) $A \cup (B \cup C) = (A \cup B) \cup C$
  b) $A \cap (B \cap C) = (A \cap B) \cap C$
- **Distributive Law**
  a) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
  b) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- **De Morgan's Law:**
  The very useful and important relationship between the basic operations of forming unions, intersections and complements of events is known as De Morgan's law. De Morgan's law states that the complement of union (intersection) of two events is the intersection (union) of their complements.
  a) $(A \cup B)^c = A^c \cap B^c$
  The complement of union of two events is the intersection of their complements.
  b) $(A \cap B)^c = A^c \cup B^c$
  The complement of intersection of two events is the union of their complements.

6.10 SUMMARY

The students will learn basic of random experiment and related terms like sample point, sample space, events, type of events, operations (complementary, union, intersection and difference) on events. The basic knowledge of random experiments and related terms is necessary to understand the concept and its use in probability theory. Students will able to perform the various operations on events using Python.

- Learn how to write the sample space related to random experiment.
- Learn how to define events from the sample space based on the characteristics of interest.
- Learn how to perform algebraic operations like complementary, union, intersection and difference.
- Learn how to perform these operations on events using Python programming.

6.11 REFERENCES

Books

1. Gupta, S.C. and Kapoor, V.K. (2014): Fundamentals of Mathematical Statistics, Sultan Chand & Sons, New Delhi, 12th Edition.
2. Hastie, Trevor, et al. (2009): The Elements of Statistical Learning, Springer

3. Ross, S.M. (2004): Introduction to Probability and Statistics for Engineers and Scientists, Academic Press
4. Navidi, W. (2011): Statistics for Engineers and Scientists, McGraw Hill, Third Edition.

6.12 QUESTIONS

Short Answer:

1. Define random experiment. Give some examples of random experiment.
2. Define Sample space and sample points with illustration.
3. What is the difference between sample space and event?
4. Define Simple and Compound events with illustration.
5. Define events and mutually exclusive events with illustration.
6. Define difference of events with illustration.
7. Define impossible event. Give some examples of impossible events.
8. Define sure event. Give some examples of sure events.
9. Define Union and intersection of events.
10. Define mutually exclusive events. Give an example of two events that are mutually exclusive and two events that are not mutually exclusive.
11. Classify the below experiments as random or non-random experiments:
    a. Two cards are drawn from pack of 52 cards and the suits (i.e., Club, Diamond, Heart, Spade) to which they belong are noted.
    b. A bowl contains 3 red and 6 blue balls. Two balls are drawn and their colors are noted.
    c. Water is heated in a bowl to a temperature of 100^{o}C for 5 minutes and we observe that water turns into steam.
    d. Inspecting an item from a production line as defective or non-defective.
    e.

Long Answer:

1. An experiment consists of tossing coins three times. Write the sample space. List the events that corresponds to
    a. At least one head
    b. At most one head
    c. More than one head
    d. Two or more heads
    e. More heads than tails.
2. An experiment consists of tossing coins three times. Write the sample space. Are the outcomes are equally likely?
3. An experiment consists of tossing a coin and a six faced die. What is the sample space of this experiment?
4. A box contains 4 Red, 4 Blue and 4 Yellow Marbles. Construct a sample space for the experiment of drawing two marbles in succession (with replacement).
5. Determine whether these events are mutually exclusive:
    a. Draw a card from pack of 52 cards. Let A be the event of getting spade card and B be the event of getting 6.

b. Draw a card from pack of 52 cards. Let A be the event of getting club card and B be the event of getting an ace.

c. Toss coin two times. Let A be the event that only one head occur and B be the event that only one tail occur.

d. Two dice are rolled. Let A be the event that the number appear on the first die is same as the number appear on the second die and B be the event that the sum of the number appear on the top face is 8.

6. Let A, B, C be three events. Write the expression for the following events:
   a. All the three event occur
   b. At least two of the events occur
   c. None of the event occurs
   d. Exactly two events occur.
7. Two dice are thrown. Write the sample space. Let A be the event that the sum of the numbers on the top of dice is 7, let B be the event that sum of the numbers on the top of dice is at most 7. Determine $A \cap B$, $A \cap B^c$, $A \cup B$, $(A \cup B)^c$.

PRACTICALS

1. Write a python code to define the sample space (S) for tossing two coins. Let A be the event that one head occurs and B be the event that at least one lead occurs. Determine $A \cup B$ and $A \cap B$.
2. Write a python code to define the sample space (S) for an experiment of tossing three coins. Let A be the event that at least two head occurs and B be the event that at least one tail occurs. Determine $A \cup B$ and $A \cap B$.
3. Let the sample space be S = {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. Let A = {2, 4, 6, 8, 10}, B = {1, 3, 5, 7, 9}, C = {3, 6, 9}, D = {5, 10}. Write python code to
   a. Determine $A \cup B$, $A \cap B$, $A \cap C$, $A \cap D$ and A – D.
   b. Find complement of A, complement of C.
   c. Check whether A and B are mutually exclusive and exhaustive.
   d. Check whether C and D are mutually exclusive.
   e. Check whether C and D are equal.

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

UNIT VII: Probability

STRUCTURE

7.0 OBJECTIVES

The main goal of this module is to teach you the basics concept of probability, different definitions of probability and various important theorems on probability. By studying this module you should be able to:

- understand how to calculate probability,
- use different theorems on probability to solve the numerical examples

Some examples are solved to understand the concept of probability and use of various theorems of probability. Few examples are done using python programming language.

7.1. INTRODUCTION

In the sense of mathematical logic, a statement is either true or false, and there is no third possibility in this case. Out of this frame, a third possibility does exist: What is the `extent' or `likelihood' of the statement being true? What is the extent or likelihood of the likelihood being false?, e.g., If a boy informs his father that he will either `pass' or `fail' in an examination, then obviously his father will not feel comfortable with this answer. He will be more satisfied by knowing the likelihood (percentage) of passing or failing in the examination. The numerical measurement of this `extent' or `likelihood' is called `Probability' of the happening or phenomenon under consideration..

In subsequent sections, you will learn about different definitions of probability and some important theorems on probability.

7.2. Classical Definition of Probability

If there are n mutually exclusive, equiprobable and exhaustive elements in a sample space S, and if r of them are favourable to occurrence of some event A, then the probability of event A is given by

$$P(A) = \frac{\text{Number of Favourable Cases to } A}{\text{Total Number of Cases}}$$
$$= \frac{r}{n}$$

This definition was given by Laplace.

The favourable cases to happening of A are always less than or equal to the total exhaustive cases, and also they cannot be negative.

Therefore, $0 \leq r \leq n$. Dividing by n, we have

$0 \leq \frac{r}{n} \leq 1$, i.e., $0 \leq P(A) \leq 1$.

If r cases are favourable to the happening of an event A, then $n - r$ cases are favourable to not happening of A.

Therefore,

$$P(A^c) = \frac{\text{Number of Favourable Cases to } A^c}{\text{Total Number of Cases}}$$
$$= \frac{n-r}{n}$$
$$= 1 - \frac{r}{n}$$
$$= 1 - P(A)$$

Thus, we have $P(A^c) = 1 - P(A)$.

7.3. Statistical or Empirical Definition of Probability

If the random experiment is repeated under essentially the same conditions for large number of time, then the limit of the ratio of number of times an event happens to the total number of trials is defined as the probability of that event. Here assume that the limit exists.

$$P(A) = \lim_{n \to \infty} \frac{r}{n}$$

7.4. Axiomatic Definition of Probability

Axiomatic approach to the probability was introduced by a Russian Mathematician Kolmogorov with the help of set theory.

Suppose S be the sample space of a random experiment. Let φ(S) be the power set of S and R be the set of all real numbers. Suppose that a set function P : φ(S) → R satisfies following axioms (postulates):

Axiom 1: For all $A \in \varphi(S), P(A) \geq 0.$

Axiom 2: $P(S) = 1.$

Axiom 3: For $A_1, A_2 \in \varphi(S), A_1 \cap A_2 = \phi, P(A_1 \cup A_2) = P(A_1) + P(A_2).$

Then $P\colon \varphi(S) \rightarrow R$ is called **probability function** and $P(A)$ is called the **probability of event** A.

- From axiom 1, probability of any event is a non-negative real number. (**Non-negativity**)
- From axiom 2, probability of certain event is 1. (**Certainty**)
- From axiom 3, probability function is additive. (**Additivity**)

Axiom 3, can be generalized as follows :

$A_1, A_2, \ldots, A_n$ are mutually exclusive events from $\varphi(S)$,

$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n).$

Properties of Probability

(a) **Non-negativity:** Probability of any event is always non-negative, i.e., $P(A) \geq 0$.

(b) **Certainty:** Probability of sure event or certain event is always 1, i.e., $P(S) = 1$.

(c) **Additivity:** If $A_1, A_2, \ldots, A_n$ are mutually exclusive events, then

$$P(A_1 \cup A_2 \cup \ldots \cup A_n) = P(A_1) + P(A_2) + \ldots + P(A_n)$$

7.5. Theorems on Probability

7.5.1 Probability of Impossible Event

The probability of impossible event is always zero, i.e., $P(\phi) = 0$.

Proof

Let ϕ be an impossible event of sample space S. Then, we have $\phi \cup S = S$ and $\phi \cap S = \phi$. Thus, ϕ and S are mutually exclusive and exhaustive events. Hence by Axiom 2 and 3,

$$P(\emptyset \cup S) = P(S)$$

$$P(\emptyset) + P(S) = P(S), \text{(Using Axiom 3)}$$

$$P(\emptyset) + 1 = 1 \text{(Using Axiom 2)}$$

$$\text{i.e., } P(\emptyset) = 0.$$

7.5.2 Probability of Complementary Event

If A and A^c are complementary events of the sample space S, then $P(A) + P(A^c) = 1$.

Proof

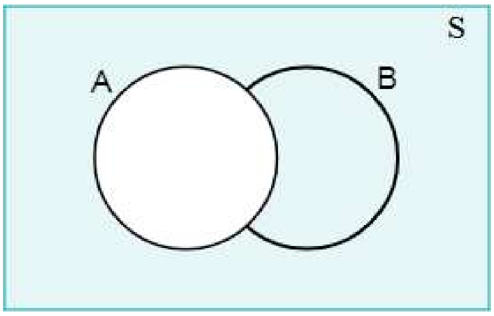

S
A
B

Figure 7.1: Complementary Event of *A*

Since *A* and A^c are complementary events of the sample space *S*, we have $A \cup A^c = S$ and $A \cap A^c = \varphi$. Hence by Axiom 2 and 3,

$$P(A \cup A^c) = P(S)$$

$$P(A) + P(A^c) = P(S), \text{(Using Axiom 3)}$$

$$P(A) + P(A^c) = 1 \text{(Using Axiom 2)}$$

$$\text{i.e., } P(A^c) = 1 - P(A).$$

Thus, we have $P(A)+ P(A^c) = 1$.

7.5.3 Addition Law of Probability

If *A* and *B* are two events of non-empty sample space *S*, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Proof

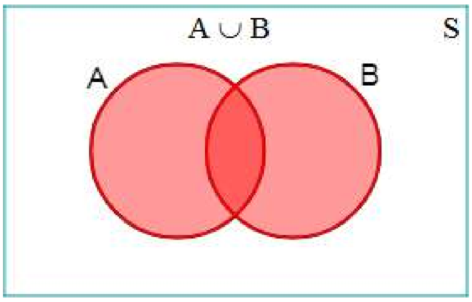


Figure 7.2: Union of *A* and *B*

The above Venn diagram shows $A \cup B$.

From the Venn diagram in Figure 7.3, it is clear that the event $A \cup B$ can be written as $A \cup B = A \cup (A^c \cap B)$ such that *A* and $A^c \cap B$ are mutually exclusive.

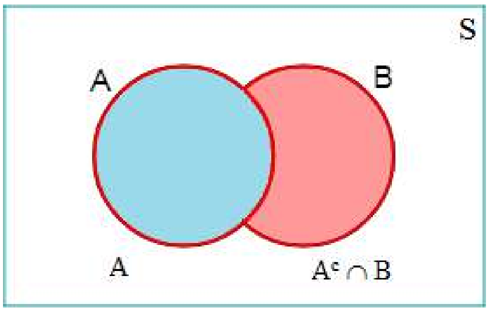


Figure 7.3: $A \cup B = A \cup (A^c \cap B)$

Using Axiom 3,

$$P(A \cup B) = P[A \cup (A^c \cap B)]$$

$$= P(A) + \ P(A^c \cap B) \qquad \dots \dots (1)$$

Similarly, from the Figure 7.4, the event *B* can be written as union of two mutually exclusive events $A \cap B$ and $A^c \cap B$ such that $B = (A \cap B) \cup (A^c \cap B)$.

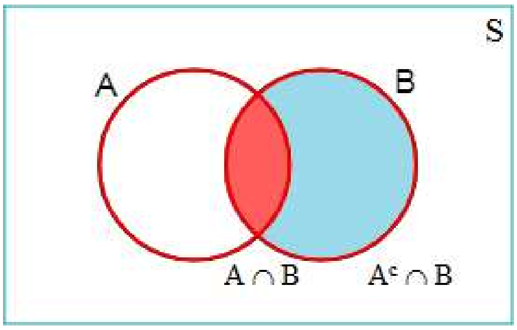


Figure 7.4: $B = (A \cap B) \cup (A^c \cap B)$

Using Axiom 3,

$$P(B) = P[(A \cap B) \cup (A^c \cap B)]$$

$$= P(A \cap B) + P(A^c \cap B)$$

$$\therefore P(A^c \cap B) = P(B) - P(A \cap B)$$

Substituting for $P(A^c \cap B)$ in equation 1, we get

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

7.5.4 Addition Law of Probability for Mutually Exclusive Events

If *A* and *B* are two mutually exclusive events of non-empty sample space *S*, then

$$P(A \cup B) = P(A) + P(B).$$

Proof

Since *A* and *B* are mutually exclusive events, $A \cap B = \phi$.

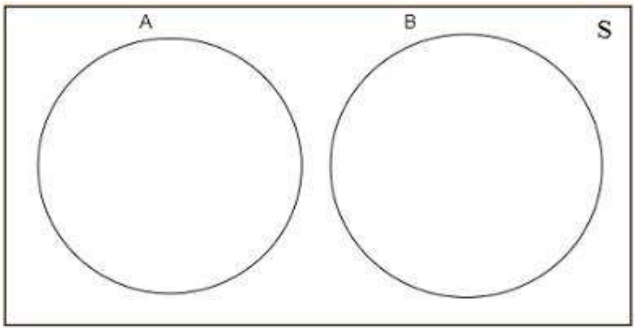


Figure 7.5: Mutually Exclusive Events

Therefore, $P(A \cap B) = P(\phi) = 0$. Hence,

$$P(A \cup B) = P(A) + P(B)$$

7.5.5 Law of Addition for Three Events

For three events *A*, *B* and *C* of non-empty sample space *S*,

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C)$$
$$-P(B \cap C) + P(A \cap B \cap C)$$

7.5.6 Some Important Results

Suppose *A* and *B* are the events of the non-empty sample space S.

(a) If $A \subset B$, then $P(A) \leq P(B)$.

(b) $P(A \cap B) \leq P(A)$ or $P(A \cap B) \leq P(B)$.

(c) $P(A \cup B) \geq P(A)$ or $P(A \cup B) \geq P(B)$.

(d) If $A \subseteq B$, then $P(A \cap B) = P(A)$.

(e) If $A \subseteq B$, then $P(A \cup B) = P(B)$.

Example 1:

Consider a random experiment of throwing a fair die. List all the elements of the sample space. Find the probability that

(i) Even number appear on the top

(ii) Odd number appear on the top

(iii) A number greater than or equal to 3 appear on the top.

Solution: The sample space S of the random experiment is S = {1, 2, 3, 4, 5, 6}.

(i) Let A be the event that even number appear on the top. Then A = {2, 4, 6}.

$$P(A) = \frac{\text{Number of favourable cases to A}}{\text{Total number of cases}} = \frac{3}{6} = \frac{1}{2}.$$

(ii) Let B be the event that odd number appear on the top. Then B = {1, 3, 5}.

$$P(B) = \frac{\text{Number of favourable cases to B}}{\text{Total number of cases}} = \frac{3}{6} = \frac{1}{2}.$$

(iii) Let C be the event that a number greater than or equal to 3 appear on the top. Then C = {3, 4, 5, 6}.

$$P(C) = \frac{\text{Number of favourable cases to C}}{\text{Total number of cases}} = \frac{4}{6} = \frac{2}{3}.$$

Python Code to calculate probability

```
# Import a Fraction function from
# fractions library
from fractions import Fraction
# define a user defined function Prob
def Prob(event, space):
    "The probability of an event, given a sample space of equiprobable outcomes."
    return Fraction(len(event & space),
                    len(space))
```

```
## Sample space when a die is thrown
S = {1,2,3,4,5,6}
## Event A is odd number appear on the top
A = {1,3,5}
## Event B is even number appear on the top
```

```
B = {2,4,6}
## Event C is a number greater than or equal to 3
## appear on the top
C = {3,4,5,6}

print("Probability of A is ", Prob(A,S))
print("Probability of B is ", Prob(B,S))
print("Probability of C is ", Prob(C,S))
```

The above code will give the following results:

```
Probability of A is 1/2
Probability of B is 1/2
Probability of C is 2/3
```

Example 2:

Consider a random experiment of tossing three coins. List all the elements of the sample space. Find the probability of getting

(i) at least one head,
(ii) at the most two heads,
(iii) no head.

Solution: Let three coins be tossed. The sample space S consists of 8 elementary events

S = {HHH, HHT, HTH, THH, HTT, THT, TTH, TTT}.

(i) Let A denote the event that at least one head occurs. Then A = {HHH, HHT, HTH, THH, HTT,THT, TTH}.

$$P(A) = \frac{\text{Number of favourable cases to A}}{\text{Total number of cases}} = \frac{7}{8}$$

(ii) Let B denote the event that at the most two head occurs. Then B = {HHT, HTH, THH, HTT,THT,TTH, TTT}

$$P(B) = \frac{\text{Number of favourable cases to B}}{\text{Total number of cases}} = \frac{7}{8}$$

(iii) Let C denote the event that no head occur. Then C ={TTT}

$$P(C) = \frac{\text{Number of favourable cases to C}}{\text{Total number of cases}} = \frac{1}{8}$$

Python Code to calculate probability

```
## import product function from itertools
from itertools import product
## generate sample space for tossing 3 coins
S = set(product(["H","T"],repeat = 3))
S

## Event A at least 1 H
A = {R for R in S if R.count("H") >=1}
A
## Event B at most 2 H
B = {R for R in S if R.count("H")<=2}
B
## Event C no H
C = {R for R in S if R.count("H")==0}
```

```
C
# probability of A
Prob(A,S)
print("Probability of A is ",Prob(A,S))
# probability of B
Prob(B,S)
print("Probability of B is ",Prob(B,S))
# probability of C
Prob(C,S)
print("Probability of C is ",Prob(C,S))
```

The above code will give the following results:

```
Probability of A is 7/8
Probability of B is 7/8
Probability of C is 1/8
```

Example 3:

In a single throw with two uniform dice, what is the probability of getting (a) a total of 9, (b) total different from 9, (c) total is greater than or equal to 8, (d) a total of 7 or 11, (e) maximum of two numbers is greater than 4.

Solution:

Let two uniform dice be thrown. The sample space consists of 36 elementary events.

S = {(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)}.

(a) Let A be the event that total of 9 occur. Therefore A = {(3, 6), (4, 5), (5, 4), (6,3)}. So, no. of elements in A is r = 4. Hence, the required probability is

$$P(A) = \frac{\text{Number of favourable cases to A}}{\text{Total number of cases}} = \frac{4}{36} = \frac{1}{9}$$

(b) The event of getting the total different from 9 is the complementary event of A. So, the required probability is

$$P(A^c) = 1 - P(A) = 1 - \frac{1}{9} = \frac{8}{9}$$

(c) Let B be the event that total of greater than or equal to 8 occur. Therefore,

B = { (2,6),
(3,5), (3,6),
(4,4), (4,5), (4,6),
(5,3), (5,4), (5,5), (5,6),
(6,2), (6,3), (6,4), (6,5), (6,6)}.

Number of favorable cases for B = 15. Therefore

$$P(B) = \frac{\text{Number of favourable cases to B}}{\text{Total number of cases}} = \frac{15}{36}$$

(d) Let C be the event that total of 7 occur and D be the event that the total of 11 occur. C ={(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)} and D = {(5,6), (6,5)}.
The required probability is C ∪ D. Moreover C and D are mutually exclusive.

$$P(C \cup D) = P(C) + P(D) = \frac{6}{36} + \frac{2}{36} = \frac{8}{36} = \frac{2}{9}$$

(e) Let E = {(x,y) | max (x,y) >4}. Then

$$E = \{ \begin{array}{r} (1,5), (1,6) \\ (2,5), (2,6), \\ (3,5), (3,6), \\ (4,5), (4,6), \\ (5,1),(5,2), (5,3), (5,4), (5,5), (5,6), \\ (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)\}. \end{array}$$

No. of cases favourable to E = 20. Therefore, the probability of E is

$$P(E) = \frac{20}{36} = \frac{5}{9}$$

Python Code to calculate probability

```
# Import a Fraction function from
# fractions library
from fractions import Fraction
# define a function Prob
def Prob(event, space):
   "The probability of an event, given a sample space of equiprobable outcomes."
   return Fraction(len(event & space),
                   len(space))
## import itertools from product library
from itertools import product
S = set(product([1,2,3,4,5,6],repeat = 2))
print("The sample space is S = ",S)
print("Total cases in S = ",len(S))
```

The above code will give the following results:

```
The sample space is S =  {(1, 3), (6, 6), (5, 6), (2, 1), (6, 2), (1, 6), (5, 1), (2, 5), (1, 2), (3, 3), (5, 5), (4, 4), (6, 3), (1, 5), (3, 6), (2, 2), (4, 1), (1, 1), (6, 4), (3, 2), (2, 6), (5, 4), (4, 5), (5, 2), (1, 4), (2, 3), (4, 2), (6, 5), (3, 5), (5, 3), (4, 6), (6, 1), (3, 1), (4, 3), (3, 4), (2, 4)}
Total cases in S =  36
```

```
## Event A total of 9 occur
A = {R for R in S if (R[0]+R[1])==9}
print(A)
print("Favourable cases to A",len(A))
# probability of A
Prob(A,S)
print("Probability of A is ",Prob(A,S))
```

The above code will give the following results:

```
{(6, 3), (5, 4), (4, 5), (3, 6)}
Favourable cases to A 4
Probability of A is 1/9
```

```
B= {R for R in S if R[0]+R[1]>=8}
```

```
print(B)
print("Favourable cases to B = ",len(B))
# probability of B
Prob(B,S)
print("Probability of B is ",Prob(B,S))
```

The above code will give the following results:

```
{(6, 4), (5, 4), (2, 6), (5, 5), (4, 6), (6, 6), (5, 6), (4, 5), (4, 4), (6, 3), (6, 2), (3, 6), (5,
3), (6, 5), (3, 5)}
Favourable cases to B = 15
Probability of B is 5/12
```

```
C = {R for R in S if R[0]+R[1]==7}
print(C)
print("Favourable cases to C",len(C))
D = {R for R in S if R[0]+R[1]==11}
print(D)
print("Favourable cases to D",len(D))
# probability of C union D
Prob(C | D,S)
print("Probability of C union D is ",Prob(C | D,S))
```

The above code will give the following results:

```
{(6, 1), (1, 6), (4, 3), (3, 4), (2, 5), (5, 2)}
Favourable cases to C 6
{(5, 6), (6, 5)}
Favourable cases to D 2
Probability of C union D is  2/9
```

```
E= {R for R in S if max(R[0],R[1])>4}
print(E)
print("Favourable cases to E",len(E))
# probability of E
Prob(E,S)
print("Probability of E is ",Prob(E,S))
```

The above code will give the following results:

```
{(6, 6), (5, 6), (6, 2), (1, 6), (5, 1), (2, 5), (5, 5), (6, 3), (1, 5), (3, 6), (6, 4), (5, 4), (2,
6), (4, 5), (6, 5), (5, 3), (3, 5), (4, 6), (6, 1), (5, 2)}
Favourable cases to E 20
Probability of E is 5/9
```

Example 4:

In a lot of 100 electric bulb 10% of them are defective. Five bulbs are selected at random.

a. What is the probability of no defective bulb among the 5 bulbs?
b. What is the probability of 2 bulbs being defective among the 5 bulbs?

Solution:

Total number of bulbs = 100. Out of 100 electric bulb, 10% bulbs are defective, i.e., 10 bulb are defective and remaining 90 are non-defective.

Out of 100 bulbs 5 bulbs can selected in $\binom{100}{5}$ ways. i.e., n = Total number of elements $= \binom{100}{5}$.

And out of 90 non-defective bulbs 5 can be selected in $\binom{90}{5}$ ways. i.e., m = $\binom{90}{5}$.

a. Let A be the event that no defective bulb among selected 5.

$$P(A) = \frac{m}{n} = \frac{\binom{90}{5}}{\binom{100}{5}}$$

b. Let B be the event that two defective bulb among selected 5.

Out of 10 defective bulbs 2 bulbs can be selected in $\binom{10}{2}$ ways and remaining 3 bulb from 90 non-defective can be selected in $\binom{90}{3}$ ways. The events are compound events.

$$P(B) = \frac{m}{n} = \frac{\binom{10}{2}\binom{90}{3}}{\binom{100}{5}}.$$

<u>Example 5:</u>

Two cards are drawn at random simultaneously from a pack of playing cards. Find the probability that

a. both the cards are spade cards,
b. both the cards are of same suit.

Solution:

As two cards are drawn from 52 cards, the sample space consists of $\binom{52}{2}$ elements.

a. Let A be the event that both the cards are space cards. Out of 13 cards of same suit, 2 cards can be selected in $\binom{13}{2}$ ways.

$$P(A) = \frac{\binom{13}{2}}{\binom{52}{2}} = \frac{13 \times 12}{52 \times 51} = \frac{1}{17}.$$

b. Let B be the event that both the cards are of same suit. Two cards of any suit can be selected in $\binom{13}{2}$ ways, but there are 4 suits. So the number of elements favourable to B $= 4 \times \binom{13}{2}$

$$P(B) = \frac{4 \times \binom{13}{2}}{\binom{52}{2}} = \frac{4 \times 13 \times 12}{52 \times 51} = \frac{4}{17}.$$

7.6. <u>Conditional Probability</u>

In a simple language the conditional probability is "What is the chance that something will happen, given that something else has already happened?".

Let *A* and *B* be two events of non-empty sample space *S*.

The probability of some event *A* when it is known that event *B* has already occurred is called conditional probability of *A* given *B*. It is denoted by *P(A|B)* and is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \ \ P(B) > 0.$$

P(A|B) can be read as *probability that A occurs given that B occurs*. The conditional probability of *A* given *B* is the joint probability of *A* and *B* divided by the marginal probability of *B*.

The probability of some event *B* when it is known that event *A* has already occurred is called conditional probability of *B* given *A*. It is denoted by *P(B|A)* and is defined by

$$P(B|A) = \frac{P(B \cap A)}{P(A)}, \ \ P(A) > 0.$$

P(B|A) can be read as *probability that B occurs given that A occurs*. The conditional probability of *B* given *A* is the joint probability of *B* and *A* divided by the marginal probability of *A*.

Result 1:

Let *A* and *B* be two events of non-empty sample space *S* such that $P(A) > 0, P(B) > 0$. Then

$$0 \le P(A|B) \le 1.$$

Result 2:

Let *A* and *B* be two events of non-empty sample space *S* such that $P(A) > 0, P(B) > 0$. Then

$$P(A^c|B) = 1 - P(A|B).$$

7.7. <u>Multiplication Law of Probability</u>

Let *A* and *B* be two events of non-empty sample space *S* such that $P(A) > 0, P(B) > 0$, then

$$P(A \cap B) = P(A|B) \cdot P(B)$$

and

$$P(B \cap A) = P(B|A) \cdot P(A).$$

That is,

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A).$$

Multiplication law of probability is also known as ***Law of Compound Probability***.

Proof:

Suppose there are *n* elements in the sample space and n_1, n_2, n_3 elements in the events $A \cap B^c$, $A \cap B$, $A^c \cap B$ respectively.

Then,

$$P(A) = \frac{n_1 + n_2}{n}, P(B) = \frac{n_2 + n_3}{n} \text{ and } P(A \cap B) = \frac{n_2}{n}.$$

To find the *P(A|B)*, we keep in view *B* as a sample space. In *B* there are $n_1 + n_3$ elements out of which n_2 are in favour of *A*.

$$P(A|B) = \frac{n_2}{n_2 + n_3} \text{ and } P(A \cap B) = \frac{n_2}{n}.$$

Therefore,

$$P(A|B) \cdot P(B) = \frac{n_2}{n_2 + n_3} \cdot \frac{n_2 + n_3}{n} = \frac{n_2}{n} = P(A \cap B).$$

That is,

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Similarly, one can prove

$$P(B \cap A) = P(B|A) \cdot P(A).$$

Result 1:

For three events *A*, *B* and *C* of the sample space *S*,

$$\begin{aligned} P(A \cap B \cap C) &= P(A) \cdot P(B|A) \cdot P(C|(A \cap B)) \\ &= P(A) \cdot P(C|A) \cdot P(B|(A \cap C)) \\ &= P(B) \cdot P(A|B) \cdot P(C|(A \cap B)) \\ &= P(B) \cdot P(C|B) \cdot P(A|(B \cap C)) \\ &= P(C) \cdot P(A|C) \cdot P(B|(A \cap C)) \\ &= P(C) \cdot P(B|C) \cdot P(A|(B \cap C)) \end{aligned}$$

Result 2:

For three events *A*, *B* and *C* of the sample space *S*,

$$P[(A \cup B)|C] = P(A|C) + P(B|C) - P[(A \cap B)|C]$$

Example:

A random sample of 200 students is classified below by sex and experiencing hypertension during examination period. If a student is selected at random from this sample, find the probability that the student is

(a) male given that the student is experiencing hypertension,
(b) experiencing hypertension given that the student is female,
(c) female given that the student is experiencing no hypertension.

Status \Gender	Male	Female	Total
Hypertension	60	45	105
No Hypertension	40	55	95
Total	100	100	200

Solution:

Let *M* be the event that male student is chosen, *F* be the event that female students is chosen, *H* be the event that the one chosen experiences hypertension and *N* be the event that the one experiences no hypertension.

(a) Probability that the student is male given that he is experiencing hypertension is

$$P(M|H) = \frac{P(M \cap H)}{P(H)}$$
$$= \frac{60/200}{105/200} = \frac{60}{105}.$$

(b) Probability that the student is experiencing hypertension given that the student is female

$$P(H|F) = \frac{P(H \cap F)}{P(F)}$$
$$= \frac{45/200}{100/200} = \frac{45}{100}.$$

(c) Probability that the student is female given that she is experiencing no hypertension is

$$P(F|N) = \frac{P(F \cap N)}{P(N)}$$
$$= \frac{55/200}{95/200} = \frac{55}{95}.$$

7.8. Independent Events

Events are said to be independent if occurrence or non-occurrence of any one of them does not depend on that of any of the remaining ones.
In particular, an event *A* is said to be independent of another event *B* if *P(A|B) = P(A)*.
This definition is meaningful only if *P(A|B)* is defined, i.e., *P(B) > 0.*

7.9. Multiplication Law of Probability For Independent Events

If A and B are events such that *P(A) > 0*, *P(B) > 0*, then *A* and *B* are independent if and only if

$$P(A \cap B) = P(A) \cdot P(B).$$

i.e., for independent events *A* and *B*, the probability that both of these occur simultaneously is the product of their respective probabilities.

Note:

Three events *A*, *B* and *C* of the sample space *S*, are **mutually independent** if

a. $P(A \cap B) = P(A)P(B)$
b. $P(B \cap C) = P(B)P(C)$
c. $P(A \cap C) = P(A)P(C)$
d. $P(A \cap B \cap C) = P(A)P(B)P(C)$

If only first three conditions are satisfied, then *A*, *B* and *C* are **pair-wise independent**.

Result 1:

If A and B are independent events of the sample space S, then

a. A and B^c are independent
b. A^c and B are independent
c. A^c and B^c are independent

Example:

Jaydev and Vijay can solve respectively 60% and 80% problems in a book. They try independently to solve a problem randomly selected from the book. Find the probability that (i) problem is solved, (ii) only Jaydev can solve the problem, (iii) only Vijay can solve the problem, (iv) none of them can solve the problem.

Solution:

Let A be the event that Jaydev can solve the problem, and B be the event that Vijay can solve the problem. Hence $P(A) = 0.60$ and $P(B) = 0.80$.
As Jaydev and Vijay solve the problem independently, we have

$$P(A \cap B) = P(A) \times P(B) = 0.60 \times 0.80 = 0.48.$$

(i) The problem is solved if both Jaydev and Vijay solve the problem. Thus the required probability is

$$\begin{aligned} P(\text{Problem is solved}) &= P(A \cup B) \\ &= 1 - P(A^c \cap B^c) \\ &= 1 - P(A^c)P(B^c) \\ &= 1 - (0.40)(0.20) \\ &= 1 - 0.08 = 0.92 \end{aligned}$$

(ii) The probability that only Jaydev can solve the problem is $P(A \cap B^c)$.

$$\begin{aligned} P(A \cap B^c) &= P(A \cap B^c) \\ &= P(A)P(B^c) \\ &= (0.60)(0.20) \\ &= 0.12 \end{aligned}$$

(iii) The probability that only Vijay can solve the problem is $P(A^c \cap B)$.

$$\begin{aligned} P(A^c \cap B) &= P(A^c \cap B) \\ &= P(A^c)P(B) \\ &= (0.40)(0.80) \\ &= 0.32 \end{aligned}$$

(iv) The probability that none can solve the problem is $P(A^c \cap B^c)$.

$$\begin{aligned} P(A^c \cap B^c) &= P(A^c \cap B^c) \\ &= P(A^c)P(B^c) \\ &= (0.40)(0.20) \\ &= 0.08 \end{aligned}$$

7.10. SUMMARY

In this unit we have introduced the concept of probability and various definitions of probability, how to compute the probabilities of events using different laws of probability. We have discussed the concept of conditional probability, multiplication law of probability and independence. To understand the theoretical concepts we have cover some numerical examples manually as well as using Python programming language.

Some of the key definitions and properties are introduced in this unit.

- Classical, statistical and axiomatic definitions or probability.
- Various Theorems on probability.
- Some important results on probability
- Examples on probability
- Python code to solve the problems on probability

7.11. REFERENCES

Books

1. Gupta, S.C. and Kapoor, V.K. (2014): Fundamentals of Mathematical Statistics, Sultan Chand & Sons, New Delhi, 12th Edition.
2. Hastie, Trevor, et al. (2009): The Elements of Statistical Learning, Springer
3. Ross, S.M. (2004): Introduction to Probability and Statistics for Engineers and Scientists, Academic Press
4. Navidi, W. (2011): Statistics for Engineers and Scientists, McGraw Hill, Third Edition.

QUESTIONS

Short Answer:

1. Define
    a. Axiomatic Definition of Probability
    b. Statistical or Empirical Definition of Probability
    c. Classical Definition of Probability
2. Given the classical definition of probability. State its limitation.
3. State the addition law of probability for two events.
4. State the addition law of probability for three events.
5. What is the probability of impossible event?
6. What is the probability of certain event?
7. Determine the type of events in each case
    a. $P(A \cap B) = 0$
    b. $P(A \cup B) = P(A) + P(B) =1$
    c. $P(A \cup B) = P(A) + P(B)$.
8. Define conditional probability.
9. State the multiplication law of probability.
10. State the condition for independence of two events.
11. A^c and B^c are independent events. What can you say about the events A and B?
12. State the conditions for pair-wise independence of three events.
13. State the conditions for mutually independence of three events.

Long Answer:

1. Prove the addition law of probability.
2. With the usual notations, prove that
    a. $P(\phi) = 0$
    b. $P(A^c) = 1 – P(A)$
    c. $P(A \cup B) = P(A) + P(B) – P(A \cap B)$
    d. For two mutually exclusive events A and B, $P(A \cup B) = P(A) + P(B)$

   e. Two or more heads
   f. More heads than tails.
3. A die is thrown once. What is the probability that
   a. a number 5 appear on the top,
   b. an even number appear on the top,
   c. an odd number appear on the top,
   d. a number less than 0 appear on the top,
   e. a number greater than or equal to 0 appear on the top.
4. Prove multiplication law of probability.
5. *A* and *B* are independent events. Prove that A^c and B^c are independent.
6. A^c and B^c are independent events. Prove that *A* and *B* are independent.
7. Two cards are drawn without replacement from a pack of 52 cards. What is the probability that (i) Both are drawn are red, (ii) First is king and second is queen, (iii) One is red and other is black?
8. A charted accountant applies for a job in two firms X and Y. He estimates that the probability of his being selected in firm X is 0.7, and being rejected at Y is 0.5, and the probability of at least one of his application being rejected is 0.6. What is the probability that he will be selected in one of the firms?
9. A die is thrown twice. Let *A, B, C* denote the following events: $A = \{(a, b) \mid a \text{ is odd}\}$, $B = \{(a, b) \mid b \text{ is odd}\}$ and $C = \{(a, b) \mid a + b \text{ is odd}\}$. Check whether *A, B* and *C* are independent or independent in pairs only.
10. Two dice are thrown. Write the sample space. Let *A* be the event that the sum of the numbers on the top of dice is 7, let *B* be the event that sum of the numbers on the top of dice is at most 7. Determine the probability of $A \cap B$, $A \cap B^c$, $A \cup B$, $(A \cup B)^c$.
11. $P(A) = 2P(B) = 3\ P(B|A) = 0.6$. Find the probability that (i) *B* does not occur, (ii) exactly one of *A* and *B* occurs, (iii) not more than one of *A* and *B* occurs, (iv) *A* and *B* do not occur together and (v) neither of *A* and *B* occurs.

PRACTICALS

1. Write a python code to define solve the example 9 from the above Long Answer questions.
2. Write a python code to define the sample space (*S*) for an experiment of tossing three coins. Let *A* be the event that at least two head occurs and *B* be the event thatat least one tail occurs. Determine $P(A)$, $P(B)$, $P(A \cup B)$ and $P(A \cap B)$.
3. Write a python code to define solve the example 10 from the above Long Answer questions.
4. Let the sample space be $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Let $A = \{2, 4, 6, 8, 10\}$, $B = \{1, 3, 5, 7, 9\}$, $C = \{3, 6, 9\}$, $D = \{5, 10\}$. Write python code to
   a. Determine $P(A \cup B)$, $P(A \cap B)$, $P(A \cap C)$, $P(A \cap D)$ and $P(A – D)$.
   b. Find $P(A^c)$, $P(C^c)$.

M.Sc. (Computer Science)

Probability & Statistical Analysis

Semester 1

UNIT VIII: STATISTICAL INFERENCE

STRUCTURE

8.0 OBJECTIVES

a. To elaborate Statistical Inference
b. To discuss Random variables
c. To discuss and Probability and Density Functions
d. To explore Moments in Python

8.6 INTRODUCTION TO STATISTICAL INTERFERENCE

In python, statistical inference is used to draw and infer conclusions from set of values in given dataset. Initially, some random samples are used and extracted from given population which is used to describe and draw relevant inferences (conclusions) for entire population (Vondrejc, 2019). Thereare numerous Inferential Statistics available which can be used along with Python and are named as:

a. Z Test and Z Scores
b. t-Test
c. F-Tests
d. Correlation Coefficients
e. Chi square

8.1.1 Need of Preliminary Libraries

Distinct preliminary set of libraries is required to get imported in Python when one wants to work with Arrays, Data frames and different tools for performing statistical analysis. Numpy and Pandas come under this category. These both are highly important and different packages where Numpy canbe used to perform distinct operations on Arrays whereas another platform Pandas is used to carry different operations on Data frames. These two libraries can be imported as:

```
import numpy as np

import pandas as pd
```

8.1.2 Z Scores and Z-test

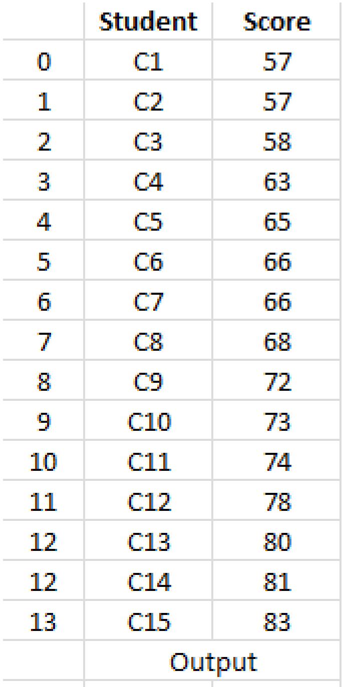
- **Z Scores**

Z Scores computes the probability of score which are calculated from normal distribution. This helps in comparing scores for two or more given normal distributions

- **Z Value**
- **Importing Dataset**

Here, dataset which contains exam scores for some of the students has been imported:

```
Z_scoresdata=pd. Read_excel(C:/Users/user/Desktop/Datasets/marks_Score.xls")

Z_scoresdata
```



	Student	Score
0	C1	57
1	C2	57
2	C3	58
3	C4	63
4	C5	65
5	C6	66
6	C7	66
7	C8	68
8	C9	72
9	C10	73
10	C11	74
11	C12	78
12	C13	80
12	C14	81
13	C15	83

Output

- **Z Score Calculation**

Following is the code by which Z scores can be calculated. Here, Z Scores have been calculated for a column with name „Score‟ column of Z Score dataset. The function „ddof‟ can be used to alter the divisor sum of squares of mean sample. Initially, ddof holds 0 but for std we can use ddof=1

```
Z_scoredata[„scores_zScore‟]=(z_scoresdata[„score‟]-z_scoredata[„score‟].mean())/z
_scoredata[„Score‟].std(ddof=1)
```

- **Z Score calculation in Python—**

```
import numpy as np

from scipy import stats

Arayr1= [[19,3,6,2,35],

[49,11,12,35,5]]

Array2 = [[49,11,11,34,5],

[12,10,10,34,22]]

Print (“array1:”, array1)

Print (“\narray2:”, array2)

Print (“z score for array1:, stats,zscore(array1))
```

```
Print(“\nz score for array1:”, stats.zscore(array1,axis=1))
```

- **Calculating Percentage**

For above considered example, percentage of people scored more than 70 can be calculated. Mean and Standard Deviation is to be taken for calculating area under curve. The code that can be used to find area under curve can be depicted as:

```
Cutoff= 70

Print(1-(scipy.stats.norm(70.5, 7.06).cdf(70)))
```

Where 7.06 is standard deviation and 70.5 is mean for finding percentage.

- **Z Test**

This test determines whether the given two datasets are similar or not.

- **Importing dataset**

  By taking initial dataset from population and some random samples of this dataset, importing dataset is initial set which can be done in following way:

```
HghtDataPop=pd.read_csv(“C:/User/ABC/Datasets/Heightof100ppl.csv”)
```

  Now, by taking random sample of above mentioned dataset:

```
HghtdataSample= pd.read_excel(“C:/User/ABC/Datasets/HeightdataSample.xls”)
```

- **Package to be imported for applying Z Test-**

```
from statesmodel.stats.weightstats import ztest
```

- **How to run Z Test**

```
Ztest(A2,b2=None,value=mean1)
```

- **Code for implementing Z test in Python**

```
    # considering an array of 55 numbers with mean 105 and std dev 16
import math
import numpy as np
from numpy.random import randn
from statsmodels.stats.weightstats import ztest
    mean = 105
    std_dev = 16/math.sqrt(55)
    alpha =0.01
    nul_mean =101
    data = sd_iq*random(55)+mean
    # printing mean and std_dev
    print('mean=%f std_dev=%f' % (np1.mean(data), np1.std(data)))
    # Z test
    ztest_Score, p1_value= ztest(data1,value = null_mean, alternative='larger')
      if(p1_value < alpha):
    print("Rejecting Null Hypothesis")
    else:
print("Failed to Reject Null Hypothesis")
```

8.1.3 t Test

t-Test is used to evaluate similarity level of groups. This can also be done by using Z test, the difference is Z Test is better to apply for the case sample size greater than 30 whereas t-Test is used for sample size less than 30.

- **Importing package for applying t-Test**

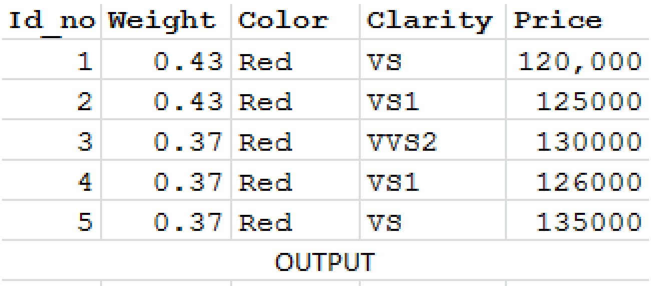
Import scipy.stats as stats

- **Importing Dataset**

For instance, considering hypothetical *Rubyjewellery* dataset which contains all necessary information of Rubyjewellery, which is sold in a store of jewellery.

Rubyjewellery-pd.read_excel("C:/Users/Login/Rubyjewellery.xls")

Rubyjewellery



Id_no	Weight	Color	Clarity	Price
1	0.43	Red	VS	120,000
2	0.43	Red	VS1	125000
3	0.37	Red	VVS2	130000
4	0.37	Red	VS1	126000
5	0.37	Red	VS	135000

OUTPUT

- **Code for t-test in Python**

```
from numpy.random import randn
from scipy.stats import ttest_ind
seed(1)
data_1 = 4 * randn(105) + 51
data_2 =4 * randn(105) + 52
# comparing samples
Stat1, p1 = t_test_ind(data_1, data_2)
print('t=%f, p1=%f' % (sta1t, p1))
```

8.1.4 F –test

F test is a statistical test which has F-distribution under null hypothesis. F test uses F statistics which actually is the ratio of two different variances; hence, they use F Distribution. Such test is applicable in comparing two regression models to check for statistical significance. Moreover, unlike Z and t-Test, where comparison is done on two datasets for inferring results, in F Test compares two variances.

8.2 CONCEPT OF RANDOM VARIABLE

A random variable in Python Statistical Inference is a variable whose values are possibly theoutcomes of random process (Julier, 2004). There are two main categories of random variables such as discrete and continuous variables.

8.2.1 Discrete Random Variable

The variable which takes only one countable number of different numbers and values and also which can be quantified. For instance, ***A*** to be possible number which comes up when rolling a fair dice. ***A*** can take different values: [1, 2, 3, 4, 5, 6]. Hence, is a discrete random variable.

- **Probability Distribution:**

It is associated with list of probabilities along with each possible value. It is also refereed as *probability function.* Mathematically, it can be interpreted as, suppose there is random variable ***Y*** which may take ***n*** different values with probability that $Y = y_i$ can be defined as $P(Y = y_i) = p_i$. In this case, probability pi need to satisfy following conditions:

1. 0 ≤ pi ≤ 1 for every i
2. p1+p2+p3+…..+ pn = 1

Bernoulli distribution, Binomial distribution and Poisson distribution can be considered as best examples of Discrete Probability Distributions.

8.2.2 Continuous Random Variable

The variable which can take infinite number of values is continuous random variable. For example, Y is taken as variable to store the height of players in a group. Such variables are interpreted as continuous random variables over interval of class, such variables are inferred by using area under curve or by integrals. Distribution of these variables is also interpreted as probability distribution functions. It can be demonstrated by using p(x), which must also satisfy following condition:

1. When curve is not having negative values p(x)>0
2. When area under curve =1

A curve that meets such requirements is taken as a density curve. Normal distribution, Exponential distribution and Beta distribution are common examples of Continuous Probability Distributions .

8.2.3 Importing some important distributions in Python

- **Uniform Distribution**

In this distribution all the outcomes of probability distribution are equally like. In Python, this distribution can be visualized by importing *uniform function* from *scipy.stats* module as:

```
     # import uniform distribution
  from scipy.stats import uniform
     # random numbers
     N=1000
     Start=5
     Width=10
Data_uniform= uniform.rvs (size=n, loca=start, scale=width)
```

- **Code for uniform distribution in Python can be written as:**

```
From numpy import random
A=random.uniform(size(3,3))
Print(a)
```

- **Normal Distribution**

It is also considered as bell curve and occurs quite naturally in various situations. It is also called as Gaussian distribution and most commonly applicable in statistical inference. It poses bell shaped density curve and is defined by mean (μ) and standard deviation (σ).Here, the density curve is symmetric and centered about its mean. In Python, it can be imported as:

```
from scipy.stats import norm

Data1_normal = norm.rvs (size=1000, loca=0, scale=1)
```

Where *scipy.stats* is a relative module, *norm.rvs ()* method, *loc* is mean of distribution, *scale* is standard deviation and *size* is total number of random variables. Random normal distribution of size (3*3) can be generated by:

```
From numpy import random

A= random.normal(size=(3,3))

Print(a)
```

- **Gamma Distribution**

It belongs to two-parametric family under continuous probability distribution. It is used rarely. Most of the times it is used in modeling in engineering areas where the variables are always positive with skewed results. This can be imported in Python as:

```
from scipy.stats import gamma

Data1_gamma = gamma.rvs (x = 10, size_s= 1000)
```

- **Exponential Distribution**

This distribution is used to describe the time interval between different events in a process of Poisson point i.e., the process in which any event occurs in a continuous and independent way with constant and average piece of rate. In python it can be imported as:

```
from sipy.stats import expon

Data1_expon = expon.rvs (scale=1, loca=0, size = 1000)
```

- **Poisson Distribution**

This distribution is used to show the event occurrence within specific period of time. For instance, users visited a website in a particular interval is a Poisson Process in itself. An event can occur n number of times and average number of events in an particular interval is depicted as lambda λ. Poisson distribution can be imported in Python as:

from scipy.stats import poisson

Data_poisson = poisson.rvs (mu= 2, size, 1000)

- **Bernoulli Distribution**

In this distribution, only two outcomes are expected to come such as failure or success and the probability of each success and failure is entirely same for all number of trials and is known as Binomial distribution. In Python, this distribution can be added as:

from scipy.stats import binom

Data1_binom = binom.rvs (n = 20, p = 0.5, size = 1000)

8.3 PROBABILITY MASS FUNCTION

This function describes the probability which is associated with given random number y (Vondrejc, 2019). The function is depicted as P(y).

Example, when rolling a die infinitely and looking on proportion 1, then 2 and so on. The random variable can correspond to outcome of a dice roll. So, random variable can take following discrete values i.e., 1, 2, 3, 4, 5 or 6. The main aim of the probability mass function is to describe possibility/probability of every possible value. In this example, there is probability to get 1, 2 and so on. For example in rolling of dice, there is equal probability of getting any number, which can be written as:

$$P(y = 1) = P\ (y = 2)$$

$$=P(y=3)$$

$$=P(y=4)$$

$$=P(y=5)$$

$$=P(y=6)$$

So, 1/6.

In this way, distribution demonstrates similar probability for every possible value, and called as uniform distribution. Probability mass function looks like:

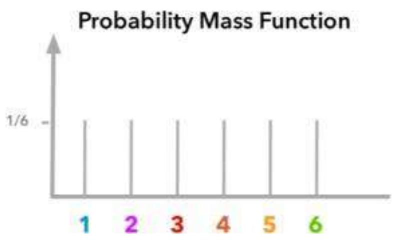


Fig. 1 Probability mass function (for dice)

Here, x-axis depicts outcome and y axis depicts probability. The code that can demonstrate this in more clear way and can be written as:

```
Numb_throws=1000
Outcomes=np.zero(numb_throws)
For I in range(numb_throws);
#when rolling a dice
Outcome=np.random.choiceA(„1‟,‟2‟,‟3‟,‟4‟,‟5‟,‟6‟)
Outcome[i]=outcome
val.cnt=np.uniquee(outcomes, return.count=True)
prop=cnt/len(outcomes)
# after rolling dice 1000 times, plotting of results
Plt.bar(val,propi)
Plt.xlabel("outcome")
Plt.ylabel("probability")
Plt.show()
Plt.close()
```

8.3.1 Properties of Probability Mass Function

Probability Mass function, if and only if,

$$\forall x \in x,\ 0 \leq P(x) \leq 1$$

Here, symbol ∀ depicts for any or for all. This represents, for every x within range of x (in case of rolling dice, set of possible values are 1, 2,3, 4,5 and 6). The probability of each outcome can vary between 0 and 1. Here, 0 represents event has not occurred and 1 represents event has occurred. In particular case of dice, the probability of each value is 1/6 i.e., between 0 and 1.

8.3.2 Probability Density Function

All the discrete variables cannot take infinite values at a certain period of time. Still, it needs to state probability which is associated with all possible outcomes. The equivalence of probability mass function with continuous variable is known as **Probability Density Function.**

- **Important property of Probability Density Function**

$$\forall x \in x, p(x) \geq 0$$

Here, p(x) needs not to be less than 1, because it is not corresponding to probability.

8.4 MATHEMATICAL EXPECTATION

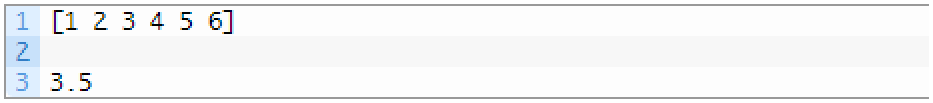
Probability describes happening of certain events and occurrence of particular event depends upon previous experience. The Mathematical Expectation represents all those events which are almost impossible for any experiment. Probability for that particular event remains 0 and Probability of an event remains 1 where both of the numerator and denominator remain equal.

Mathematical Expectation is also referred as expected value, that uses the notation E[X]. This can be computed as probability weighted sum of all possible values which can be drawn by as follows:

E[X]= sum (x1*p1, x2*p2*………*xnpn)

Following is the code with 6-element vector and calculation of mean:

```
From numpy import array

From numpy import mean

M1 = array_example ([1,2,3,4,5,6])

Print (M1)

Result = mean (M1)

Print (result)
```



```
[1 2 3 4 5 6]

3.5
```

- **Properties of mathematical Expectation-**

1. For two variables X and Y, total sum of these two variables is equals to sum of mathemeatical expectationof X and Y respectively i.e., E(X+Y)= E(x)+E(Y).
2. For two independent variables, mathematical expectation would be the product of those considered variables.
3. Mathematical Expectation of constant sum and considered function of any random variable is equals to sum of that constant of function of given random variable.
4. The mathematical expectation of sum and product of a constant and function of any random variable and another constant is all equals to sum of product of that constant and mathematical expectation of given function and random variable and constant.
5. Linear combination of all random variables and a constant is all equals to sum of product of n constants and mathematical expectation of all n numbers.

8.5 MOMENTS

Moments are used to calculate the n moments about mean for given sample i.e., all the array elements with particular axis of that given array (Fan, 2016). In Python, Moments can be calculated as:

```
From scipy import stats

import numpy as np

arry = np.array ([1,27, 31, 21, 13, 9],[12,8,4,8,7,10])

print (“0th moments is :”, stats.moment(arry,moment=0))
```

8.5.1 Moment generating function and characteristic function

```
    from random import choice
    import matplotlib.pyplot as plt
    import numpy as np
    exp_value = lambda values: sum(values) / len(values)
    std_deviation = lambda values, exp_value: np.sqrt(sum([(v - exp_value)**2 for v in values]) /
    len(values))

 muu, sigmaa = 40, 1
    population = np.random.normal(muu, sigmaa, 100000)
    mean = expected_value(population)
    print(
'''population: Expected_value: {0} Standard_deviation: {1}
       '''.format(mean, standard_deviation(population, mean))
    )
    plt.hist(population, 70, density=True)
    plt.show()

    randomly_select_items = [choice(population) for _ in range(63)]
    mean = exp_value(randomly_selected_items)
    s_d = std_deviation(randomly_selected_items, mean)
    print(format(mean, s_d))
  plt.hist(randomly_selected_items, 10, density=True)
    plt.show()
    xsss = np.arange(63, 44, 0.001)
    actual_ys = norm.pdf(xs, muu, sigmaa)
    ys = norm.pdf(xs, mean, s_d)
    plt.plot(xs, actual_ys, label='actual_population_distribution')
    plt.plot(xs, ys, label='sample_distribution')
    plt.legend()
    plt.show()
```

8.6 PRACTICE QUESTIONS

Q1. What is Statistical Inference?

Q2. What are different Inferential Statistics available can be used along with Python? Q3. What are random variable?

Q4. Explain difference between F-test and t-test along with suitable example.

Q5. How Z scores and Z test is calculated. Demonstrate with suitable dataset.

Q6. How moments can be imported and calculated in Python.

Q7. Explain different properties of Mathematical Expectation.

Q8. Explain different distributions. Which distribution is frequently used and why?

REFERENCES

[1] Vondrejc, J. & Matthies, H. G. (2019) “Accurate Computation of Conditional Expectation for Highly Nonlinear Problems”, ,SIAM/ASA Journal on Uncertainty Quantification 7(4), pp. 1349-1368.
[2] Julier, S. J. & Uhlmann, J. K. (2004) “Unscented Filtering and Non linear Estimation”, Proceedings of IEEE, 92, pp. 401-422.
[3] Ale, A. , Kirk, P. & & Stumpf, M.P. (2013) “A General moment expansion method for stochastic kinetic models”, Journal of Chemistry Physics, 138 (17).
[4] Fan, S. , Geissmann, Q. , Lakatos, E. & Lukauskas, S. (2016) “Means: Python package for Moment Expansion Approximation, Inference and Simulation”, Bioinformatics, 32 (18).